UNIVERSITE DE NEUCHATEL

FACULTÉ DES LETTRES ET SCIENCES HUMAINES

Méthodes statistiques d'analyses longitudinales dans le domaine de l'immigration. Application au parcours de vie des étudiants africains en Suisse avec l'utilisation de données administratives

THÈSE DE DOCTORAT

Présentée à la

Faculté des lettres et sciences humaines
de l'Université de Neuchâtel

pour l'obtention du grade de
Docteur en sciences humaines et sociales

par

Mamadou Pathé Barry

Directeur de thèse

Prof. Etienne Piguet

Codirectrice de thèse

Dr. Katia Iglesias

Jury

Dr. Alina Matei, Unine
Institut de statistique

Prof. Philippe Wanner, Unige
Institut de démographie et socioéconomie

Neuchâtel, le 10 septembre 2019

UNIVERSITÉ DE
NEUCHÂTEL

Faculté des Lettres et Sciences Humaines
Espace Tilo-Frey 1
CH - 2000 Neuchâtel
Tél : +41 (0)32 718 17 00
Fax : +41 (0)32 718 17 71
secretariat.lettres@unine.ch

IMPRIMATUR

La Faculté des lettres et sciences humaines de l'Université de Neuchâtel, sur les rapports de M. Etienne Piguet, co-directeur de thèse, professeur, Université de Neuchâtel ; Mme Katia Iglesias, co-directrice de thèse, professeure HES, Haute Ecole de Santé de Fribourg ; Mme Alina Matei, professeure titulaire, Institut de statistique, Université de Neuchâtel ; M. Philippe Wanner, professeur, Institut de démographie et socioéconomie, Université de Genève autorise l'impression de la thèse présentée par M. Mamadou Pathé Barry en laissant à l'auteur la responsabilité des opinions énoncées.

i.A. P. Schn

Neuchâtel, le 10 septembre 2019

Le doyen
Pierre Alain Mariaux

PROF. PIETER SCHNYDER

Table des matières

Remerciements

Cette thèse est l'aboutissement d'un long parcours académique qui a commencé en Guinée pour s'achever en Suisse. Ce long parcours académique est indissociable de mon parcours migratoire en Suisse qui n'a pas été parmi les plus simples.

C'est avec beaucoup d'émotions que je redige ces quelques lignes pour remercier toutes les personnes qui, de près ou de loin, ont contribué à mon éducation ou qui m'ont soutenu lors des interminables démarches administratives visant à m'offrir le droit de réaliser le rêve qui me hantait dès mon jeune âge : faire un doctorat.

J'aimerais remercier le Professeur Etienne Piguet, mon directeur de thèse, pour l'encadrement, les conseils ainsi que pour la confiance qu'il m'a accordée. Je remercie également la Dr. Katia Iglesias, codirectrice de ce travail pour son aide, les conseils, ainsi que pour toute cette énergie positive qu'elle m'a transmise tout le long de cette thèse. J'ai aussi été l'assistant du Dr. Katia pour les cours de statistique, j'ai eu beaucoup de plaisir de travailler avec elle et j'ai beaucoup appris durant ces années d'assistanat. J'ai pu exercer avec toute sa confiance ma passion pour l'enseignement et l'encadrement des étudiants.

Mes remerciements vont également à la Professeure Janine Dahinden qui m'a offert l'oppotunité de faire cette thèse et qui m'a accompagné dans les démarches administratives auprès des autorités migratoires tout en ayant la patience d'attendre jusqu'à l'issue finale de ces démarches administratives. Au cours de ces longues et périlleuses démarches administratives, l'intervention, au nom de la solidarité universitaire, du Professeur Minh Son Nguyen a été l'ouverture de la porte vers le doctorat. Je vous prie de trouver ici l'expression de ma profonde reconnaissance.

Je remercie également le Professeur Philippe Wanner et la Dr. Alina Matei d'avoir accepté de faire partie du jury et pour leurs remarques très constructives.

Le chapitre 13 de cette thèse a été consacré aux données et ces mêmes données ont été utilisées pour répondre à mes questions de recherche. Je remercie l'Office fédéral de la statistique (OFS) pour la mise à disposition des données sans lesquelles cette thèse n'allait pas voir le jour.

Je remercie mon père, ma mère ainsi que mes frères et soeurs pour leur soutien et encouragements tout au long de ces longues études. Je remercie tout particulièrement mon grand frère Alpha Amadou Barry qui m'a attiré dès mon jeune âge vers les sciences ainsi que pour tout l'encadrement dont j'ai bénéficié de sa part. J'exprime aussi ma profonde reconnaissance à Christine Venard pour tout son soutien, ses conseils, sa présence et ses encouragements.

Mes remerciements vont également à mes collègues du laboratoire d'études des processus sociaux pour la bonne ambiance ainsi que pour les moments très sympathiques que nous avons passé ensemble.

Merci à mes amis, Christelle Cocco, Allan Estivalet, Issa Somparé, Vera Von Flüe et Odile Kaudjhis Molo pour leur présence, leurs encouragements ainsi que pour leurs conseils. Je suis très chanceux de les avoir à mes côtés.

Le principe des vies liées a voulu que je rencontre mon épouse Maeda après mes études de master. Cette rencontre a été le début de la construction du parcours de vie familial par la naissance de mes filles Kindy en 2013 et Nafina en 2016. Concilier la vie familiale et une thèse nécessite une organisation sans faille, un immense défi à relever et de grands sacrifices. Je remercie mon épouse Maeda qui en a beaucoup fait en aménageant des horaires de travail de nuit ou les weekends pour que je puisse avancer sur la thèse. Je remercie infiniment ma famille sur les deux continents et cette thèse leur est dédiée.

Résumé

Nombreux sont les étudiants qui décident de quitter leurs pays pour aller se former à l'étranger. Cette mobilité étudiante n'est pas un fait nouveau, c'est un phénomène qui s'inscrit dans une tradition académique aussi ancienne que l'histoire des universités elles-mêmes.

Dans cette thèse, nous nous intéressons au parcours de vie des étudiants internationaux de manière générale et particulièrement à celui des étudiants africains qui ont quitté leurs pays d'origine dans le but de poursuivre une formation universitaire en Suisse. L'augmentation très rapide de l'effectif des étudiants internationaux et africains dans le monde et en Suisse nous a poussés à nous intéresser à cette problématique dans une perspective de parcours de vie. Nous commençons ainsi par une introduction à l'analyse des parcours de vie pour ensuite présenter les modèles statistiques les plus couramment utilisés dans ce domaine. Ces modèles se subdivisent en trois groupes : non paramétriques, paramétriques et le modèle semi-paramétrique de Cox.

L'objectif principal de la thèse est de transposer ces méthodes statistiques, qui sont très largement utilisées dans d'autres domaines, dans le but d'analyser l'immigration avec des données longitudinales administratives.

Les modèles sont présentés avec des exemples d'application. Ces derniers sont comparés entre eux afin de mettre en lumière les avantages et les inconvénients de chaque modèle ainsi que les questions de recherche auxquelles ils permettent de répondre. Nous montrons aussi comment choisir le modèle le plus adapté parmi tous les modèles présentés. Les modèles de régression logistiques à temps discret sont ensuite abordés dans le but de remédier à certaines insuffisances des modèles paramétriques et du modèle de Cox. Nous présentons aussi les différentes sources de données longitudinales pour ensuite introduire les données administratives livrées par l'Office fédéral de la statistique (OFS). Ces données proviennent des bases de données des étrangers (ZAR, ZEMIS) et de la base de données des étudiants (LABB). Ces deux bases de données ont été fusionnées dans le but de mettre en commun les parcours migratoire et académique des étudiants internationaux et des étudiants africains.

Nous décrivons ensuite l'évolution de l'effectif des étudiants africains en Suisse durant ces dernières années (pour la période allant de 1997 à 2014), les nationalités les plus représentées, les domaines d'études privilégiés par ces étudiants ainsi que les universités suisses qui accueillent le plus grand nombre d'étudiants africains. Il ressort de nos analyses que l'effectif des étudiants africains en Suisse par nationalité est très inégal ; l'Afrique du Nord à elle seule représente près de la moitié de l'effectif total des étudiants africains en Suisse. En Afrique subsaharienne, le Cameroun est le pays le plus représenté dans les universités suisses suivi par le Sénégal, la Côte d'Ivoire et Madagascar, mais dans des proportions moindres. Les étudiants africains en Suisse s'orientent généralement vers les branches techniques et économiques au niveau du bachelor et sont nombreux à suivre ces formations dans les Hautes Ecoles Spécialisées (HES). Au niveau du master, il ressort que ces étudiants sont nombreux à s'orienter vers les sciences exactes et naturelles, les branches techniques et économiques dans les universités et les écoles polytechniques fédérales. Les universités romandes sont celles qui accueillent le plus d'étudiants africains tandis qu'en Suisse alémanique, c'est l'Université de Bâle.

A l'issue de cette analyse, nous nous sommes intéressés aux facteurs explicatifs du fait que les étudiants internationaux et les étudiants africains prolongent le séjour en Suisse après les études. Parmi ces facteurs explicatifs, nous relevons en particuliers des caractéristiques démographiques, migratoires et académiques.

Mots clés : Etudiants africains, étudiants internationaux, étudiants étrangers, données administratives, hautes écoles universitaires (HEU), parcours de vie, mobilité des étudiants

Abstract

Many students leave their country to study abroad. This student mobility is not a new phenomenon ; it is part of an academic tradition that is as old as the history of universities themselves. In this thesis, we are interested in the life courses of international students in general and African students in particular, who have left their home country in order to obtain a university degree in Switzerland. The very rapid increase in the number of international and African students in the world and in Switzerland has led us to take an interest in this issue from a life course perspective. We begin with an introduction to life course analysis and then present the most commonly used statistical models in this area. These models are subdivided into three groups : non-parametric, parametric and semi-parametric model. The main objective of the thesis is to transpose these statistical methods, which are widely used in other fields, to analyze immigration with longitudinal administrative data. The models are presented with application examples. These examples are compared with each other to highlight the advantages and disadvantages of each model as well as the research questions they can answer. The study aims at showing how to choose the most suitable model among all the presented models. The discrete time logistic regression models are then discussed in order to address some of the weaknesses of the parametric models and the Cox model. First, we present the different sources of longitudinal data and introduce administrative data delivered by the Swiss Federal Statistical Office (FSO), which come from foreigners' databases (ZAR, ZEMIS) and the student database (LABB). These two databases have been merged to link the migration and academic paths of international and African students. Then, we describe the evolution of the number of African students in Switzerland over the past years (1997 to 2014), the most represented nationalities, the fields of study favored by these students, and the Swiss universities that host the largest number of African students. Our analysis shows that the number of African students in Switzerland is very unequal as far as nationality is concerned ; North Africa alone accounts for almost half of all African students in Switzerland. In sub-Saharan Africa, Cameroon is the most represented country in Swiss universities, followed by Senegal, Ivory Coast and Madagascar, but with lower numbers. African students in Switzerland are oriented towards technical and economic branches/fields at Bachelor level, and many of them follow these courses at the University of Applied Sciences and Arts Western Switzerland (in French HES.SO). At Master's level, it appears that many of these students choose Natural Sciences, Technical and Economic branches at Universities and Federal Polytechnic schools. Most African students are hosted by French-speaking universities, and in the German-speaking part of Switzerland, the University of Basel hosts most of them. At the end of the analysis, we are interested in the reasons why international and African students extend their stay in Switzerland after graduation. Among the explanatory factors, we note in particular, demographic, migratory and academic factors. Our results show a positive and significant impact of marriage, gender, entry cohort, continent of origin, and a negative impact of high school attended on the probability of extending stay in Switzerland after graduation for international students. As for African students, our results show a negative and significant impact of age, high school attended, length of stay, and a positive and significant impact of gender, cohort of admission and marriage on the chance of extending the stay in Switzerland after studies.

Keywords : African students, international students, foreign students, administrative data, universities and institutes of technology (UIT), Life Course, students mobility

Introduction

Les étudiants sont de plus en plus nombreux à franchir les frontières pour passer une à plusieurs années d'études dans un pays étranger. Selon les chiffres de l'OCDE [1], plus de 4.5 millions d'étudiants suivent une formation de niveau supérieur dans un pays étranger (OCDE, 2014). Le début du siècle est donc marqué par une importante augmentation du nombre d'étudiants étrangers dans le monde. L'Australie, l'Autriche, le Luxembourg, la Nouvelle Zélande, le Royaume-Uni et la Suisse sont les pays où le nombre d'étudiants en mobilité internationale est le plus élevé en pourcentage de l'effectif total de l'enseignement tertiaire. En valeur absolue, les effectifs les plus élevés d'étudiants en mobilité internationale sont originaires de Chine, de la Corée du Sud et de l'Inde. Les étudiants asiatiques représentent 53% de l'effectif mondial des étudiants en formation à l'étranger (OCDE, 2014).

Notons cependant qu'il existe un immense déséquilibre entre les pays du Nord et ceux du Sud : les migrations étudiantes suivent principalement les directions Nord-Nord ou Sud-Nord avec des flux migratoires beaucoup plus importants dans la direction Nord-Nord. Cela s'explique d'une part, par l'existence de programmes visant à encourager la mobilité des étudiants européens (Erasmus), d'accords internationaux entre les pays du Nord et, d'autre part, par le fait que certains étudiants étrangers originaires du Sud sont considérés par les services de l'immigration des pays du Nord comme une population à risque (Latreche, 2001; Keller-Gerber, 2017). Par population à risque, on sous-entend une catégorie d'étudiants qui émigre sous prétexte d'études d'une part, d'autre part, ce sont tous les étudiants qui après les études ne rentrent pas dans leur pays d'origine en cherchant à s'établir dans le pays d'accueil ou en migrant vers d'autres pays que le pays d'origine à la fin des études.

L'intérêt commercial a renforcé les liens diplomatiques avec les pays émergents, ce qui fait que certains pays de l'Amérique latine et d'Asie font partie des axes géographiques prioritaires pour la mise en place de conventions d'échanges (Terrier, 2010). Cette augmentation rapide du nombre d'étudiants internationaux à travers le monde a poussé certains chercheurs à s'intéresser à la problématique de la mobilité des étudiants en l'analysant sous plusieurs aspects : fuite des cerveaux, apport culturel, activité économique lucrative pour le pays d'accueil, forme déguisée d'immigration pour contourner les difficultés administratives liées à l'immigration classique. D'autres études se sont intéressées à la proportion d'étudiants diplômés qui restent dans le pays d'accueil quelques années après l'obtention du diplôme, ou aux facteurs qui poussent les étudiants à rentrer dans le pays d'origine ou à rester dans le pays d'accueil. Ces études mentionnent comme facteurs de rétention les opportunités de carrières dans le pays d'accueil ou les troubles politiques et sociaux dans le pays d'origine. Dans cette recherche, nous nous intéressons au cas particulier de la Suisse à travers des questions de recherche que nous nous posons sur le parcours de vie des étudiants internationaux d'une part, et, d'autre part, sur le parcours de vie des étudiants africains entrés en Suisse au bénéfice d'un permis B pour études en s'intéressant aux facteurs explicatifs du fait de prolonger le séjour en Suisse après les études.

On peut dès lors se poser la question suivante : comment étudier le parcours de vie de ces étudiants ?

La réponse à cette question fait appel à trois autres questions qui sont : 1) que veut-on mesurer dans le parcours de vie de ces étudiants (la question de recherche) ?, 2) comment va-t-on mesurer ce qui nous intéresse dans ce parcours de vie (les méthodes statistiques) ?, 3) de quoi dispose-t-on pour mesurer ce qu'on veut (les données) ?

Dans cette thèse, l'objectif principal est la recherche du meilleur modèle statistique pour analyser des données longitudinales administratives dans le but de répondre à nos questions de recherche. Les questions de recherche seront formulées, des méthodes statistiques seront comparées, les données que nous allons utiliser seront aussi

1. Organisation de coopération et de développement économiques

présentées. Ensuite, nous choisirons le modèle le plus adapté à nos données pour répondre à nos questions de recherche.

On s'intéressera plus précisement aux variables explicatives du fait que les étudiants internationaux et les étudiants africains prolongent le séjour en Suisse après les études. En effet, la politique migratoire suisse est une politique restrictive. Nous partons du pricipe que pour prolonger le séjour en Suisse après les études, certaines nationalités offrent plus d'avantages que d'autres. Les étudiants internationaux issus des Etats tiers n'avaient que deux possibilités [2] jusqu'en 2011 pour rester en Suisse après leurs études : le mariage ou la maladie. C'est à partir de 2011 qu'une troisième possibilité a été offerte à ces étudiants de pouvoir obtenir un permis de six mois après les études dans le but de chercher un travail. Nous partons donc du principe que le mariage est une variable importante dans le fait de pouvoir prolonger le séjour en Suisse après les études. D'autres variables comme le sexe, l'âge, les hautes écoles fréquentées ainsi que la cohorte d'entrées en Suisse sont aussi prises en compte pour comprendre le phénomène étudié.

Nous formulons ainsi notre première question de recherche qui est consacrée aux étudiants internationaux de manière générale (y compris les étudiants africains) comme suit :

Le fait que les étudiants internationaux arrivés en Suisse entre 1997 et 2009 prolongent le séjour en Suisse après leurs études dépend-il du mariage de ces étudiants en Suisse durant la période d'observation, de l'âge de ces étudiants à l'arrivée en Suisse, du sexe, de la haute école fréquentée [3] *(Unifr, Unige, Unil, Unine, Uniba, UniZH, HES , EPFZ, EPFL), du continent d'origine des étudiants (Afrique, Asie, Europe, Amérique) et de la cohorte d'entrées (entrées entre 1997 et 2000, entrées entre 2001 et 2005, entrées entre 2006 et 2009) ?*

La deuxième question de recherche qui ne concerne que les étudiants africains est formulée comme suit :

Le fait que les étudiants africains arrivés en Suisse entre 2003 et 2012 prolongent le séjour en Suisse après leurs études dépend-il du mariage de ces étudiants en Suisse après leur arrivée, de l'âge, du sexe des étudiants, des hautes écoles fréquentées (Unifr, Unige, Unil, Unine, Uniba, UniZH, HES , EPFZ), de la cohorte d'entrées en Suisse (entrées entre 2003 et 2005, entrées entre 2006 et 2008, entrées entre 2009 et 2012) et de la région d'origine [4] *de ces étudiants (AO, AE, AN, AC, AA) ?*

Dans cette thèse, nous ne nous contentons pas de répondre directement à nos questions de recherche, nous commençons par faire une revue des différentes méthodes statistiques couramment utilisées dans l'analyse de parcours de vie dans le but de trouver le modèle le plus adapté pour répondre à nos questions de recherche.

Par ailleurs, le premier apport de cette thèse résidera dans la vulgarisation de ces méthodes dans le but de les rendre accessible à un public non statisticien, notamment aux chercheurs dans les sciences humaines et sociales. Les différentes méthodes qui seront présentées sont largement diffusées dans la recherche médicale sous l'appellation *d'analyse de survie* mais peu utilisées dans les sciences humaines et sociales. Le deuxième apport de la thèse est leur application dans le but d'analyser l'immigration de manière générale et particulièrement la mobilité des étudiants internationaux et africains en Suisse à l'aide de données administratives. Le troisième aspect fait référence aux données disponibles.

En effet, celles-ci sont longitudinales dans le sens où les étudiants une fois arrivés en Suisse sont enregistrés et suivis d'année en année durant tout le long de leur séjour en Suisse. Ainsi, un étudiant qui a séjourné en Suisse pendant six ans par exemple apparaîtra sur six lignes différentes dans notre base de données, chaque ligne représentant une année de séjour. Une fois les méthodes statistiques et les données présentées, nous partons donc à la recherche du modèle le plus adapté pour répondre à nos questions de recherche.

La recherche du meilleur modèle nous a donc conduit vers les modèles de régression logistique à temps discret dont fait partie le modèle logit à temps discret. Ce dernier est le modèle le plus compatible avec nos données ; il sera donc utilisé pour répondre à nos deux questions de recherche.

2. Une troisième possibilité existe mais elle est très difficile à faire valoir : 'a la fin des études, l'étudiant trouve un poste et l'employeur prouve aux autorités que cette personne est la seule à disposer des compétences nécessaires pour occuper ce poste. Ceci étant très compliqué à prouver, nous n'avons pas pris en compte, dans cette thèse cette possibilité de prolonger le séjour en Suisse après les études.

3. Unifr= Université de Fribourg, Unige = Université de Genève, Unil= Université de Lausanne, Unine= Université de Neuchâtel, Uniba= Université de Bâle, UniZH = Université de Zurich, HES = Hautes Ecoles Spécialisées , EPFZ = Ecole Polytechnique Fédérale de Zurich, EPFL = Ecole Polytechnique Fédérale de Lausanne

4. AO = Afrique de l'Ouest, AE= Afrique de l'Est, AN = Afrique du Nord, AC= Afrique Centrale, AA=Afrique Australe

Dès lors, pour arriver à ce résultat, nous avons subdivisé cette thèse en cinq parties dans le but de répondre à nos questions de recherche.

Dans la première partie, nous avons procédé à une revue de la littérature sur la mobilité internationale des étudiants dans le monde et en Suisse avant de présenter nos questions de recherche. Il ressort de cette revue de la littérature que les différentes études sur la mobilité des étudiants ne touchent pas leur parcours de vie et elles sont souvent basées soit sur des données d'enquêtes, soit sur des données récoltées auprès d'institutions internationales ou au niveau des pays d'accueil de ces étudiants.

Dans la deuxième partie, nous abordons les méthodes statistiques. Nous avons d'abord procédé à une introduction à l'analyse des parcours de vie en présentant des exemples de parcours de vie puis les différentes fonctions à la base de l'analyse des parcours de vie dans le but d'introduire les différents modèles ciblés. Les modèles statitistiques ciblés sont ceux qui sont le plus couramment utilisés pour analyser l'immigration de manière générale dans une perspective de parcours de vie. Ces méthodes se subdivisent en modèles descriptifs et en modèles explicatifs. Dans les modèles descriptifs, nous avons présenté les méthodes de Kaplan-Meier et actuarielle avec des exemples d'application avant de comparer les deux méthodes et de mentionner le type de questions de recherche auxquelles ces méthodes permettent de répondre.

Dans les modèles explicatifs, nous avons présenté les méthodes de régression paramétrique et semi-paramétrique de Cox. Dans les méthodes paramétriques, les régressions exponentielle, de Weibull, de Gompertz, log-normale, log-logistique et gamma sont présentées avec un exemple d'application pour chacune de ces méthodes ainsi qu'une comparaison des différents résultats obtenus d'une méthode à l'autre.

En raison du délai d'attente très long avant la livraison des données, les bases de données qui ont été utilisées pour développer les exemples pour les différentes méthodes proviennent des livres ou parfois sont des données fictives. Lorsque ces bases de données ne concernent pas le domaine de l'immigration, de nouvelles variables ont été crées, d'autres ont été recodées dans le but de ramener la problématique à celle des étudiants internationaux.

Nous mettrons également en évidence le fait que ces modèles se sudivisent en modèles à risque proportionnel et en modèles à temps de sorties accélérées. Dans les modèles à risque proportionnel, la variable « dépendante » est le risque alors que dans les modèles à temps de sorties accélérées, la variable « dépendante » est la durée de séjour. Les modèles à hasard proportionnel permettent de répondre à des questions de recherche qui ont pour objectifs de connaître l'impact d'une série de variables explicatives sur le risque d'occurrence d'un phénomène donné. Les modèles à temps de sorties accélérées permettent de répondre à des questions de recherche qui s'intéressent à l'impact d'une série de variables explicatives sur la durée de séjour. Le troisième modèle présenté est le modèle semi-paramétrique de Cox qui est une combinaison de modèles non paramétriques et paramétriques ; ce modèle est également un modèle de régression dont la variable « dépendante » est aussi le risque. Pour mettre en pratique ces méthodes, il nous faut des données qui puissent nous permettre de suivre les individus dans le temps, c'est-à-dire des données longitudinales. Le caractère annuel des données nous a poussés à introduire une autre famille de modèles appelée modèles de régressions logistiques à temps discret. Nous ressortons également le fait que cette famille de modèles est plus adaptée aux données annuelles qui présentent un risque important que plusieurs personnes expérimentent l'événement étudié au même moment.

Dans la troisième partie, nous avons d'abord présenté les différentes sources de données longitudinales ainsi que les avantages et les inconvénients de chaque type de données. Ensuite, les données administratives de l'OFS qui seront utilisées dans cette recherche sont aussi présentées ainsi que les différents travaux de préparation de ces données. En effet, ces données proviennent de deux sources différentes qui sont : les données sur les étrangers et les données sur les étudiants. Ces deux sources de données contiennent un identifiant unique qui a été construit à partir du numéro AVS anonymisé ; ce qui rend possible la mise en commun de ces deux bases de données. La méthode de mise en commun des données et la méthode d'extraction de la base de données des étudiants internationaux et africains sont également présentées dans cette section. Ces données sont individuelles et sont récoltées au 31 décembre de chaque année. Nous avons ensuite testé le potentiel d'application de ces différentes méthodes à nos données administratives à travers leur application sur des questions de recherche.

Après la présentation des différents modèles statistiques et les données, nous abordons nos questions de recherche dans la quatrième partie dans le but de répondre à ces dernières. Nous commençons d'abord par

reprendre quelques analyses descriptives de mon article publié dans la revue Géo-Regards sur les étudiants africains en Suisse (Barry, 2017). Cet article présente l'évolution de l'effectif des étudiants africains en Suisse au niveau du bachelor, du master et du doctorat, les filières d'études privilégiées par ces étudiants ainsi que les hautes écoles suisses qui accueillent le plus d'étudiants africains en Suisse. Cette analyse descriptive est complétée par une analyse similaire sur l'évolution dans les hautes écoles suisses de l'effectif des étudiants suisses au niveau du bachelor, du master et du doctorat. Ces résultats corroborent le fait qu'il y a de plus en plus de Suisses qui font des études universitaires, malgré le fait, que cette proportion soit encore inférieure à la moyenne de l'OCDE. On remarque à travers cette analyse que l'effectif des étudiants suisses a augmenté de manière significative ces dernières années dans les différents niveaux de formation. Nous introduisons ensuite nos questions de recherche dont la première porte sur les étudiants internationaux de manière générale et la deuxième sur les étudiants africains.

Dans la première question de recherche, les résultats montrent un impact fort et significatif du mariage sur la chance pour les étudiants internationaux de prolonger le séjour en Suisse après leurs études. Nous observons également un impact négatif et significatif des hautes écoles fréquentées sur les chances de prolonger le séjour en Suisse après les études par rapport au fait de fréquenter l'Université de Berne. Nous obtenons aussi un impact positif de l'âge et du genre sur les chances de prolonger le séjour en Suisse après les études pour les étudiants internationaux alors que la durée de séjour a un impact négatif sur cette chance.

Dans la deuxième question de recherche, qui est consacrée aux étudiants africains, nous observons également un impact fort et significatif du mariage sur les chances de prolonger le séjour en Suisse après les études. En effet, il ressort de nos resultats que le fait d'être marié par rapport au fait de ne pas l'être, toutes choses étant égales par ailleurs, la chance de prolonger le séjour en Suisse après les études pour les étudiants africains est environ 176 fois plus susceptible de se produire.

Dans la cinquième partie de cette thèse, nous effectuons une discussion générale sur les méthodes statistiques, les données ainsi que sur les résulats obtenus sur nos questions de recherche. Cette discussion parlera de la complexité des méthodes statistiques présentées dans cette thèse ainsi que des difficultés de les vulgariser. En ce qui concerne les données, la discussion mettra l'accent sur les données administratives. On verra également que la protection des données peut porter atteinte à la recherche dans la mesure où nous n'avons pas pu mener cette recherche à la hauteur de nos ambitions car certaines variables sensibles n'ont pas été livrées par l'OFS. Nous terminerons par faire quelques recommandations et ensuite nous dégagerons quelques pistes de recherches futures.

Première partie

Les étudiants internationaux dans le monde et en Suisse

Chapitre 1

Revue de la littérature sur les étudiants internationaux dans le monde et en Suisse

Dans cette thèse, il s'agit de l'analyse de parcours de vie des étudiants internationaux en Suisse de manière générale, et, particulièrement du parcours de vie des étudiants africains. Pour introduire cette notion de parcours de vie, considérons les exemples suivants de quatre étudiants africains qui viennent en Suisse pour poursuivre des études universitaires.

Le premier étudiant s'appelle Abdoul, il est d'origine sénégalaise, il est né en 1998 au Sénégal et a obtenu son baccalauréat scientifique au Sénégal en 2015. Abdoul est arrivé en Suisse en 2016 en étant célibataire et est immatriculé à l'Université de Genève pour le bachelor en sciences économiques. A son arrivée en Suisse, Abdoul s'est rendu au contrôle des habitants pour s'annoncer auprès des autorités genevoises. Il a été enregistré dans le registre des étrangers, il a reçu par la suite un permis de séjour pour études communément appelé permis B court séjour, il a également reçu un numéro AVS et poursuit ses études universitaires en Suisse.

Considérons ce deuxième exemple concernant une étudiante camerounaise du nom de Michelle, née en 1996 au Cameroun et qui a obtenu son baccalauréat dans le domaine des sciences sociales en 2015. Elle est arrivée en Suisse en 2018 dans le but de faire un bachelor à la Faculté des lettres et sciences humaines de l'Université de Neuchâtel. A son arrivée en Suisse, elle s'est aussi annoncée aux autorités, elle a reçu un permis de séjour pour études, un numéro AVS et elle poursuit ses études de bachelor dans la même université, elle est célibataire.

Le troisième exemple que nous allons prendre est celui d'un étudiant marocain célibataire, du nom de Khaled qui est né au Maroc en 1990 et qui a obtenu son baccalauréat dans une branche technique en 2010, puis un master en informatique au Maroc en 2015. Khaled est arrivé en Suisse en 2016 dans le but de faire un doctorat en informatique à l'Université de Lausanne. A son arrivée, il a aussi fait les démarches nécessaires auprès des autorités dans le but d'obtenir un permis de séjour et il a aussi reçu un numéro AVS.

Le dernier exemple que nous allons prendre pour illustrer la notion de parcours de vie est celui d'un étudiant Guinéen célibataire, du nom de Sadio qui est né en Guinée en 1986 et qui a obtenu son baccalauréat en sciences mathématiques en 2005. Sadio est arrivé en Suisse en 2006 dans le but de faire un bachelor en mathématiques à l'EPFL. A son arrivée, en Suisse, il a aussi fait les démarches nécessaires auprès des autorités dans le but d'obtenir un permis de séjour et il a aussi reçu un numéro AVS.

Etudier le parcours de vie de ces quatre étudiants africains en Suisse consiste à fixer un instant de début d'observation et un instant de fin d'observation en s'intéressant aux changements qui surviennent dans la vie de ces étudiants ainsi qu'à la chronologie de ces événements durant cette période d'observation. On s'intéressera par exemple à la survenue d'un mariage durant la période d'observation, à l'obtention d'un diplôme quelconque (bachelor, master, doctorat), à l'obtention du premier emploi après les études, par exemple. Dans cet exemple, la date de début d'observation peut être la date de début des études en Suisse et la date de fin peut être la fin des études ou par exemple huit ans après la date de début des études en Suisse.

Ces quatre étudiants ont en commun le fait d'être des étudiants africains qui ont tous choisi de venir en Suisse pour faire des études universitaires. Ils ont tous fait les mêmes démarches en Suisse auprès des autorités dans le

but d'obtenir un permis de séjour. Les initiatives prises par ces étudiants de venir étudier en Suisse s'inscrivent dans une problématique bien connue, qui date d'époques lointaines et qui s'inscrit dans le cadre de la mobilité internationale des étudiants.

La mobilité des étudiants n'est pas un phénomène nouveau ; c'est un phénomène aussi ancien que l'histoire des universités elles-mêmes. Le recrutement d'étudiants venus de l'étranger par les universités a été vu depuis la fin du XIXe siècle comme le point de départ du rayonnement futur des universités sur le plan international (Verger, 1991). Les Universités de Paris, d'Oxford, de Montpellier et de Bologne, qui sont les universités les plus anciennes, ont acquis une réputation dans des contrées bien plus lointaines en accueillant des étudiants venant d'horizons divers depuis le XIIe siècle. La mobilité étudiante au Moyen Âge était étroitement liée au statut social qui se mesurait en termes de distance entre la région d'origine de l'étudiant et l'université de destination. Seuls les étudiants issus de familles riches pouvaient se permettre d'aller étudier dans les régions très lointaines (Garneau, 2007). Les étudiants les plus pauvres devaient se contenter de fréquenter les universités régionales ou les universités les plus proches (Verger, 1991). L'augmentation du nombre d'universités à la fin du Moyen Âge a permis de résoudre ce problème en offrant plus de choix aux étudiants à revenus modestes de se rendre dans d'autres universités augmentant ainsi le nombre d'étudiants en mobilité internationale à travers l'Europe.

Nous commencerons tout d'abord par faire une distinction entre étudiants en mobilité internationale et étudiants étrangers. Par étudiants en mobilité internationale, on sous-entend les étudiants qui ont quitté leur pays d'origine pour se rendre dans un autre pays dans l'intention d'y suivre des études. L'institut de statistique de l'UNESCO, l'OCDE et Eurostat définissent donc les étudiants en mobilité internationale comme ceux qui suivent une formation dans un autre pays que celui dont ils sont résidents ou dans lequel ils étaient scolarisés auparavant. Par étudiants étrangers, on entend tout étudiant qui n'est pas ressortissant du pays où il suit une formation. Ces derniers peuvent ainsi, dans certains cas, être nés dans ce pays et n'avoir pas connu de migration. Les étudiants en mobilité internationale constituent donc un sous-ensemble des étudiants étrangers.

Selon les chiffres de l'OCDE, en 2012 plus de 4.5 millions d'étudiants suivaient une formation de niveau supérieur dans un pays étranger (OCDE, 2014). Entre 1975 et 2012, le nombre d'étudiants en mobilité dans le monde a été multiplié par plus de 5.5, passant de $800'000$ étudiants en 1975 à $4'500'000$ en 2012. Cette mobilité étudiante suit principalement les directions Nord-Nord et Sud-Nord. La direction Nord-Nord s'explique par la volonté de certains Etats de renforcer leurs liens commerciaux, diplomatiques à travers la mise en place d'accords de coopérations universitaires visant à faciliter la mobilité des étudiants, à entretenir des liens sociaux et culturels (Terrier, 2009a; Endrizzi, 2010). Dans l'Union Européenne, on peut citer par exemple la mise en place du programme Erasmus en 1987 qui a pour but de favoriser, faciliter et encourager les échanges d'étudiants entre pays européens [1] (Ballatore et Blöss, 2008). La plupart des programmes d'échanges ou des accords universitaires concernent des pays du Nord ou des pays émergents (Terrier, 2009a). Les mouvements d'étudiants du Sud vers le Nord s'expliquent principalement par le fait que les étudiants ressortissants des pays pauvres sont à la recherche de formations et de compétences que les universités locales ne peuvent pas leur offrir. L'inégal niveau de développement entre le Sud et le Nord, le manque d'offre de formation de qualité dans les branches techniques, informatiques et économiques, le prestige d'étudier dans un établissement post-secondaire de renommée internationale (Efionayi et Piguet, 2014) sont, entre autres, les facteurs explicatifs de l'accroissement du nombre d'étudiants du Sud désirant poursuivre une formation dans les universités du Nord. Il ressort de l'étude coordonnée par Efionayi et Piguet en 2014 dans trois universités ouest-africaines, Côte d'Ivoire (Abidjan), Sénégal (Saint-Louis) et Niger (Niamey) que les étudiants de ces universités se plaignaient des conditions précaires d'études, des ressources mises à disposition (manque d'ordinateurs, de salles, de documentations) ainsi que de l'ingérence de l'Etat dans les affaires académiques. L'étude mentionne aussi le fait que les grèves et l'instabilité politique perturbent le déroulement des cours comme ce fut le cas en Côte d'Ivoire où des étudiants ont vu des années invalidées ou ont connu des années blanches suite aux troubles politiques que le pays a traversés en 2009.

Les problèmes des universités africaines ne se résument pas seulement à ces facteurs, ils sont aussi dus au fait

1. Le programme ERASMUS n'a pas d'influence directe sur les étudiants internationaux car les participants restent immatriculés dans l'industrie d'origine ; on parle de « credit mobility », mais ces étudiants peuvent avoir un impact indirect sur l'accroissement de la mobilité internationale analysée ici.

que la plupart des universités africaines ne sont pas un « produit » de l'Afrique elle-même, mais résultent du passé colonial (Terrier, 2009a). Cet état de fait pose aussi des problèmes dans la gestion de ces universités de manière générale, notamment celui du surpeuplement, de la dégradation rapide des infrastructures existantes (Missine, 1968), de l'encadrement et des grèves à répétition (Efionayi et Piguet, 2014). L'ensemble de tous ces facteurs fait que les étudiants africains sont nombreux à vouloir poursuivre leur formation dans les universités occidentales. Leur effectif est néanmoins en baisse : selon l'Unesco, en 2013, la part des étudiants africains dans la mobilité étudiante mondiale était de 10.5%, soit un effectif de $373'303$ étudiants (Campus France, 2016). En 2011, ce chiffre s'élevait à $412'516$ étudiants soit une diminution de 10.6% ; quant à la mobilité mondiale des étudiants, celle-ci augmentait de 2.6% dans la même période. Cette diminution du nombre d'étudiants s'observe aussi au sein de l'Union Européenne avec une diminution de la proportion d'étudiants africains accueillis entre 2012 et 2013 de 11%, ce qui correspond à un effectif de $22'000$ étudiants en moins. Cette diminution pourrait être la conséquence de la politique migratoire de plus en plus restrictive de certains pays européens en matière d'accueil d'immigrés de manière générale, y compris les étudiants. Cette tendance à la diminution du nombre d'étudiants africains sur le plan mondial s'observe aussi en France qui fut longtemps, et est encore aujourd'hui, par son histoire coloniale, le premier pays d'accueil des étudiants africains au monde. Les étudiants africains restent tout de même les étudiants les plus mobiles avec un taux de mobilité [2] de 3.5%, soit plus du double de la moyenne mondiale de 1.7% en 2005 (Erlich, 2012). Avec cette diminution du nombre d'étudiants africains dans les pays d'accueil habituels, on a assisté ces dernières années à une diversification des destinations (Asie, Afrique et Moyen-Orient) et à un accroissement rapide de la mobilité pour étude dans la direction Sud-Sud notamment en direction de l'Afrique du Sud, du Ghana et des pays d'Afrique du Nord. L'Afrique du Sud est devenue en 2013 le deuxième pays d'accueil des étudiants africains au monde juste derrière la France avec un effectif de $33'053$ étudiants et devant le Royaume-Uni ($32'454$) et les Etats-Unis ($32'212$). Les étudiants accueillis en Afrique du Sud proviennent majoritairement des pays limitrophes. Ces pays, y compris l'Afrique du Sud, sont organisés autour de la SADC (Southern Africa Developpement Community). Cette institution prévoit que chaque pays membre de l'organisation réserve au moins 5% des inscriptions annuelles aux étudiants des autres pays membres (Unesco, 2012). Le Ghana se place à la neuvième place des pays qui accueillent le plus d'étudiants africains dans le monde avec un effectif de $10'009$ étudiants et le Maroc à la douzième place avec un effectif de $6'958$ étudiants africains en 2013 (Campus France, 2016). La Chine et le Moyen-Orient ont mis en place ces dernières années une politique d'attraction des étudiants africains avec l'octroi de nombreuses bourses d'études à destination de ces pays. En ce qui concerne la Chine, celle-ci, pour attirer les étudiants africains vers les universités chinoises et pour faciliter leur intégration, a ouvert des écoles dispensant des cours de mandarin dans quelques pays africains.

Dans cette recherche, nous nous intéressons uniquement aux étudiants africains en mobilité internationale y compris ceux de l'Afrique du Nord et donc à ceux qui ont quitté leurs pays d'origine dans le but de poursuivre une formation universitaire en Suisse. Nous excluons les ressortissants de pays africains vivant et établis en Suisse, ayant la nationalité suisse ou ayant suivi leur scolarité en Suisse. Cette étude inclut dès lors sous le label court « étudiants internationaux » toutes les personnes d'origine africaine qui sont entrées en Suisse au bénéfice d'un permis B pour études (HES [3] , Universités ou EPF [4]).

Les étudiants africains représentent environ 10.5% de l'effectif des étudiants en mobilité dans le monde et la part des étudiants africains en Suisse représentait environ 6% de l'effectif total des étudiants internationaux en 2013 (OFS, 2015). La Suisse, par la qualité de son enseignement, est une destination attractive pour les étudiants africains.

Nous avons donc pu observer que les étudiants sont de plus en plus nombreux à quitter leurs pays d'origine pour passer une à plusieurs années d'études dans un pays étranger. Cet effectif élevé d'étudiants en mobilité à travers le monde a poussé certains chercheurs à s'intéresser à cette problématique.

Les études qui existent sur la mobilité internationale des étudiants l'ont abordée sous différents aspects : économique, fuite des cerveaux, immersion dans une nouvelle culture (Erasmus), immigration sous prétexte d'études,

2. Pour un pays donné, le taux de mobilité se calcule en faisant le rapport entre le nombre d'étudiants partant à l'étranger et le nombre total d'étudiants de ce pays

3. Haute Ecole Spécialisée

4. Ecole Polytechnique Fédérale

rayonnement du pays d'accueil.

En effet, la mobilité des étudiants suit la direction Nord-Nord ou Sud-Nord avec un nombre important d'étudiants originaires des pays en développement qui souhaitent se former dans les universités du Nord. Cet effectif important d'étudiants originaires de pays en voie de développement a poussé certains pays développés à revoir leur politique en matière d'accueil des étrangers de manière générale et particulièrement les étudiants internationaux, soit en supprimant des bourses d'études, soit en introduisant directement les étudiants dans leur politique migratoire globale. En effet, ces pays développés voient en la mobilité des étudiants une manière de contourner les difficultés administratives liées à l'immigration classique (Latreche, 2001; Keller-Gerber, 2017; Ewers et Lewis, 2008; Rakotonarivo, 2013).

L'aspect économique de la mobilité internationale réside dans le fait que ces étudiants sont des acteurs économiques importants pour le pays d'accueil dans la mesure où ces étudiants doivent payer des frais de scolarité importants et doivent supporter d'autres charges non négligeables dans le pays d'accueil. Findlay (Findlay, 2011) mentionne le fait que dans un système d'enseignement supérieur globalement compétitif, le recrutement d'étudiants internationaux à travers le monde n'est pas causé seulement par la volonté de ces étudiants et de leurs familles de vouloir qu'ils se forment dans de grandes écoles à travers le monde. Il mentionne le fait que ce recrutement d'étudiants internationaux est devenu une activité hautement lucrative atteignant un montant de cinq milliards de livres en 2006 pour le Royaume-Uni seulement (Findlay, 2011), ce montant était d'environ 14.5 milliards de dollars aux USA pour l'année académique 2006-2007 (Kim et al., 2011). D'autres chercheurs voient l'accueil des étudiants internationaux comme étant un commerce comme tout autre (Yang, 2003).

La mobilité internationale des étudiants a aussi été étudiée sous l'aspect de la fuite des cerveaux. Cette vision de la mobilité internationale des étudiants est considérée comme une perte de l'élite intellectuelle et scientifique des pays en voie de développement au profit des pays développés (Gaillard et Gaillard, 2002; Mendy, 2014).

Ces différentes études ne touchent pas le parcours de vie de ces étudiants et sont souvent basées soit sur des données d'enquêtes, soit sur des données récoltées auprès d'institutions internationales ou au niveau des pays d'accueil de ces étudiants. Les analyses statistiques effectuées sont souvent descriptives.

Dans cette étude, nous nous intéressons aux méthodes utilisées pour analyser l'immigration de manière générale et particulièrement la mobilité étudiante dans une perspective de parcours de vie. Cela nécessite des données longitudinales et des méthodes statistiques adaptées à ce type de données. Ce qui nous a poussés à approfondir davantage la revue de la littérature dans le but de faire un état des lieux de ce qui existe d'un point de vue méthodologique pour analyser les migrations internationales.

La plupart des recherches quantitatives portant sur les migrations internationales d'Afrique subsaharienne sont basées sur les données produites par les principaux pays de destination sur les effectifs d'étudiants étrangers inscrits dans un établissement d'enseignement supérieur. Ces statistiques sont compilées dans la base de données de l'ISU (Institut Statistique de l'UNESCO) sur l'éducation supérieure (Kabbanji et al., 2013). La littérature sur les mobilités étudiantes s'est essentiellement centrée sur les facteurs qui influencent la décision de poursuivre des études supérieures à l'étranger ainsi que sur le choix du pays de destination et de l'institution d'éducation supérieure dans ce pays (Kabbanji et al., 2013). Ces études sont basées surtout sur des données d'enquêtes transversales ne permettant pas d'examiner les trajectoires individuelles migratoires, scolaires, professionnelles et familiales de la population étudiante.

La connaissance de cette population reste très partielle car peu d'études traitent de la migration étudiante, les recherches existantes se focalisent souvent sur une seule nationalité (Terrier, 2009b). Les études concernent surtout la France, le Canada, l'Angleterre, les USA, mais ne traitent pas des trajectoires de ces étudiants dans ces pays.

Ce manque d'informations motive l'intérêt de faire cette recherche afin de comprendre les trajectoires des étudiants internationaux en Suisse d'une part et d'autre part de dégager d'autres possibilités d'études futures autour du même thème.

L'utilisation de données longitudinales quantitatives nous permettra de combler le manque de littérature sur le devenir de la population d'étudiants internationaux et africains en Suisse. Il est utile à ce sujet de mentionner, dans le cadre d'un débat purement méthodologique, quelques études qui utilisent des méthodes d'analyses de parcours de vie bien que toutes ces études ne s'inscrivent pas dans la thématique de la mobilité des étudiants.

La première est une étude longitudinale des migrations internationales étudiantes ghanéennes, et sénégalaises (Kabbanji et al., 2013). Trois questions sont abordées dans cette étude : en quoi se distinguent les étudiants internationaux ghanéens et sénégalais des autres étudiants qui restent dans leur pays d'origine en termes de parcours scolaires, professionnels et migratoires ? En quoi se distinguent-ils en termes de caractéristiques socio-démographiques ? Quels sont les déterminants individuels et familiaux de l'accès à l'enseignement supérieur dans leur pays d'origine et à l'étranger ?

Pour répondre à ces questions de recherche, les méthodes statistiques suivantes ont été utilisées. L'analyse de séquence est utilisée pour décrire les trajectoires scolaires, professionnelles et migratoires de différents groupes de l'échantillon (étudiants/non-étudiants, nationaux/internationaux). Cette analyse de séquences a permis de comparer les trajectoires scolaires, professionnelles et migratoires des étudiants ghanéens et sénégalais. La méthode d'appariement optimal est utilisée pour identifier une typologie des parcours scolaires, professionnels et migratoires des étudiants nationaux et internationaux. Celle-ci a permis dans un premier temps de mesurer la dissemblance entre chaque paire de séquences de l'échantillon puis, de construire une typologie en créant des groupes de séquences similaires. Il s'agit ensuite de comparer les séquences à travers le calcul des distances basé sur les taux de transition observés d'un état à un autre, puis de regrouper les séquences proches (Robette, 2011). Ensuite, la méthode de classification ascendante hiérarchique a été utilisée pour regrouper les individus les plus semblables selon le critère de Ward. Celui-ci permet de minimiser l'hétérogénéité intra-classes et de maximiser l'hétérogénéité inter-classes. L'analyse biographique en temps discret a été utilisée pour explorer les déterminants des migrations internationales étudiantes ghanéennes et sénégalaises. Cette méthode est appropriée car les données sont longitudinales d'une part, d'autre part, elle permet également d'introduire une dimension temporelle : on peut évaluer par exemple le risque qu'un individu ayant certaines caractéristiques vive un événement qui nous intéresse. Des régressions logistiques multinomiales en temps discret ont été effectuées pour identifier l'impact de différentes variables, toutes choses étant égales par ailleurs, sur la probabilité de vivre un ou des événements précis. Les analyses d'optimal matching et les analyses de survie se distinguent également par leur objet d'étude et le but poursuivi. Dans le cas des analyses de séquences d'états par optimal matching, on s'intéresse à des trajectoires composées d'états et l'on cherche à distinguer des typologies, c'est-à-dire à former des groupes d'individus dont les trajectoires sont similaires. Avec les analyses de survie, c'est principalement un événement qui est au centre de la problématique, et on cherche quels sont les facteurs qui expliquent le mieux l'augmentation du risque que cet événement se produise.

Cette étude mentionne également la distinction entre étudiants internationaux et étudiants étrangers. Elle définit les étudiants internationaux comme étant les étudiants qui ont quitté leurs pays d'origine dans le but de suivre une formation à l'étranger. Par étudiants étrangers, on entend tout étudiant qui n'est pas ressortissant du pays où il suit une formation ; mais ils peuvent être résidents de ce pays.

La deuxième étude (Piguet et Ravel, 2002) s'intitule « Les demandeurs d'asile sur le marché du travail suisse 1996-2000 ».

Des méthodes de régressions logistiques sont utilisées pour mesurer l'effet respectif des différentes caractéristiques individuelles des demandeurs d'asile sur l'insertion sur le marché du travail pour établir l'effet d'une caractéristique isolée sur la probabilité d'occupation. L'idée centrale étant de calculer le taux d'activité des différents groupes de personnes en fonction d'un facteur donné tout en maintenant les autres facteurs constants. Cette étude s'intéresse aussi au parcours de vie des demandeurs d'asile au fil du temps car les données à disposition pour l'étude sont des données de panel. Il était donc possible de savoir à quelle date une personne est entrée sur le marché du travail, combien de temps elle a été occupée ou de savoir si d'autres caractéristiques se sont modifiées, par exemple, l'état civil ou le statut du séjour. Les données de panel offraient donc la possibilité de décrire la dynamique d'ensemble de la population de demandeurs d'asile ainsi que celle du marché du travail.

La troisième étude s'intitule « Entre contraintes institutionnelle et domestique : les parcours de vie masculin et féminin en Suisse »(Levy et al., 2006).

Dans cette étude, la trajectoire de chaque individu y est décrite par une séquence d'états dont la durée est exprimée en année. La méthode de l'appariement optimal fournit une matrice de distances entre les séquences individuelles, comparées une à une moyennant l'utilisation de « coûts ». Ces distances sont ensuite soumises

à une analyse en clusters qui permet d'identifier d'éventuels regroupements de séquences qui indiquent l'existence de types de trajectoires. Pour expliquer l'émergence de parcours de vie distincts, une série de régressions logistiques a été réalisée en utilisant les types de parcours de vie comme variables dépendantes et les variables structurelles comme variables indépendantes.

La quatrième étude s'intitule « Swiss-Swedish Joint Study on Cohort-Based Asylum Statistics » (European Commission, 1998).

Le but de cette étude est de parvenir à une description comparative des concepts clés et des définitions de base relatives aux demandeurs d'asile turcs et somaliens en Suède et en Suisse, de décrire la structure, le contenu et les limites des données des registres sur l'asile. Eurostat voulait mettre à la disposition des pays européens une méthode d'analyse de données pour que ces pays puissent appliquer ces méthodes sur leurs propres données en matière d'asile. Les bases de données existantes sur l'asile en Suède et en Suisse ont été analysées en utilisant une approche longitudinale. Afin d'étudier l'aspect temporel inhérent à chaque processus d'asile, des biographies d'asile ont été établies. Une biographie d'asile est caractérisée par une séquence unique d'événements consécutifs dans le processus d'asile de chaque personne. Trois types de variables sont utilisées dans cette étude : des variables d'identification (numéro d'identification personnel, le numéro de dossier), des variables « géographiques-démographiques » et des variables « événements de l'asile ». Les variables « géographiques-démographiques » contiennent des informations sur le genre, la date de naissance et la nationalité. Les variables « événements de l'asile » contiennent des informations sur la date de la demande d'asile, le lieu de la demande et les dates de la prise de décision des autorités en ce qui concerne la procédure et l'issue de la demande d'asile. L'étude mentionne aussi l'autonomie des pays en matière de traitement des demandes d'asile. Cette autonomie entraîne des différences significatives entre les pays européens sur la façon de demander l'asile, sur les possibilités de recours, et sur la durée de la procédure. Lors du traitement de données nationales sur l'asile, l'étude recommande de tenir compte de ces différences entre pays. Dans le but de rendre la comparaison possible sur le plan international, une harmonisation du cadre sur les concepts, les définitions et la manière de récolter les données est nécessaire.

D'un point de vue méthodologique, l'utilisation de données longitudinales du type données de panel impose de prendre en compte le fait que chaque individu est observé à plusieurs reprises et que, par conséquent, les observations d'un individu ne sont pas indépendantes entre elles (Mueller, 2011).

D'autres études beaucoup plus proches de notre thématique ont été réalisées dans d'autres pays européens ainsi qu'aux Etats-Unis. Certaines de ces études traitent de l'intégration des étudiants internationaux dans le marché de l'emploi du pays d'accueil, tandis que d'autres, traitent du taux de prorogation des permis de séjour après les études et des facteurs qui poussent ces étudiants à rester dans le pays d'accueil. La recherche de Mosneaga et de Winther (Mosneaga et Winther, 2013) menée au Danemark a permis de montrer que pour les étudiants issus des pays non européens, la principale motivation qui pousse cette catégorie d'étudiants à rester au Danemark après l'obtention du diplôme réside dans les opportunités de carrière. D'autres études ont mentionné comme facteurs de rétention les troubles politiques et l'instabilité économique dans le pays d'origine (Tansel et Güngör, 2003). Quant aux motivations de retourner dans le pays d'origine après les études, les raisons familiales sont le plus souvent invoquées par les étudiants (Sykes et Ni Chaoimh, 2013).

L'étude de Bijwaard, et Wang (Bijwaard et Wang, 2016) est très intéressante et se rapproche le plus de notre thématique dans la mesure où elle utilise des données administratives provenant du registre central de l'immigration des Pays-Bas combinées avec celles du service de naturalisation. Cette étude aborde l'importance du mariage et de l'insertion dans le marché de l'emploi des étudiants internationaux sur leur décision de partir après les études. D'un point de vue méthodologique, l'estimateur de Kaplan-Meier est utilisé pour déterminer la probabilité de retour, de mariage ainsi que celle de trouver un emploi selon le lieu d'origine des étudiants.

D'autres études ont été réalisées en Finlande ainsi qu'aux Etats-Unis et s'intéressaient aux taux de prorogation des permis de séjour après les études. La méthodologie utilisée dans ces études consiste à calculer la proportion d'étudiants internationaux qui restent en Finlande et aux Etats-Unis cinq ans après l'obtention du diplôme.

L'étude menée en Finlande par le Centre International de la Mobilité (CIMO) (CIMO, 2016) a permis de montrer que la proportion d'étudiants internationaux toujours enregistrés en Finlande et actif professionnellement cinq ans après les études est de 34%. La même étude a permis de montrer que les étudiants africains sont ceux

qui sont les plus actifs professionnellement après les études avec une proportion d'actifs de 64% seulement un an après les études. C'est presque le double de la proportion des étudiants internationaux actifs cinq ans après les études. Les données mobilisées pour cette recherche sont aussi des données administratives provenant de plus de quarante registres administratifs différents, ainsi que du registre de la population, les registres des entreprises, le registre de l'administration fiscale et celui des étudiants.

Une étude similaire a été réalisée aux Pays-Bas dans le but de déterminer la proportion d'étudiants internationaux qui restent aux Pays-Bas cinq ans après l'obtention du diplôme (NUFFIC, 2016). Dans cette étude, les cohortes des diplômés en 2007, 2008 et 2009 ont été suivies un an, trois ans, cinq ans et sept ans après l'obtention du diplôme. Chaque cohorte contenait environ 12'000 étudiants internationaux diplômés et les résultats obtenus dans cette étude montrent qu'il y a entre 36% et 42% des étudiants internationaux qui sont restés aux Pays-Bas cinq ans après l'obtention du diplôme. L'étude mentionne aussi le fait que les étudiants issus des pays non membres de l'espace économique européen et qui sont inscrits dans une université dans les filières des sciences appliquées, en ingénierie, en sciences ou dans les branches de la santé avaient une proportion de diplômés qui restaient aux Pays-Bas cinq ans après les études plus élevée que la moyenne. L'étude mentionne aussi un taux de participation dans les différentes cohortes qui est supérieur à 75%.

Aux Etats-Unis, Finn et Pennington (Finn et Pennington, 2003) ont réalisé une étude portant sur la proportion des doctorants qui restent aux USA cinq ans après l'obtention du diplôme, selon la filière dans laquelle le doctorat a été obtenu. Il ressort de cette étude qu'en 2005, 68% des titulaires d'un doctorat dans le domaine des sciences et de l'ingénierie étaient encore présents aux USA cinq ans après l'obtention du diplôme, cette proportion se situait à 34% pour les titulaires d'un doctorat en sciences économiques. Les données utilisées dans ces études sont des données administratives provenant de l'administration fiscale et des données de sondage (National Science Foundation's Survey of Earned Doctorates) et des données de l'administration de la sécurité sociale (SSA).

Certaines études se sont intéressées aux facteurs poussant les étudiants à rester ou à quitter le pays d'accueil après les études. D'autres études se sont focalisées sur les avantages de l'accueil des étudiants internationaux en termes d'enrichissement culturel pour les pays d'accueil (Andrade, 2006; Olivas, 2002; Aslanbeigui, 1998).

L'étude qui se rapproche de plus de notre recherche est celle de Dongbin (Kim et al., 2011) qui analyse les tendances des décisions de rester aux Etats-Unis des doctorants internationaux après l'obtention du diplôme de doctorat tout en dégageant les facteurs « push » et « pull » . D'un point de vue méthodologique, l'étude s'intéresse aux variables qui influencent le plus la décision des étudiants internationaux de rester aux Etats-Unis après l'obtention du diplôme de doctorat, selon la filière d'étude et la nationalité. Une régression logistique a été réalisée dans le but de déterminer les meilleurs prédicteurs de la décision de rester aux USA après le doctorat. Les données utilisées dans cette recherche sont des données d'enquêtes. L'enquête est réalisée annuellement par le NSF (The National Science Foundation) dans toutes les institutions d'enseignement supérieur accréditées aux USA et elle concerne toutes les personnes ayant obtenu un doctorat dans une de ces institutions. Ces données sont complétées par celles de l'enquête SED (The Survey of Earned Doctorates) réalisée depuis 1957.

Après cette revue de la littérature sur les étudiants internationaux que nous avons explorée, nous pouvons maintenant aborder sommairement la politique migratoire de la Suisse avant d'introduire les objectifs visés par cette thèse. L'intérêt d'aborder cette politique migratoire réside dans le fait que les étudiants issus de la communauté extra-européenne ne bénéficient pas des mêmes conditions d'accueil et de séjour que les étudiants issus des pays de l'Union Européenne.

En effet, les étudiants internationaux et étrangers non européens étaient soumis, à l'image des autres étrangers en Suisse, à la Loi fédérale du 26 mars 1931 sur le séjour et l'établissement des étrangers (LSEE). Cette loi régissait la politique d'admission des étrangers et laissait le soin aux autorités de déterminer le nombre d'autorisations de séjour à accorder, y compris pour les études selon un certain nombre de critères tels que la part de la population étrangère résidante (Bolzman et Guissé, 2015). L'ordonnance sur la limitation du nombre d'étrangers (OLE) qui vise à maintenir un équilibre entre la population suisse et la population étrangère résidente va aussi avoir un impact sur l'admission et l'accueil des étudiants issus de la communauté non européenne. C'est à partir de 2008 qu'ils ont été soumis à la loi sur les étrangers (LEtr) qui règle les entrées et les sorties de Suisse ainsi que le séjour et le regroupement familial des étrangers. Les étudiants internationaux et étrangers

des états tiers désirant se former dans les établissements d'enseignement supérieur suisses doivent signer un papier certifiant qu'ils vont quitter la Suisse à la fin de leurs études. Le départ de la Suisse à la fin des études était systématique, ce qui n'offrait quasiment aucune chance aux étudiants internationaux non européens, y compris les africains, de prolonger le séjour en Suisse après leurs études. Le caractère temporaire de leur permis de séjour (permis B pour études) renouvelable chaque année et sous réserve [5] rappelle constamment le caractère temporaire du séjour. Dès lors, on peut se poser la question suivante : qu'est ce qui permettait aux étudiants africains et aux étudiants internationaux non européens de prolonger le séjour en Suisse après leurs études ?

Comme le mentionnent Guissé et Bolzman (Bolzman et Guissé, 2015), les étudiants internationaux y compris les africains n'avaient que deux possibilités pour rester en Suisse : le mariage ou la maladie. Soit ils se mariaient à une personne de nationalité suisse, soit à un ressortissant de l'union européenne vivant en Suisse ou à un étranger établi en Suisse. Dans ce cas, ils changeaient de statut et pouvaient prolonger le séjour en Suisse après les études. S'ils souffraient d'une maladie grave qui pouvait mettre leur vie en danger en cas de retour dans le pays d'origine, ils pouvaient aussi rester en Suisse pour pouvoir bénéficier des soins appropriés. Il a fallu attendre l'année 2011 pour voir une autre possibilité de prolonger le séjour en Suisse après les études s'offrir aux étudiants internationaux non européens suite à l'initiative dite de Neirynck. Cette modification de la loi sur les étrangers est le fruit de l'initiative du 19 mars 2008 (entrée en vigueur le 1er janvier 2011) du conseiller national vaudois Jacques Neirynck visant à accorder un permis de six mois [6] aux diplômés des hautes écoles suisses issus d'états tiers dans le but de chercher un travail.

A partir de 2011, les étudiants ressortissants d'états tiers ont donc théoriquement trois possibilités de rester en Suisse qui sont : le mariage, la maladie ou trouver un travail dans un délai de six mois après l'obtention du diplôme. Guissé et Bolzman (Bolzman et Guissé, 2015) mentionnent le fait qu'il ne soit presque pas possible de trouver un travail dans un délai de six mois dans la mesure où même les personnes ayant fait toute leur scolarité en Suisse, y compris les Suisses peinent à trouver le premier emploi dans un délai de six mois. Cette hypothèse semble se vérifier car en analysant les données du Secrétariat d'état aux migrations (SEM) concernant les personnes qui ont pu obtenir un permis de travail après les études, on remarque un faible effectif des personnes qui en ont bénéficié comme le montre la figure 1.1.

On peut dès lors se poser la question de savoir quel est l'impact du mariage sur le fait de rester ou de quitter la Suisse après leurs études pour les étudiants des états tiers ? La politique migratoire suisse étant restrictive, les étudiants internationaux issus de l'Union Européenne et des pays de l'AELE ont plus de chance de prolonger le séjour en Suisse après les études que les étudiants issus des états tiers [7] en raison de l'existence d'accords entre les pays européens. La nationalité est donc une variable importante dans le fait de pouvoir prolonger le séjour en Suisse après les études et cela indépendamment du sexe dans la mesure ou la loi sur les étrangers s'applique indifféremment aux hommes et aux femmes. Les universités suisses sont réputées pour la qualité de leur enseignement et par les centres de recherche qu'elles représentent. Ce qui nous amène également à poser la question du rôle des universités fréquentées sur la probabilité de prolonger le séjour en Suisse après les études.

A partir de ce qui précède, nous voulons répondre à nos deux questions de recherche ; la première concernant le parcours de vie des étudiants internationaux en Suisse de manière générale et la deuxième ne concernant que le parcours de vie des étudiants africains en Suisse.

Le choix de la période allant de 1997 à 2009 pour la première question de recherche s'explique d'une part, par la législation suisse en matière d'accueil des étrangers en vigueur jusqu'en 2011, d'autre part, par une question de données sur les étudiants internationaux en Suisse que nous aborderons antérieurement. En effet, les étudiants internationaux entrés en Suisse durant cette période n'avaient principalement que le mariage comme possibilité, le motif médical n'est pas considéré dans cette étude.

La variable continent d'origine des étudiants nous permettra de comprendre le rôle de la nationalité sur la probabilité de prolonger le séjour en Suisse après les études.

5. Pas de changement du plan d'étude initial (en cas de changement, il faudra prouver le lien entre cette nouvelle formation et l'ancienne : l'étudiant doit respecter le délai des études et doit être inscrit dans une institution d'enseignement supérieur

6. Ce permis de six mois commence à partir de la date de validation des derniers examens et non à partir de la date d'expiration du dernier permis de séjour délivré

7. Par Etat tiers, on entend les pays européens et non européens qui ne font pas partie des accords sur la libre circulation des personnes.

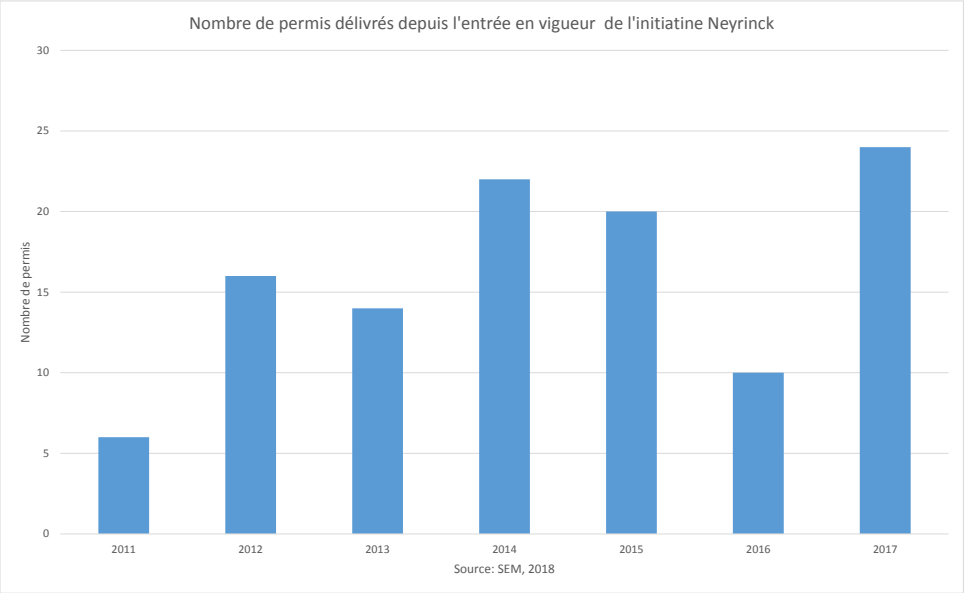


FIGURE 1.1 – Nombre de permis délivrés depuis l'entrée en vigueur de l'initiative Neirynck jusqu'en 2017

Pour pouvoir répondre à ces questions de recherche, nous avons besoin de méthodes statistiques et de données ; la deuxième et la troisième partie de cette thèse seront respectivement consacrées aux méthodes statistiques et aux différents types de données. Pour illustrer la nécessité d'avoir des méthodes statistiques pour répondre à nos questions de recherche, nous allons reprendre notre exemple de départ sur les quatre étudiants africains de nationalités sénégalaise (Abdoul), camerounaise (Michelle), marocaine (Khaled) et Guinéenne (Sadio). Dans l'exemple sur ces étudiants, nous pouvons mettre en évidence des variables administratives, démographiques, académiques et migratoires.

La variable administrative qui est représentée par le numéro AVS identifie les individus de manière unique (un numéro AVS par personne). Les variables démographiques sont : le sexe, le pays de naissance, l'âge (qui peut être calculé connaissant la date de naissance), l'état civil et la nationalité.

Les variables académiques sont : l'année d'obtention du baccalauréat, le pays dans lequel il a été obtenu, la filière dans laquelle il a été obtenu, les universités fréquentées, les filières d'études et l'année de début des études en Suisse. Le nombre de variables académiques augmente pour chaque étudiant à chaque fois qu'un événement se produit alors que le nombre de lignes par étudiant augmente selon la durée de séjour. Par exemple, les variables suivantes peuvent s'ajouter aux variables académiques initiales : l'année d'obtention du bachelor, la filière dans laquelle il a été obtenu, l'institution dans laquelle il a été obtenu, l'année d'obtention du master, le domaine dans lequel il a été obtenu, l'institution dans laquelle il a été obtenu. Nous entendons par événement dans cet exemple, l'obtention d'un diplôme quelconque (bachelor, master, doctorat). Comme variables de migrations, on peut citer l'année d'arrivée en Suisse, le canton de résidence et la commune de résidence.

Les informations sur ces quatre étudiants peuvent donc être résumées en lignes et en colonnes. Les colonnes représenteront les variables et les lignes les individus (de manière indirecte la durée de séjour car chaque ligne représente une année). La dimension longitudinale des données réside dans le fait que ces informations sont enregistrées de manière répétée sur plusieurs années pour les mêmes individus : c'est ce qui nous offre la possibilité d'observer des changements dans le parcours de vie de ces étudiants. Ces changements peuvent être observés au niveau démographique (changement d'état civil, changement de nationalité), au niveau académique (obtention d'un diplôme, changement de filière d'études ou d'université), au niveau migratoire (changement de

canton, de commune, départ de la Suisse).

Pour pouvoir analyser ces données longitudinales, nous avons besoin de modèles statistiques qui puissent prendre en compte la dimension temporelle des données ainsi que des observations incomplètes. Pour comprendre la notion d'observations incomplètes, considérons le cas de Michelle, l'étudiante camérounaise de notre exemple de départ. Supposons qu'on s'intéresse à l'obtention du diplôme de bachelor par Michelle entre l'année d'arrivée en Suisse en 2018 et trois ans après son arrivée soit en 2021. Si Michelle n'obtient pas son diplôme de bachelor entre 2018 et 2021 ou si on la perd de vue avant qu'elle obtienne son diplôme de master dans la période définie, on parlera d'observations incomplètes.

Par ailleurs, à partir des exemples sur ces étudiants africains (Abdoul, Michelle, Khaled et Sadio), nous mettons en évidence le fait que les théories des parcours de vie mobilisées dans cette recherche s'inscrivent dans quelques concepts du paradigme du parcours de vie d'Elder (Elder et Steinmetz, 1976; Elder et Kenneth, 1987; Elder, 1998; H., 1994), et plus précisément à la temporalité des événements et au principe des vies liées. La temporalité des événements fait référence à l'importance du temps auquel un événement se produit dans la vie d'une personne ainsi que les potentiels effets que cet événement peut avoir sur la trajectoire de vie de la personne concernée. Cette prise en compte de la temporalité des événements suppose que l'on ait à disposition des données récoltées au cours du temps. Le principe des vies liées met l'accent sur le fait que le parcours de vie d'une personne n'est pas indépendant de celui des autres personnes qui l'entourent ou de celui des individus que cette personne rencontre dans sa vie. Dans le cadre de cette recherche, le principe de vies liées est tenu en compte à travers le mariage dans la mesure où par exemple, notre étudiante Michelle qui est venue en Suisse pour étudier peut voir son parcours de vie influencé par une rencontre durant les études qui va aboutir à un mariage, puis à la naissance des enfants, et enfin à un établissement en Suisse. Le parcours de vie de Michelle ne doit ainsi pas être considéré de manière indépendante de celui de son époux.

Pour étudier le parcours de vie des ces étudiants internationaux et africains dans le but de répondre à nos questions de recherche, nous avons donc besoin de modèles statistiques et de données longitudinales. Les modèles statistiques doivent pouvoir tenir compte à la fois du temps (l'aspect longitudinal des données) ainsi que des observations incomplètes.

Nous devons cependant mentionner le fait qu'il existe plusieurs modèles pour analyser des données longitudinales dans une perspective de parcours de vie et que ces modèles prennent des appelations différentes selon les domaines d'études. Le domaine médical étant celui dans lequel ces modèles sont le plus répandus.

Pour illustrer la large diffusion de ces méthodes dans le domaine médical, on peut citer quelques études dont celle de Rudolph (Rudolph et al., 2018) qui s'intéresse à la durée qui s'écoule jusqu'au diagnostic du sida ou jusqu'au décès, dans une cohorte de femmes séropositives, en comparant des modèles non paramétriques et paramétriques. Cette étude fait ressortir le fait que lorsque le vrai modèle paramétrique est connu, alors ce dernier offre une meilleure estimation des paramètres que les autres modèles paramétriques ou non paramétriques. L'étude de Kargarian-Marvasti (Kargarian-Marvasti et al., 2017) s'est 'intéressée quant à elle aux facteurs ayant un impact positif sur la durée de la neuropathie chez les patients atteints de diabète de type 2 en comparant le modèle de Cox, aux modèles paramétriques. Cette étude fait ressortir le fait que le meilleur modèle pour analyser le phénomène étudié est le modèle log-normale par rapport aux autres modèles paramétriques ou au modèle de Cox. L'étude de Habibi (Habibi et al., 2018) s'est s'intéresssée à la comparaison de modèles statistiques d'analyse de survie pour analyser les facteurs de pronostics chez les patients atteints de cancers gastriques. Les résultats obtenus dans cette étude montrent que le modèle log-normale est le plus adapté parmi les modèles paramétriques et le modèle de Cox pour analyser les facteurs de pronostics des patients atteints de cancers gastriques. On peut aussi mentionner l'étude de Hoseini (Hoseini et al., 2017) qui compare les modèles de Weibull, log-normale et le modèle de Cox pour analyser la durée de vie de patientes atteintes du cancer du sein au Rafsanjan (Iran). Cette étude a montré que le modèle log-normal était plus adapté pour analyser le phénomène étudié que le modèles de Weibull et de Cox. Une dernière étude dans le domaine de la médecine que nous allons citer est celle de Zare (Zare et al., 2013) qui compare les modèles paramétriques et le modèle de Cox pour étudier la survie de patients atteints de cancers gastriques et qui sont sur le point de subir une intervention chirurgicale à l'Institut iranien du cancer. Cette étude a montré que les modèles log-logistique, de Gompertz et le modèle log-normale étaient plus adaptés que le modèle de Cox pour anaylser la durée de vie de ces patients atteints de cancers gastriques. Ces diffétentes études ont utilisé le critère AIC (Akaike information criterion)

pour comparer les différents modèles dans le but de choisir le meilleur. Ces exemples montrent que ces modèles sont très utilisés dans le domaine médical.

Dans cette thèse, ous aborderons également les différentes sources de données longitudinales en mettant une attention particulière sur les données administratives qui feront l'objet de nos futures analyses. Nous présenterons les modèles les plus couramment utilisés dans l'analyse des parcours de vie ainsi que le type de données adaptées à ces modèles. Tous les modèles qui seront présentés dans cette recherche seront suivis par des exemples d'application dans le but de les illustrer. Nous tenons aussi à préciser qu'en raison du temps très long qu'à pris la livraison des données, nous avons jugé nécessaire de prendre des données dans des livres ou sur Internet dans le but de mettre en application les différents modèles. Certaines données, même si elles ne concernent pas la mobilité des étudiants, seront utilisées en recodant des variables ou en créant de nouvelles variables dans le but de ramener la thématique à la mobilité des étudiants. Ces différents exemples développés avant la livraison des données par l'OFS ont été maintenus dans la thèse, ce qui explique que certains exemples ne traitent pas de la mobilités des étudiants. La source des données sera indiquée et les travaux de recodage des variables clairement présentés. Auparavant, une introduction à l'analyse des parcours de vie sera faite dans la section qui va suivre.

Deuxième partie

Méthodes statistiques d'analyse de parcours de vie

Chapitre 2

Introduction à l'analyse des parcours de vie

Pour aider à mieux comprendre la notion de parcours de vie, nous allons reprendre l'exemple de nos quatre étudiants africains, Abdoul, Michelle, Khaled et Sadio qui sont arrivés en Suisse pour poursuivre des études universitaires.

Dans cet exemple, nous avons des étudiants africains qui ont décidé de venir en Suisse pour faire des études universitaires dans quatre hautes écoles universitaires différentes. Le cas de ces étudiants n'est pas unique dans la mesure où il y a des étudiants venant de tous les coins du monde qui ont eu à un moment donné de leurs parcours académiques l'idée d'avoir une expérience internationale en allant étudier hors des frontières nationales en choisissant la Suisse comme destination. En récoltant les informations sur ces étudiants et en les mettant ensemble, on obtient des données (une base de données).

A partir des informations recueillies sur nos quatre étudiants internationaux, on peut constituer une base de données qui prendra la structure suivante :

Tableau 2.1 – Exemple de construction d'une base de données

Etudiant	AVS	Entrée	Né en	Sexe	Nationalité	Etat civil	Uni	Filière
Abdoul	756.xx.	2015	1998	M	Sénégal	Célibataire	Unige	Economie
Michelle	756.xx.	2018	1996	F	Cameroun	Célibataire	Unine	Lettres
Khaled	756.xx.	2016	1990	M	Maroc	Célibataire	Unil	Informatique
Sadio	756.xx.	2006	1986	M	Guinée	Célibataire	EPFL	Mathématiques

Dans cette base de données, nous avons mentionné les prénoms des étudiants dans le but de faciliter la lecture de celle-ci en reconnaissant les caractéristiques individuelles de chaque personne comme décrit dans la courte biographie de ces quatre étudiants. Les prénoms et noms des étudiants ne doivent normalement pas figurer dans la base de données tout comme les vrais numéros AVS des étudiants pour une question de confidentialité. Cette notion de confidentialité sera abordée ultérieurement dans la partie consacrée aux données (chapitre 13). Les informations contenues dans cette base de données (tableau 2.1) sont seulement celles de la première année après l'arrivée en Suisse. Si ces étudiants sont toujours en Suisse cinq ans après, nous aurons une base de données de cinq lignes par personne (soit 15 lignes au total) avec les mêmes caractéristiques individuelles sauf s'il y a changement au niveau d'une caractéristique individuelle pour une personne donnée. Cette notion de changement qui porte le nom d'événement est essentielle dans l'analyse des parcours de vie d'une part, d'autre part, le fait que ces étudiants soient suivis pendant plusieurs années introduit la dimension temporelle des données. Cet exemple fait donc ressortir deux éléments importants dans l'analyse des parcours de vie : les notions d'événement et de temps.

Supposons maintenant que nous ayons une très grande base de données à l'image de celle que nous avons contruite avec nos étudiants Abdoul, Michelle, Khaled et Sadio et qui contient plusieurs étudiants internationaux sur plusieurs années.

L'analyse des parcours de vie consiste à s'intéresser à l'occurrence d'un ou de plusieurs événements d'intérêt au cours du temps ainsi qu'à la chronologie de ces événements : le mariage, le décès ou l'obtention d'un diplôme

par exemple. Cela suppose donc de disposer d'un instant de départ qui correspond au début d'observation et d'un instant de fin qui correspond à la fin d'observation ; on s'intéressera à la distribution de l'occurrence de l'événement d'intérêt entre le début d'observation et la fin d'observation en ciblant l'instant précis auquel les individus soumis à observation ont connu l'événement d'intérêt. Dans l'analyse des parcours de vie, nous sommes confrontés à l'existence d'observations incomplètes causées par le fait que tous les individus ne connaîtront pas l'événement étudié dans la mesure où certaines personnes sortiront de l'observation avant la fin de l'étude sans avoir connu l'événement d'intérêt et d'autres ne connaîtront pas l'événement d'intérêt jusqu'à la fin de la période d'observation. La variable d'intérêt est donc le temps ou en d'autres termes, la durée qui s'écoule entre le début de l'observation et la survenue de l'événement étudié. Il peut s'agir par exemple du temps qui s'écoule entre l'arrivée d'un étudiant international en Suisse et la date de son mariage ou la date d'obtention de son premier diplôme, du temps qui s'écoule entre la date d'un mariage et la survenue du divorce ou le temps qui s'écoule entre le mariage et la naissance du premier enfant.

Considérons l'exemple fictif suivant (tableau 2.2), concernant des étudiants internationaux soumis à des examens finaux pour l'obtention du diplôme de bachelor dans une université donnée. L'événement d'intérêt est l'obtention du diplôme de bachelor ; le tableau ci-dessous résume les données.

Tableau 2.2 – Exemple de données

Année	1	2	3	4	5	6	7	8	9	10
Nombre de diplômes	8	6	12	15	9	13	18	9	10	20

Dans cet exemple, on observe que la période d'observation est de dix ans (première ligne du tableau). Dans la première année, huit étudiants ont obtenu le diplôme de bachelor, dans la troisième année, douze étudiants l'ont obtenu et dans la dixième année, vingt étudiants ont obtenu leurs diplômes de bachelor. Dans cet exemple, notre variable d'intérêt c'est le temps (la durée jusqu'à l'obtention du diplôme). Cette variable temps est une variable discrète [1] pouvant présenter des informations incomplètes lorsque tous les individus soumis à observation ne connaissent pas l'événement étudié avant la fin de l'étude. On parle alors d'observations incomplètes ou de censures (nous privilégierons cette appellation qui est la plus répandue). Une donnée est dite censurée lorsque nous ne disposons que d'une information partielle la concernant ; les données pouvant être censurées à gauche ou à droite.

Une donnée est dite censurée à gauche lorsque nous ne disposons d'aucune information antérieure à la date de début de l'étude pour un ou plusieurs individus. Dans le cadre de la mobilité des étudiants, une donnée censurée à gauche est le cas d'un étudiant international dont on ne connaît pas le parcours migratoire ou académique avant son arrivée en Suisse. La seule façon d'avoir des informations sur son parcours migratoire antérieur serait de procéder à une enquête rétrospective. Il est cependant important de préciser que dans le cadre de cette recherche, il n'existe pas de données censurées à gauche dans la mesure où le point de départ de l'étude c'est la date d'entrée en Suisse dans le but de poursuivre des études supérieures. Le parcours des étudiants avant leur arrivée en Suisse ne nous intéresse pas ; dès lors, il n'y a pas de censure à gauche.

Une donnée sera dite censurée à droite lorsque nous ne disposons que d'une information incomplète concernant un individu parce que ce dernier est sorti de l'étude sans connaître l'événement étudié ou qu'on a perdu sa trace sans qu'il n'ait connu l'événement étudié. Dans le cadre de la mobilité des étudiants, si on s'intéresse à la survenue du mariage d'un étudiant international entre le début de son séjour en Suisse (t_0) et t_5 (5 ans après son arrivée en Suisse), on parlera de censure à droite lorsque cet étudiant ne s'est pas marié après cinq ans de séjour en Suisse (après t_5) ou lorsque l'étudiant disparaît de l'étude avant t_5 sans être marié au préalable.

La censure peut aussi être classée en deux types : la censure fixe et la censure aléatoire (Hill et al., 1991). La censure fixe regroupe tous les individus qui n'ont pas connu l'événement étudié pendant la durée d'observation

1. En effet, l'obtention du diplôme est comptabilisée à la fin de l'année académique. La durée de l'année académique est définie entre le 30 septembre et le 29 août. Cela veut dire qu'elle ne couvre pas tous les mois de l'année mais seulement une certaine période de temps durant l'année et par conséquent la variable temps ne peut pas être considérée comme un continuum (ce qui correspondrait à une variable aléatoire continue). Notre variable temps correspond donc à une variable aléatoire discrète.

qui peut être exprimée en jours, mois, semaine ou année (le temps de censure est le même pour tous). Par censure aléatoire [2], on entend tous les individus qui sortent d'observation avant la fin de l'étude sans connaître l'événement étudié. Ces individus ont des temps de censure différents et propres à chacun d'eux, ces temps de censures étant différents de la durée de l'étude elle-même. Dans l'exemple cité ci-dessus concernant la durée qui s'écoule depuis l'arrivée d'un étudiant international en Suisse et son mariage après cinq ans de séjour, tous les étudiants qui ne se sont pas mariés après cinq ans de séjour ont des durées de censure fixes et tous ceux qui sont sortis de l'étude avant cinq ans sans être mariés auront des temps de censures aléatoires. Par sortir de l'étude, on sous-entend des étudiants qui ont quitté la Suisse avant la fin des études, donc sans obtenir un diplôme, ou des étudiants qui disparaissent des régistres.

Pour illustrer cette notion de censure, considérons les étudiants internationaux suivants : Paul, André, Martine, Vanessa, Olivier, Tim et Sandra, comme présenté dans le graphique 2.1.

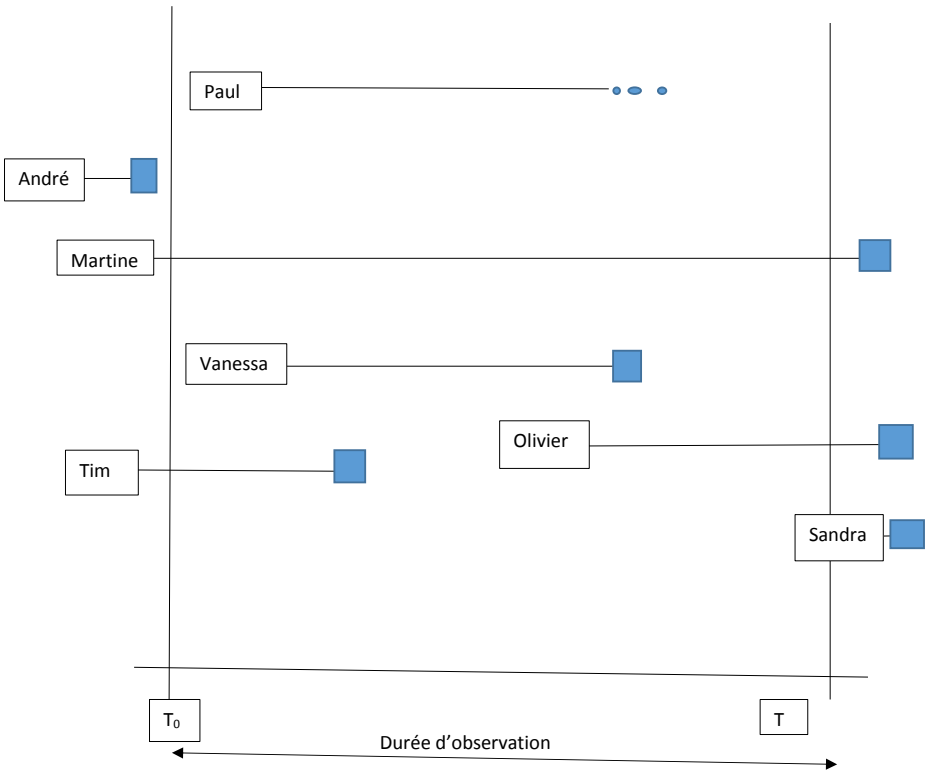


FIGURE 2.1 – Graphique illustrant les différents cas de censures

Paul a commencé ses études durant la période d'observation et a disparu de l'étude durant cette même période d'observation ; c'est le cas d'un étudiant qui a interrompu ses études.

André est censuré à gauche, cela veut dire qu'on ne connait ni la date de début de ses études ni la fin de ces dernières. On connait seulement que ces dates sont situées avant le début de la période d'observation.

Martine a commencé ses études avant le début de la période d'observation et ne les a pas terminées durant la période d'observation. Le parcours de cette étudiante est censurée à gauche et à droite.

2. Par censure aléatoire, on sous-entend le fait que les sorties d'observation des individus sont indépendantes.

Vanessa a commencé ses études durant la période d'observation et les a terminées durant cette période.

Olivier a commencé ses études durant la période d'observation et il a terminé ses études après cette période d'observation ; son parcours est censuré à droite.

Tim a commencé ses études avant la période d'observation et les a terminées durant la période d'observation. Son parcours étudiant est censuré à gauche.

Le parcours de Sandra est censurée à droite. La date de début de ses études se situe en dehors de la période d'observation.

Pour faire une analyse de parcours de vie, nous avons besoin d'une population qui est soumise à l'étude, de la durée jusqu'à l'événement étudié (variable temps) et de la durée écoulée jusqu'à la censure (variable censure).

Les méthodes d'analyses de données de durée (données de survie) ont été développées au départ pour mesurer la fiabilité et la longévité de matériel électriques et électroniques (Dupont, 2006) sous l'appelation d'analyse de fiabilité. Dans ce cas, l'événement étudié est la survenance d'une panne lorsque ce sont des machines qui sont testées (Ray, 1988) ou tout simplement la fin de vie lorsqu'il s'agit d'une ampoule par exemple. On s'intéresse à la survenance d'une panne depuis la mise en fonction d'une machine jusqu'à une date donnée qui est fixée à l'avance. Bien qu'il s'agisse de machine, on parle également de censure dans la mesure où les machines qui ne connaitront pas de panne jusqu'à la fin de la période d'observation seront censurées à droite.

Ces méthodes ont été importées dans le domaine de la médecine sous l'appellation *d'analyse de survie* et dans la démographie sous l'appellation *d'analyse de biographie* (Hill et al., 1991). Dans le domaine de la médecine, l'intérêt porte sur la durée de survie d'un patient après un traitement donné (opération chirurgicale, traitement d'un cancer). L'événement d'intérêt est généralement le décès. On parlera de censure lorsque l'on n'observe pas de décès chez certaines personnes après la période d'observation (Ray, 1988). Dans le domaine démographique, on peut s'intéresser à divers événements comme par exemple, le mariage, la naissance du premier enfant, le divorce.

Ces méthodes, quels que soient leurs domaines d'application utilisent les mêmes techniques statistiques et les mêmes fonctions pour décrire la durée. Pour pouvoir faire une analyse de survie, on doit connaître les informations suivantes sur chaque individu soumis à observation : la date du début d'observation (date à laquelle l'individu a été placé sous observation), la date de fin d'observation, on s'intéressera ensuite aux dates d'occurrence de l'événement d'intérêt au cours du temps. A partir de ces informations, on peut déterminer la durée d'observation, le temps qui s'est écoulé avant que l'individu ne connaisse l'événement étudié, les individus qui sont sortis de l'étude sans connaître l'événement étudié, puis déterminer la distribution de la durée de survie et faire des comparaisons entre groupes.

Dans le but d'analyser le temps de survie (durée de séjour), les modèles statistiques d'analyse de parcours de vie utilisent plusieurs fonctions qui sont reliées entre elles. Ces fonctions dépendent toutes du temps et seront présentées dans les sections suivantes à travers de cas concrets. La relation entre ces fonctions est valable pour tous les modèles qui seront présentés.

Les modèles d'analyse de parcours de vie se regroupent en trois types de modèles qui sont : les modèles non-paramétriques, les modèles paramétriques et les modèle semi-paramétriques (Cox). Dans les modèles non paramétriques, les méthodes de Kaplan-Meier et actuarielle seront présentées. Dans les modèles paramétriques, les distributions exponentielle, de Weibull, de Gompertz, lognormale, loglogistique et gamma seront présentées. Le modèle de Cox (modèle semi-paramétrique) sera aussi présenté dans le cadre de cette partie consacrée aux modèles d'analyse des parcours de vie. Les modèles non paramétriques sont des modèles descriptifs alors que les modèles paramétriques et le modèle de Cox sont des modèles explicatifs (ce sont des modèles de régression). Le choix de ne présenter que ces méthodes s'explique pour deux raisons. La première raison est le fait que dans la littérature, ces méthodes apparaissent comme celles qui sont les plus couramment utilisées dans l'analyse des parcours de vie. La deuxième raison réside dans la nature de nos questions de recherche qui ont été présentées dans la première partie. Ces questions de recherche montrent que nous avons besoin de modèles explicatifs qui puissent tenir compte du temps tout en mettant en relation une variable dépendante avec plusieurs variables explicatives.

Avant de détailler les différentes fonctions à la base de l'analyse des parcours de vie, nous allons d'abord présenter les caractéristiques de la variable temps.

2.1 Variable durée ou variable temps

Considérons une population au sein de laquelle les individus sont soumis au même risque [3] de connaître un événement donné au temps T. Le temps T est une variable positive dont la distribution est au centre de l'analyse des parcours de vie. Sa loi de probabilité peut être déterminée selon un certain nombre de fonctions.

Ces fonctions peuvent être représentées en temps continu ou en temps discret. Les fonctions de survie, de densité, de répartition, de risque et de risque cumulé permettent de décrire la distribution de la variable temps et d'étudier les parcours de vie que ce soit dans les modèles non-paramétriques, ou paramétriques ou dans le modèle semi-paramétrique de Cox. Ces fonctions s'écrivent différemment selon le type de modèle ciblé.

Avant d'introduire ces différentes fonctions, on va reprendre l'exemple de nos quatre étudiants africains (Abdoul, Michelle, Khaled et Sadio).

Si l'on s'intéresse à la probabilité que Michelle obtienne son diplôme de bachelor à chaque instant t, on utilisera la fonction de densité ou fonction de probabilité pour calculer cette probabilité. Cette fonction de densité nous permet de connaître cette probabilité pour chaque instant t donné.

Si l'on s'intéresse à la probabilité que Michelle obtienne son diplôme de bachelor avant une année donnée, on utilisera une fonction statistique qui s'appelle la fonction de répartition. Cette fonction de répartition nous permettra, par exemple, de calculer la probabilité que la durée des études de bachelor de Michelle soit inférieure à 4 ans.

La fonction de survie ou durée de séjour nous permettra de calculer la probabilité que Michelle obtienne son diplôme de bachelor après un certain nombre d'années d'études. Avec cette fonction, on peut par exemple calculer la probabililté que Michelle obtienne son diplôme de bachelor après trois ans ou après quatre ans d'études.

Si on s'intéresse par exemple à la proportion d'étudiants africains qui a obtenu le diplôme de bachelor à un instant t donné, on utilisera la fonction de risque. Cette proportion sera calculée en faisant le rapport entre le nombre d'étudiants africains ayant obtenu le diplôme de bachelor et le nombre d'étudiants soumis au risque de l'obtenir. Pour calculer la proportion d'étudiants africains ayant obtenu le diplôme de bachelor entre un instant t et un instant $t+1$, on utilisera la fonction de risque cumulé. Cette fonction de risque cumulé fait la somme des risques aux différents instants entre t et $t+1$ pour les étudiants concernés.

On va à présent montrer comment calculer ces différentes fonctions en développant à chaque fois un exemple avec des données et un exemple littéral.

2.1.1 Distribution de probabilité

Considérons notre exemple du tableau 2.2 sur l'obtention du diplôme de bachelor par des étudiants internationaux dans une Université donnée. La proportion d'étudiants internationaux ayant obtenu le diplôme de bachelor chaque année est calculée dans la troisième ligne du tableau 2.3.

Tableau 2.3 – Exemple de calcul de la proportion d'étudiants ayant obtenu leur diplôme chaque année

Année	1	2	3	4	5	6	7	8	9	10
Nombre de diplômes	8	6	12	15	9	13	18	9	10	20
Proportion	$\frac{8}{120}$	$\frac{6}{120}$	$\frac{12}{120}$	$\frac{15}{120}$	$\frac{9}{120}$	$\frac{13}{120}$	$\frac{18}{120}$	$\frac{9}{120}$	$\frac{10}{120}$	$\frac{20}{120}$

Au total, nous avons 120 étudiants, ce qui correspond au nombre de diplômes décernés. Pour chaque instant t, la proportion d'étudiants ayant obtenu le diplôme de bachelor est égal au rapport entre le nombre de diplômes décernés à chaque instant et le nombre total de diplômes décernés.

3. Le risque auquel nous faisons référence ici n'est pas le risque dans le sens de se mettre en danger, mais c'est le risque en tant que fonction statistique qui dans certains cas peut être interprété comme une probabilité.

En temps discret, la loi de probabilité nous permet de déterminer à chaque instant t la probabilité que l'événement étudié se produise. Lorsque le temps est considéré comme continu, la fonction de densité nous permettra de déterminer la probabilité que l'événement étudié se produise dans chaque intervalle de temps. En temps discret, la distribution de T est décrite par la probabilité p_t de connaître l'événement étudié et lorsque le temps est continu, la distribution de T est définie par la fonction de densité notée $f(t)$.

En temps discret, la distribution de probabilité de la durée T s'écrit comme :

$$p_t = P(T = t)$$

Nous nous contenterons seulement de la représentation en temps discret de la distribution de probabilité. La fonction de densité en temps continu contient certains éléments que nous n'avons pas encore introduits à ce stade.

2.1.2 Fonction de répartition

En considérant à nouveau l'exemple du tableau 2.2, la fonction de répartition notée $F(t)$ ou F_t représente la probabilité d'obtenir son diplôme de bachelor avant ou à l'instant t donné. Par exemple, la probabilité d'obtenir son diplôme de bachelor pour un étudiant international avant le temps $t = 3$ est calculée dans la troisième ligne du Tableau 2.4.

Tableau 2.4 – Exemple de calcul de la fonction de répartition

Année	1	2	3	4	5	6	7	8	9	10
Nombre de diplômes	8	6	12	15	9	13	18	9	10	20
$F(3)$ ou F_3	$\frac{8}{120}+\frac{6}{120}+\frac{12}{120}=\frac{26}{120}$									

On peut donc écrire la probabilité d'obtenir son diplôme de bachelor avant le temps $t = 3$ comme $F(3) = F_3 = p_1 + p_2 + p_3 = \frac{8}{120}+\frac{6}{120}+\frac{12}{120}=\frac{26}{120}$.

La fonction de répartition représente pour un temps t fixé, la probabilité de vivre l'événement étudié avant l'instant t. Cette probabilité représente la proportion d'individus ayant connu l'événement étudié depuis le début de l'étude jusqu'à un instant t donné. En temps continu tout comme en temps discret, il existe un lien entre la fonction de densité ou loi de probabilité et la fonction de répartition (Allison, 2010; Blossfeld et al., 2009).

$$F(t) = P(T \leq t) \tag{2.1}$$

Par exemple, si nous voulons calculer la fonction de répartition entre les temps $t = 0$ et $t = 2$ pour un événement donné, on procédera de la manière suivante :

$F(t) = P(T \leq 2) = P(T \leq 0) + P(T \leq 1) + P(T \leq 2)$

Lorsque le temps est discret, la fonction de répartition s'écrit comme :

$$F(t) = \sum_{t=1}^{n} p_t$$

Le signe $\sum_{t=1}^{n} p_t$ indique juste que nous devons additionner les probabilités entre les instants de début $(t = 1)$ et de fin $(t = n)$ qui nous intéressent. Par exemple, $\sum_{t=1}^{3} p_t = p_1 + p_2 + p_3$.

Pour mieux illustrer le calcul de la fonction de répartition, nous allons développer un exemple de calcul littéral [4]. Considérons le tableau 2.5 ci-dessous en temps discret dans lequel t_1, t_2, t_3... t_{10} représentent les durées, et p_1, p_2, p_3...p_{10} représentant les risques de connaître un événement quelconque aux temps t_1...t_{10}.

Le but est de calculer la valeur de la fonction de répartition notée $F(6)$ au temps $t = 6$. $F(6) = P(t \leq 6)$ se calcule en faisant la somme des probabilités élémentaires de t_1 jusqu'à t_6.

Tableau 2.5 – Exemple littéral de calcul de la fonction de répartition

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
$F(6) = p_1 + p_2 + p_3 + p_4 + p_5 + p_6$									

2.1.3 Fonction de survie ou fonction de séjour

Dans notre cas, la fonction de survie ou fonction de séjour est la fonction qui détermine la durée de séjour. A l'image des autres fonctions déjà présentées, on va reprendre l'exemple précédent et cette fois l'intérêt porte sur la détermination de la durée de séjour notée $S(t)$. La fonction de séjour représente la probabilité de ne pas obtenir son diplôme de bachelor après un certain temps t.

Tableau 2.6 – Exemple de calcul de la fonction de séjour

Année	1	2	3	4	5	6	7	8	9	10
Nombre de diplômes	8	6	12	15	9	13	18	9	10	20
N restant	120	112	106	94	79	70	57	39	30	20
$S(t)$	$\frac{120}{120}$	$\frac{112}{120}$	$\frac{106}{120}$	$\frac{94}{120}$	$\frac{79}{120}$	$\frac{70}{120}$	$\frac{57}{120}$	$\frac{39}{120}$	$\frac{30}{120}$	$\frac{20}{120}$

Dans cet exemple, on observe qu'au temps $t = 1$, on avait 120 étudiants soumis à observation, parmi lesquels 8 étudiants ont obtenu leurs diplômes de bachelor la même année. Au temps $t = 2$, on avait plus que 112 (120-8) étudiants soumis à observation car huit étudiants avaient obtenu leur diplôme au temps $t = 1$.

La fonction de survie notée $S(t)$ qui représente la probabilité de n'avoir pas encore obtenu son diplôme est calculée dans la dernière ligne du tableau 2.6.

Dans cet exemple, on remarque que la survie vaut 1 (120/120) à la première année et celle-ci décroît au fur et à mesure que les années passent pour se situer à $20/120$ à la dixième année.

Lorsque le temps est discret, la fonction de survie s'écrit de la manière suivante :

$$S_t = 1 - F_{t-1}$$

avec F_t représentant la fonction de répartition à l'instant t.

Considérons l'exemple littéral de calcul de la fonction de séjour présenté dans le tableau 2.7 ci-dessous en temps discret et dans lequel t_1, t_2, t_3... t_{10} représentent les durées, et F_1, F_2, F_3...F_{10} représentent les valeurs de la fonction de répartition à chaque instant. Les survies aux instants allant de t_1 à t_6 sont calculées dans la troisième ligne du tableau.

4. Le calcul littétal signifie faire des calculs avec des lettres à la place des chiffres

Tableau 2.7 – Exemple littéral de calcul de la fonction de séjour

t_1	t_2	t_3	t_4	t_5	t_6
F_1	F_2	F_3	F_4	F_5	F_6
$S_1 = 1$	$S_2 = 1 - F_1$	$S_3 = 1 - F_2$	$S_4 = 1 - F_3$	$S_5 = 1 - F_4$	$S_6 = 1 - F_5$

2.1.4 Risque instantané ou taux de hasard

En reprenant l'exemple précédent, la fonction de risque notée $h(t)$ ou h_t représente pour tout instant t, le rapport entre le nombre d'étudiants internationaux ayant obtenu le diplôme de bachelor et le nombre total d'étudiants internationaux soumis aux examens à ce même instant t. Cette fonction de risque est calculée dans le tableau 2.8.

Tableau 2.8 – Exemple de calcul de la fonction de risque

Année	1	2	3	4	5	6	7	8	9	10
Nombre de diplômes	8	6	12	15	9	13	18	9	10	20
N restant	120	112	106	94	79	70	57	39	30	20
$h(t)$ ou h_t	$\frac{8}{120}$	$\frac{6}{112}$	$\frac{12}{106}$	$\frac{15}{94}$	$\frac{9}{79}$	$\frac{13}{70}$	$\frac{18}{57}$	$\frac{9}{39}$	$\frac{10}{30}$	$\frac{20}{20}$

La taux de hasard ou fonction de risque instantané notée $h(t)$ décrit la proportion des individus qui ont connu l'événement étudié au temps t parmi les individus exposés au risque de connaître cet événement à ce même instant t. Dans notre exemple, à la première année, on a huit étudiants qui ont obtenu le diplôme sur un total de 120 étudiants soumis au risque de l'obtenir, ce qui donne un risque $h_1 = \frac{8}{120}$. Dans la deuxième année, on n'a plus que $112(120 - 8)$ étudiants soumis au risque d'obtenir le diplôme et six étudiants l'ont obtenu à la même année, ce qui donne un risque $h_2 = \frac{6}{112}$. Les risques aux autres instants se calculent de la même manière. Cette fonction, à l'image des autres, peut aussi s'écrire en temps continu ou en temps discret.

En temps continu, la fonction de risque s'écrit de la manière suivante (Allison, 2010) :

$$h(t) = \frac{f(t)}{S(t)} \tag{2.2}$$

La fonction de risque est calculée en faisant le rapport entre la fonction de densité et la fonction de séjour.

En temps discret, elle s'écrit comme :

$$h_t = p_t / S_t$$

Dans ce cas, le risque de connaître l'événement étudié au temps t est égal au rapport entre la probabilité de connaître cet événement au temps t et la survie correspondante à ce même instant t. Par exemple, le risque d'obtenir le diplôme de bachelor au temps $t = 1$ a été calculé en faisant le rapport entre la probabilité d'obtenir le diplôme de bachelor au temps $t = 1$ et la survie correspondante à ce même instant $t = 1$.

Le risque de connaître l'événement étudié à chaque instant t est calculé dans la dernière ligne du tableau 2.9.

2.1.5 Taux de risque cumulé ou taux de hasard cumulé

La fonction de risque cumulé notée $H(t)$ décrit le nombre moyen d'événements qui se seraient produits si les individus étudiés étaient soumis au risque de connaître cet événement durant toute la période d'observation. La fonction de risque cumulé peut aussi s'écrire en temps continu ou en temps discret.

Tableau 2.9 – Exemple littéral de calcul de la fonction de risque instantané

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
$h_1 = \frac{p_1}{S_1}$	$h_2 = \frac{p_2}{S_2}$	$h_3 = \frac{p_3}{S_3}$	$h_4 = \frac{p_4}{S_4}$	$h_5 = \frac{p_5}{S_5}$	$h_6 = \frac{p_6}{S_6}$	$h_7 = \frac{p_7}{S_7}$	$h_8 = \frac{p_8}{S_8}$

En temps continu, le taux de hasard cumulé s'écrit comme (Courgeau et Lelièvre, 1989) :

$$H(t) = -ln(S(t)) \tag{2.3}$$

Où $ln(.)$ désigne le logarithme népérien.

Lorsque le temps est discret, la fonction de risque cumulé s'écrit comme étant la somme jusqu'à un instant t donné des risques élémentaires de connaître l'événement étudié :

$$H(t) = \sum_{t=1}^{n} h_t$$

Dans notre exemple, au temps $t = 3$ (à la troisième année), $H(3) = h_1 + h_2 + h_3 = \frac{8}{120} + \frac{6}{112} + \frac{12}{106}$.

Si l'on veut calculer le risque cumulé depuis l'instant de départ jusqu'à un temps t donné, on va additionner les risques élémentaires depuis l'instant de départ jusqu'à cet instant t.

Reprenons l'exemple précédent, cette fois-ci notre objectif est de calculer le risque cumulé jusqu'au temps $t = 8$. Comme le montre le tableau 2.10, ce risque cumulé au temps t_8 noté $H(8)$ est égal à la somme des risques élémentaires h_1 + h_2 + h_3 + ... + h_8.

Tableau 2.10 – Exemple littéral de calcul du risque cumulé

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
$H(8) = h_1 + h_2 + h_3 + h_4 + h_5 + h_6 + h_7 + h_8$							

2.2 Application

Considérons un exemple fictif de 300 étudiants soumis au risque de refaire des examens de rattrapage dans une faculté donnée. Chaque année, le nombre d'étudiants qui refont des examens de rattrapage a été enregistré et on a obtenu le tableau 2.11. Le but de cet exemple est de calculer les différentes fonctions décrites dans les sections précédentes afin d'étudier la distribution dans le temps des examens de rattrapage pour ces 300 étudiants. Dans cet exemple, les données ne sont pas censurées.

Dans ce tableau, les colonnes à disposition sont les trois premières colonnes, à savoir, la colonne de temps « t », la colonne « Nombre d'étudiants » et la colonne « Nombre d'examens ». Les cinq autres colonnes ont été calculées. Les détails de ces calculs seront présentés pour chaque type de fonction dans les sections qui vont suivre. Tous les calculs dans les exemples qui vont suivre seront extraits de ce tableau, le but étant d'expliquer comment les différentes fonctions de ce tableau ont été calculées.

Tableau 2.11 – Exemple d'application

t	Nombre d'étudiants	Nombre d'examens	p_t	F_t	S_t	h_t	H_t
1	300	20	0.067	0.067	1	0.067	0.067
2	280	30	0.1	0.167	0.933	0.107	0.174
3	250	90	0.3	0.467	0.833	0.36	0.534
4	160	14	0.047	0.513	0.533	0.088	0.621
5	146	30	0.1	0.613	0.487	0.205	0.827
6	116	15	0.05	0.663	0.387	0.129	0.956
7	101	35	0.117	0.78	0.337	0.347	2.303
8	66	13	0.043	0.823	0.22	0.197	1.5
9	53	7	0.023	0.847	0.177	0.132	1.632
10	46	19	0.063	0.91	0.153	0.413	2.045
11	27	12	0.04	0.95	0.09	0.444	2.489
12	15	15	0.05	1	0.05	1	3.489

2.2.1 Distribution de probabilité

En temps discret, la probabilité est définie par $p_t = P(T = t)$. Cette probabilité se calcule simplement comme étant le rapport entre le nombre de cas favorables et le nombre de cas possibles pour chaque instant t donné. Par exemple, au temps $t = 1$, la probabilité associée au nombre d'étudiants qui font des examens de rattrapage vaut $\frac{20}{300} = 0.067$ (il y a eu 20 étudiants qui ont passé des examens de rattrapages sur les 300 étudiants soumis à ce risque), au temps $t = 6$, cette probabilité vaut $\frac{15}{300} = 0.05$. Les autres probabilités se calculent de la même manière.

2.2.2 Fonction de répartition

La fonction de répartition notée F(t) ou F_t représente pour un temps t fixé, la probabilité de vivre l'événement étudié avant l'instant t. Cette probabilité se calcule de la manière suivante :

$$F(t) = \sum_{t=1}^{n} p_t$$

Par exemple, $F(2) = \sum_{t=1}^{2} p_t = p_1 + p_2 = 0.067 + 0.1 = 0.167$

Cette valeur équivaut à la somme des deux premières lignes de la colonne p_t du tableau 2.11. Reprenons l'exemple littéral précédent concernant le calcul de la fonction de répartition et appliquons cet exemple aux données du tableau 2.11. Le but est de calculer la valeur de la fonction de répartition au temps $t = 6$. On écrit $F(6) = \sum_{t=1}^{6} p_t$ se calcule en faisant la somme des probabilités élémentaires de t_1 jusqu'à t_6. Cette fonction de répartition est calculée dans le tableau 2.12.

Tableau 2.12 – Exemple de calcul de la fonction de répartition de t_1 à t_6

t	1	2	3	4	5	6
p_t	0.067	0.1	0.3	0.047	0.1	0.05
F_t	$F_6 = 0.067 + 0.1 + 0.3 + 0.047 + 0.1 + 0.05 = 0.664$					

2.2.3 Fonction de survie ou fonction de séjour

En temps discret, la survie à l'instant t est égale au complémentaire à 1 de la fonction de répartition à l'instant précédent $(t - 1)$. Considérons le tableau 2.13 ci-dessous en temps discret extrait du tableau 2.11 des données sur les étudiants qui refont des examens de rattrapage. La première ligne de ce tableau représente les durées, la

deuxième ligne représente les valeurs de la fonction de répartition à chaque instant t, les survies aux instants allant de t_1 à t_4 sont calculées dans la troisième ligne.

$$S_t = 1 - F_{t-1}$$

Tableau 2.13 – Exemple de calcul de la fonction de séjour de t_1 à t_4

t	1	2	3	4
F_t	0.067	0.167	0.467	0.513 1
S_t	$S_1 = 1$	$S_2 = 1 - 0.067 = 0.933$	$S_3 = 1 - 0.167 = 0.833$	$S_4 = 1 - 0.467 = 0.533$

2.2.4 Risque instantané ou taux de hasard

En reprenant l'exemple littéral précédent, si l'objectif était de calculer le risque ou le risque instantané, nous utiliserions p_t et S_t, qui désignent respectivement la probabilité à l'instant t et la valeur de la fonction de survie à ce même instant t.

La valeur du risque instantané de refaire des examens de rattrapage pour chaque instant t n'est autre que le rapport entre la probabilité de refaire des examens de rattrapage à chaque instant t et la survie à ce même instant t. Le risque instantané de refaire des examens de rattrapage à chaque instant t est calculé dans la dernière ligne du tableau 2.14.

Tableau 2.14 – Exemple de calcul de la fonction de risque instantané de t_1 à t_4

t	1	2	3	4
p_t	0.067	0.1	0.3	0.047
$S(t)$	1	0.933	0.833	0.533
h_t	$h_1 = \frac{0.067}{1} = 0.067$	$h_2 = \frac{0.1}{0.933} = 0.107$	$h_3 = \frac{0.3}{0.833} = 0.36$	$h_4 = \frac{0.047}{0.533} = 0.088$

2.2.5 Risque cumulé ou taux de hasard cumulé

Reprenons l'exemple précédent, cette fois-ci notre objectif est de calculer le risque cumulé de refaire des examens de rattrapage jusqu'au temps $t = 8$ avec les mêmes données.

Le risque cumulé au temps t_8 noté H_8 ou $H(8)$ est égal à la somme des risques élémentaires h_1 + h_2 + h_3 + $\ldots$ + h_8. Le tableau 2.15 ci-dessous, extrait du tableau 2.11, présente les détails de calcul de ce risque cumulé pour les temps allant de t_1 à t_8.

Tableau 2.15 – Exemple de calcul du risque cumulé de t_1 à t_8

t	1	2	3	4	5	6	7	8
h_t	0.067	0.107	0.36	0.088	0.205	0.129	0.347	0.197
H_t	$H_8 = 0.068 + 0.107 + 0.36 + 0.088 + 0.205 + 0.129 + 0.347 + 0.197 = 1.5$							

Nous remarquons à partir de cet exemple l'interdépendance qu'il y a entre ces différentes fonctions. La connaissance de la fonction de probabilité a permis de calculer la fonction de répartition qui à son tour a permis de calculer la fonction de séjour. Connaissant la fonction de séjour, et la fonction de probabilité, nous avons pu calculer la fonction de risque en faisant le rapport entre la fonction de probabilité et la fonction se séjour. Connaissant la fonction de risque, on en déduit la fonction de risque cumulé en faisant simplement la somme des risques élémentaires jusqu'au temps qui nous intéresse. Cette interdépendance des fonctions de base à l'analyse des

parcours de vie est valable pour tout modèle ; qu'il soit non paramétrique, paramétrique ou semi-paramétrique (Cox).

Nous commencerons d'abord par les méthodes non paramétriques de l'analyse des parcours de vie en présentant dans le chapitre suivant la méthode de Kaplan-Meier et la méthode actuarielle.

Chapitre 3

Méthodes non paramétriques d'analyse des parcours de vie : la méthode d'estimation de Kaplan-Meier et la méthode actuarielle

Ce chapitre a pour objectif de présenter les méthodes non paramétriques d'analyse des événements de parcours de vie. Il présente les méthodes d'estimation de Kaplan-Meier et la méthode actuarielle. Le terme non paramétrique implique qu'aucune hypothèse n'est faite en ce qui concerne la distribution des risques d'occurrence de l'événement d'intérêt au cours du temps (Courgeau et Lelièvre, 1989). Cela implique qu'à un instant t donné, le risque est estimé de manière indépendante du risque estimé à l'instant précédent. En d'autres termes, les méthodes non paramétriques impliquent qu'aucune hypothèse n'est faite en ce qui concerne les différences de rythme d'occurrence des événements au cours du temps entre les diverses populations (Le Goff, 2012). Ces deux méthodes sont à vocation descriptives et permettent d'étudier la distribution au cours du temps de la durée de séjour et du risque d'occurrence d'un événement d'intérêt au cours du temps et permettent également de faire de la comparaison de groupes. Le choix de l'une de ces deux méthodes par rapport à l'autre repose sur divers éléments comme la taille de l'échantillon, le rythme d'occurrence de l'événement étudié, de la durée d'observation ou de la connaissance ou non de l'instant exact auquel l'événement d'intérêt a lieu.

L'utilisation de la méthode de Kaplan-Meier est recommandée lorsque nous disposons de données portant sur un faible effectif ou lorsque l'unité de temps considérée est petite. Cette méthode est à déconseiller si de nombreux individus connaissent l'événement à chaque instant ou si les durées ont été mesurées sur une unité de temps large, en l'occurrence en année. L'estimateur de Kaplan-Meier présente certaines propriétés très intéressantes notamment la normalité asymptotique car c'est un estimateur convergent et asymptotiquement sans biais.

Dans la méthode de Kaplan-Meier, les intervalles de temps sont définis par des dates d'événements observés alors que pour la méthode actuarielle les intervalles de temps sont fixés et sont de même longueur (Kankoé, 2010; Blossfeld et al., 2009). La méthode de Kaplan-Meier calcule la survie à chaque fois que l'événement attendu se produit. La fréquence de survenue des événements observés étant aléatoire, la survie est calculée sur des intervalles de temps qui n'ont pas la même longueur ; c'est ce qui fait qu'on obtient une courbe en marche d'escalier pour laquelle l'étendue des marches varie d'un palier à un autre. Dans la méthode de Kaplan-Meier, la survie calculée en un point est conservée jusqu'au point de calcul suivant (jusqu'à la prochaine réalisation de l'événement observé). La méthode de Kaplan-Meier ne présuppose aucune hypothèse en ce qui concerne l'occurrence de l'événement étudié dans les intervalles étudiés alors qu'elle est supposée constante dans les intervalles fixés pour la méthode actuarielle (Villar et al., 2008).

La méthode actuarielle présuppose que les sujets censurés et les occurrences de l'événement étudié se distribuent uniformément dans l'intervalle d'une part, d'autre part que les sujets censurés sont exposés au risque en moyenne durant la moitié de l'intervalle (Courgeau et Lelièvre, 1989; Le Goff et al., 2013). Par exemple, lorsque dans un intervalle de temps donné, il y a dix individus qui ont connu l'événement d'intérêt, la méthode actuarielle fait l'hypothèse que cinq individus l'ont connu dans le premier demi-intervalle et que les cinq autres individus l'ont connu dans le deuxième demi-intervalle. Cette hypothèse fait que lorsque nous ne connaissons

pas les dates exactes d'occurrence de l'événement étudié, la méthode actuarielle sera plus adaptée que la méthode de Kaplan-Meier. De plus, lorsque nous avons de grands effectifs, la méthode actuarielle peut donner des résultats plus fiables que la méthode de Kaplan-Meier. Dans le cas contraire donc, lorsque les effectifs sont petits, on privilégiera la méthode de Kaplan-Meier.

Le fait que l'on ait la possibilité de regrouper les événements par période (par intervalle de temps) fait que la méthode actuarielle demande moins de calcul que la méthode de Kaplan-Meier.

L'estimateur actuariel est aussi asymptotiquement non biaisés et cela quelle que soit la forme de la fonction de survie. Le choix d'une méthode plutôt que l'autre dépend du rythme d'occurrence de l'événement étudié ; si plusieurs individus connaissent l'événement à chaque intervalle de temps, le tracé de l'estimateur de Kaplan-Meier devient peu lisible et le temps de calcul long. Il serait donc préférable dans ce contexte de privilégier la méthode actuarielle (Hill et al., 1991; Allison, 2010).

Pour mettre en pratique ces méthodes, les données sont organisées de la même manière ; on a besoin de la variable temps, de la variable censure et de la variable événement qui prendra la valeur 1 si l'individu a connu l'événement et la valeur 0 sinon. Dans le cas de la compraison de groupes, on aura aussi besoin de la variable « groupe ». Les méthodes de Kaplan-Meier et actuarielle nous pemettront de connaître la distribution de la durée de séjour et de comparer des groupes.

Reprenons l'exemple de nos quatre étudiants africains, Abdoul, Michelle, Khaled et Sadio. Si l'on veut connaître la distribution de la durée des études (durée du bachelor ou durée du master par exemple) de ces étudiants, on pourra utiliser la méthode de Kaplan-Meier ou la méthode actuarielle.

Supposons maintenant que nous ayons un échantillon d'étudiants africains que l'on a regroupé par région d'origine des étudiants (Afrique de l'Ouest, Afrique du Nord, Afrique de l'Est, Afrique Centrale et Afrique Australe). Les méthodes de Kaplan-Meier et actuarielle permettront de comparer par exemple la durée de séjour des étudiants selon leurs régions d'origine.

Les méthodes de Kaplan-Meier et actuarielle sont les deux méthodes non paramétriques d'analyse de parcours de vie les plus répandues et seront présentées dans les sections suivantes.

3.1 La méthode d'estimation de Kaplan-Meier

L'estimateur de Kaplan et Meier (KM), aussi appelé « Product Limit Estimations (PLE) » en anglais est construit à partir de la maximisation de la vraisemblance [1]. La maximisation de la vraisemblance est une méthode statistique qui permet d'obtenir une estimation de paramètres qui maximise (augmente) la probabilité d'obtenir les données observées. Le fondement de cette méthode repose sur l'estimation de la distribution de la fonction de séjour $S(t)$, c'est-à-dire, la distribution au cours du temps de la probabilité de ne pas avoir connu l'événement étudié (Le Goff et al., 2013).

Kaplan et Meier (Kaplan et Meier, 1958) se sont préoccupés pour la première fois de l'estimation des données censurées à droite et ont mis au point un estimateur qui tient compte de cette censure. En observant des échéances $t_1 < t_2 < t_3 ... < t_k$, la vraisemblance des observations est formée pour chaque t_i par :

$$L_i = h_i^{d_i}(1 - h_i)^{N_i - d_i} \tag{3.1}$$

Où :

d_i représente le nombre d'échéances en t_i

N_i est la population soumise au risque juste avant t_i

h_i représente le taux de hasard instantané d'occurrence en t_i

A partir de l'équation (3.1), on obtient l'estimateur du maximum de vraisemblance qui représente pour chaque t_i la proportion $\widehat{h}_i$ des individus qui ont connu l'événement étudié à l'instant t_i.

1. La vraisemblance, est la probabilité que le modèle ait pu généré les données observées

Cette proportion vaut :

$$\widehat{h}_i = \frac{d_i}{N_i} \tag{3.2}$$

($1 - \widehat{h}_i$ correspond à la proportion de personnes n'ayant pas connu l'événement).

Lorsque le temps est considéré comme étant une variable aléatoire discrète et si t_i représente un instant au cours duquel au moins un événement a été observé, alors la probabilité de survie au temps t_i est égale au produit de la probabilité d'avoir survécu avant t_i et de la probabilité conditionnelle de survivre au temps t_i. La probabilité conditionnelle représentant la probabilité de survivre au temps t_i sachant que les individus étaient survivants en t_i (Le Goff et al., 2013).

Cela s'écrit comme suit :

$$S(t_i) = S(t_{i-1})P(T > t_i/T \geq t_i) = S(t_{i-1})(1 - h_i) \tag{3.3}$$

L'estimateur de la fonction de séjour de Kaplan-Meier notée $\widehat{S(t)}$ s'écrit de la manière suivante (Courgeau et Lelièvre, 1989) :

$$\widehat{S(t)} = \prod_{t_i<t} (1 - \widehat{h}_i) \tag{3.4}$$

La variance de l'estimateur de Kaplan-Meier en fonction du temps est connue sous le nom de formule de Greenwood (Courgeau et Lelièvre, 1989; Hosmer et Lemeshow, 1999).

$$Var(\widehat{S(t)}) = \widehat{S(t)}^2 \sum_{t_j} \frac{d_j}{N_j(N_j - d_j)} \tag{3.5}$$

La formule de Greenwood nous permettra de comparer la population qu'on observe à une autre. L'estimateur de Kaplan-Meier satisfait la propriété de la normalité asymptotique (Meier et al., 2004).

Une fois que $\widehat{S(t)}$ est connue, on peut estimer les autres fonctions, comme la fonction de risque cumulé ou la fonction de répartition.

Comme déjà mentionné dans le chapitre 2, la fonction de risque cumulé vaut :

$$H(t) = -ln(S(t)) \tag{3.6}$$

La fonction de risque cumulé nous permet de voir comment le risque évolue au cours du temps et offre aussi la possibilité d'avoir une idée sur le rythme d'occurrence de l'événement étudié.

On peut également calculer la fonction de répartition, qui n'est autre que le complémentaire à un de la fonction de séjour. Cette fonction de répartition nous indique la probabilité de connaître l'événement étudié entre le début d'observation et un instant t quelconque (Le Goff et al., 2013). Elle s'écrit de la manière suivante (voir chapitre 2) :

$$F(t) = 1 - S(t) \tag{3.7}$$

Tableau 3.1 – Exemple de calcul de l'estimateur de Kaplan-Meier

t_i	d_i	N_i	C_i	$1 - \frac{d_i}{N_i}$	$\widehat{S(t)}$	$\widehat{F(t)}$
0	0	31	0	1	1	0.000
8	1	31	0	0.968	0.968	0.032
17	2	30	0	0.933	0.903	0.097
25	3	28	0	0.893	0.806	0.194
30	1	25	0	0.96	0.774	0.226
38	2	24	0	0.917	0.71	0.290
45	2	22	0	0.909	0.645	0.355
70	3	20	0	0.850	0.548	0.452
83	4	17	0	0.765	0.419	0.581
90	1	13	0	0.923	0.387	0.613
109	4	12	0	0.667	0.258	0.742
130	5	8	0	0.375	0.097	0.903
140	1	3	0	0.667	0.065	0.935
164	2	2	0	0	0	1

3.1.1 Exemple d'application de la méthode de Kaplan-Meier

Considérons l'exemple suivant dans lequel nous avons 31 étudiants internationaux soumis au risque de quitter la Suisse au terme de leurs études (l'événement d'intérêt c'est le départ de la Suisse). Dans la première colonne du tableau 3.1, nous avons le temps (t_i), dans la deuxième, le nombre de départs observés (d_i), dans la troisième colonne, le nombre d'individus soumis au risque de quitter la Suisse (N_i), et, dans la quatrième colonne, le nombre de sujets censurés (C_i).

Dans ce tableau , la colonne cinq désignée par $(1 - \frac{d_i}{N_i})$ représente la proportion d'étudiants n'ayant pas quitté le pays, la colonne six désignée par $\widehat{S(t)}$ représente les valeurs de la survie estimée (durée de séjour) pour chaque instant t, et la dernière colonne $\widehat{F(t)}$ représente les valeurs de la fonction de répartition à chaque instant t donné. La fonction de répartition dans ce cas représente la probabilité de quitter la Suisse avant un temps t quelconque.

La première ligne du tableau se lit de la manière suivante : à l'instant $t = 0$, aucun événement ne s'est produit, donc, aucun étudiant n'a quitté le pays ($d_i = 0$), la totalité des étudiants sont soumis au risque de quitter le pays ($N_i = 31$), aucun sujet n'est censuré ($C_i = 0$), personne n'a quitté le pays ($1 - \frac{d_i}{N_i} = 1$) et la probabilité de départ de la Suisse avant $t = 0$ est nulle.

La troisième ligne qui correspond à l'instant $t = 17$ nous dit qu'à cet instant, deux étudiants ont quitté la Suisse ($d_i = 2$), 30 étudiants sont soumis au risque de départ, aucun étudiant n'est censuré ($C_i = 0$), la proportion d'individus n'ayant pas quitté le pays est de 0.933 soit 93.3%, la survie à cet instant vaut 0.903 et la probabilité de quitter le pays avant l'instant $t = 17$ est de 0.097.

Les autres lignes du tableau se lisent de la même manière. On remarque dans ce tableau, une des propriétés de la fonction de survie ; elle vaut 1 à l'instant $t = 0$ et 0 lorsqu'on tend vers de grandes valeurs de t.

Sur la Figure 3.1, nous avons les représenté graphiquement de la courbe de survie et la fonction de répartition de l'exemple développé ci-dessus. On observe sur ce graphique que la courbe de survie est une fonction décroissante du temps, elle vaut 1 à l'instant de départ puis décroit progressivement et tend vers 0 au fur et à mesure que t augmente. Cette courbe de survie nous donne pour chaque instant t, la valeur de la survie correspondante (probabilité de ne pas quitter la Suisse). Par exemple, lorsque $t = 25$, la probabilité de ne pas quitter la Suisse vaut 0.8 et lorsque $t = 85$, cette probabilité vaut 0.4. La courbe de la fonction de répartition elle, est

une fonction croissante du temps et donne pour chaque temps t, la probabilité de quitter la Suisse avant ou à ce temps t. Par exemple, la probabilité de départ avant $t = 25$ vaut 0.2 et la probabilité de départ avant $t = 90$ vaut 0.6.

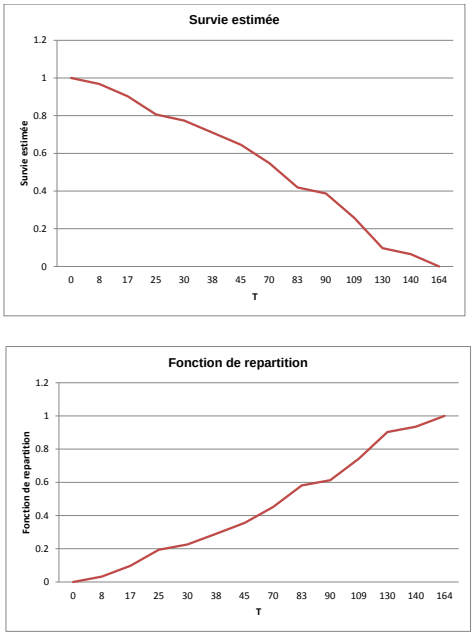


FIGURE 3.1 – Graphiques de la fonction de survie et de répartition des données de la Table 3.1

3.2 La méthode actuarielle

Dans la méthode actuarielle, on fait l'hypothèse que le risque d'occurrence de l'événement étudié est constant dans l'intervalle de temps t_i et t_{i+1} (Le Goff et al., 2013; Allison, 2010). On suppose aussi une répartition uniforme dans cet intervalle du nombre d'individus ayant connu l'événement étudié et les sorties d'observation (Le Goff et al., 2013). Considérons un échantillon d'étudiants africains par exemple, et on s'intéresse à l'obtention d'un permis de séjour de six mois en vue de trouver un travail après les études universitaires. Supposons que l'intervalle de temps défini soit l'année et que douze étudiants aient obtenu le permis dans le premier mois (janvier) et que douze autres étudiants l'aient obtenu au mois de décembre. Au total, 24 étudiants ont obtenu le permis au cours de l'année ; la méthode actuarielle suppose qu'au cours de cette année, deux permis ont été délivrés par mois car on répartit uniformément les échéances au cours de cet intervalle de temps exprimé en année. Ce principe est aussi valable pour les personnes qui sont sorties de l'étude, s'il y a douze personnes qui sortent de l'étude au cours d'un mois donné et personne le reste de l'année, on supposera qu'un étudiant est sorti de l'étude par mois au cours de cette année.

Soit les k intervalles de temps suivants $[0, t_1[$, $[t_1, t_2[$, ... $[t_{k-1}, \infty[$;

d_i, le nombre d'occurrence de l'événement étudié dans le $i^{ème}$ intervalle $[t_{i-1}, t_i[$ en sachant que $t_0 = 0$ et $t_k = \infty$.

n_{i-1}, le nombre d'individus n'ayant pas connu l'événement étudié au temps t_{i-1},

C_i, le nombre de sujets censurés dans l'intervalle $[t_{i-1}, t_i[$, et

r_i, le nombre de sujets à risque dans l'intervalle $[t_{i-1}, t_i[$.

Dans le but de simplifier les calculs, on suppose très souvent que les censures sont réparties uniformément dans l'intervalle, c'est-à-dire que les sujets censurés sont exposés en moyenne un demi-intervalle. La contribution des individus à risque pour l'intervalle $[t_{i-1}, t_i[$ est donc $\frac{C_i}{2}$.

Le nombre d'individus à risque pour l'intervalle $[t_{i-1}, t_i[$ est alors :

$$r_i = n_{i-1} - \frac{C_i}{2} \tag{3.8}$$

Ainsi, la probabilité conditionnelle $p_i = P(T \leq t_i \mid T > t_{i-1})$ de connaître l'événement étudié dans l'intervalle $[t_{i-1}, t_i[$ sachant que l'on ne l'avait pas connu en t_{i-1} est estimée par :

$$\widehat{p_i} = \frac{d_i}{r_i} = \frac{d_i}{n_{i-1} - \frac{C_i}{2}} \tag{3.9}$$

L'estimateur actuariel de la fonction de survie s'écrit :

$$\widehat{S(t)} = \prod_{i|t_i \leq t} \left(1 - \frac{d_i}{r_i}\right) \tag{3.10}$$

Une estimation de la variance de la fonction de survie s'obtient à partir de la formule de Greenwood et s'écrit comme suit :

$$\widehat{Var}(\widehat{S}(t)) = \widehat{S(t)}^2 \sum_{i|t_i \leq t} \frac{d_i}{r_i(r_i - d_i)} \tag{3.11}$$

Le risque cumulé est estimé par la relation suivante (Le Goff et al., 2013) :

$$H(t_i) \cong \sum_{i=1}^{k} p_k \tag{3.12}$$

3.2.1 Exemple d'application de la méthode actuarielle

Les données qui sont utlisées pour développer cet exemple proviennent de l'article de Kankoé Sallah, (Kankoé, 2010), le contexte a été ramené à celui des étudiants internationaux.

Dans cet exemple, nous avons 210 étudiants internationaux qui ont fait une demande d'obtention d'un permis provisoire de six mois dans le but de chercher un travail après la fin de leurs études. Dans la première colonne du tableau 3.2, nous avons le temps (t_i), qui représente la durée écoulée entre la date de la demande et la date d'obtention d'une réponse de la part des autorités, dans la deuxième colonne, le nombre d'individus censurés (C_i), dans la troisième colonne, le nombre d'individus ayant obtenu le permis souhaité (d_i) et dans la quatrième colonne, le nombre d'individus soumis au risque de l'obtenir (r_i). L'événement d'intérêt, est « obtenir le permis provisoire » . Nous devons obligatoirement disposer des informations sur ces quatre colonnes, les colonnes à calculer sont donc les colonnes cinq (p_i) et six $(\widehat{S(t)})$. La colonne cinq représente la probabilité d'obtenir le permis provisoire ; et la colonne six, la survie estimée ou la durée de séjour estimée.

Tableau 3.2 – Exemple de calcul de l'estimateur actuariel

t_i (semaines)	c_i	d_i	r_i	p_i	$\widehat{S(t)}$
0	pas défini	pas défini	pas défini	pas défini	1
3	0	0	210	$1 - \frac{0}{210 - \frac{0}{2}}$	1
9	10	40	210=(210-0-0)	$0.805 = 1 - \frac{40}{210 - \frac{10}{2}}$	0.805
12	30	10	160 = (210-10-40)	$0.931 = 1 - \frac{10}{160 - \frac{30}{2}}$	0.749
18	10	20	120=(160-30-10)	$0.826 = 1 - \frac{20}{120 - \frac{10}{2}}$	0.619
21	20	0	90	$1 = 1 - \frac{0}{90 - \frac{20}{2}}$	0.619
23	0	20	70	$0.714 = 1 - \frac{20}{70 - \frac{0}{2}}$	0.442
27	18	3	50	$0.927 = 1 - \frac{3}{50 - \frac{18}{2}}$	0.410
36	8	2	29	$0.92 = 1 - \frac{2}{29 - \frac{8}{2}}$	0.377

Dans ce tableau, à l'instant $t_i = 0$, aucun étudiant n'a obtenu le permis provisoire et la probabilité de survie vaut donc 1.

A l'instant $t_i = 3$, nous n'avons aucune donnée censurée $(C_i = 0)$, personne n'a obtenu le permis provisoire $(d_i = 0)$, le nombre d'individus soumis au risque de l'obtenir est toujours le même (210 = 210-0-0). La probabilité de survie (la probabilité de ne pas obtenir le permis provisoire) vaut 1 et la survie aussi vaut 1 (voir colonne 6). Au temps $t_i = 9$, nous avons 10 individus censurés, 40 étudiants ont obtenu le permis provisoire, la probabilité de l'obtenir vaut 0.805 et la survie à cet instant vaut $1 * 0.805 = 0.805$. Au temps $t_i = 12$, nous avons 30 personnes censurées, 10 étudiants qui ont obtenu le permis provisoire, le nombre de personnes soumises au risque de l'obtenir à cet instant est donc de $(210 - 10 - 40 = 160)$ car au temps $t_i = 9$, 10 individus étaient censurés et 40 avaient obtenu le permis provisoire. La probabilité d'obtenir le permis est donc de 0.931 pour une survie ou durée de séjour de $0.805 * 0.931 = 0.749$. Ainsi, au temps $t_i = 36$, nous avons 8 individus censurés, 2 étudiants qui ont obtenu le permis pour une probabilité de 0.920, soit une survie de 0.377. Le reste du tableau se lit et s'interprète de la même manière.

Une représentation graphique de ces résultats est présentée dans la Figure 3.2. Sur ce graphique, on voit que la fonction de survie est une fonction décroissante du temps, elle vaut 1 au début et décroît progressivement en tendant vers 0 au fur et à mesure que t augmente. A chaque instant t donné, on peut obtenir la valeur de la survie correspondante en cherchant l'intersection entre le temps concerné et la courbe de survie. Par exemple, à l'instant $t = 21$, la survie correspondante vaut environ 0.6. La courbe de la fonction de répartition, elle, nous permet de connaître la probabilité d'obtenir le permis provisoire avant un temps quelconque t. Par exemple, on peut lire sur le graphique que la probabilité d'obtenir le permis provisoire avant $t = 9$ est de 0.2 et la probabilité de l'obtenir avant $t = 18$ est de 0.38.

Pour les deux méthodes présentées, il est possible de faire des tests statistiques dans le but de comparer des groupes. Les différents tests statistiques sont présentés dans la partie qui va suivre.

3.3 Tests statistiques

Dans les analyses non paramétriques, on peut s'intéresser à la comparaison de courbes de survie entre deux ou plusieurs groupes. Plusieurs tests sont proposés dans les logiciels courants (SPSS, Stata ou SAS par exemple) pour faire cette comparaison de distribution de courbes de séjour. A l'image de tous les tests en statistique, nous

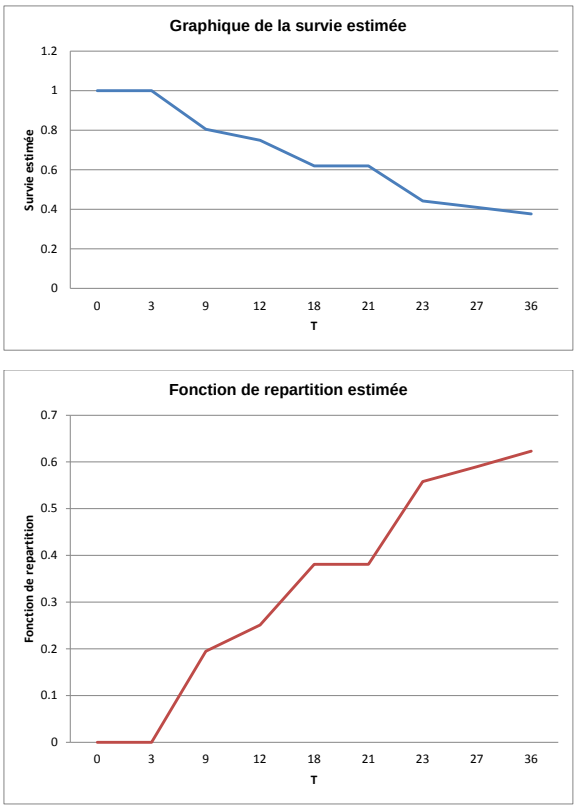


FIGURE 3.2 – Graphique de la fonction de survie et de répartition des données du Tableau 3.2

avons besoin d'une hypothèse nulle, d'une hypothèse alternative et d'une statistique de test afin de pouvoir réaliser ces tests statistiques.

Dans ces différents tests, l'hypothèse nulle à tester postule l'égalité des fonctions de séjour ou fonctions de risque entre les différents groupes à comparer. En d'autres termes, on teste l'égalité entre la proportion d'individus ayant connu l'événement d'intérêt dans deux groupes.

Le test le plus utilisé est le test du Log-rank qui, dans la comparaison de deux groupes suit une loi de χ^2 à un degré de liberté (Hill et al., 1991). Le principe de ce test consiste à comparer le nombre d'occurrence de l'événement étudié dans un groupe au nombre attendu d'occurrence de l'événement étudié dans le même groupe sous hypothèse d'indépendance (sous H_0) puis de calculer une statistique de test.

Les hypothèses nulle et alternative peuvent s'exprimer soit à l'aide des fonctions de séjour (Hill et al., 1991; Kleinbaum, 1996) soit à l'aide des fonctions de risque (Cleves et al., 2010).

Ces hypothèses s'écrivent, en fonction de la durée de séjour, de la manière suivante :

$$H_0 : S_1(t) = S_2(t) \quad \text{(Egalité des durées de séjour dans les deux groupes)}$$
$$H_1 : S_1(t) \neq S_2(t) \quad \text{(Les durées de séjour dans les deux groupes sont différentes)}$$

Ces hypothèses s'écrivent de la manière suivante en fonction du risque :

$H_0 : h_1(t) = h_2(t)$ (Egalité pour chaque t_i du risque d'occurrence de l'événement dans les deux groupes)

$H_1 : h_1(t) \neq h_2(t)$ (Le risque d'occurrence de l'événement dans les deux groupes pour chaque t_i est différent)

Il existe aussi d'autres tests pour faire cette comparaison comme le test de Wilcoxon (Breslow), le test de Tarone-Ware, le test de Peto-Peto-Prentice ou le test de Fleming-Harrington.

3.3.1 Statistique de test

Pour réaliser ces tests statistiques, nous avons besoin de calculer une grandeur U qui s'appelle statistique de test. Cette statistique de test mesure, pour chaque instant t_i, l'écart entre le nombre d'occurrence réellement observé de l'événement étudié et le nombre d'occurrence attendu de ce même événement sous hypothèse d'indépendance multiplié par une grandeur w_i qui représente un poids (Le Goff et al., 2013; Hill et al., 1991; Cleves et al., 2010).

Cette statistique de test s'écrit :

$$U = \sum_i w_i(O_i - E_i) \tag{3.13}$$

Avec :

O_i : les échéances observées,
E_i : les échéances attendues sous l'hypothèse H_0 ,
wi : la pondération ou poids.

De cette statistique de test, on déduit plusieurs autres tests plus ou moins similaires en fonction de la valeur prise par la pondération w_i (Hosmer et Lemeshow, 1999; Le Goff et al., 2013).

Test du Log-Rank : $w_i = 1$

Le test du Log-Rank suppose que tous les poids sont identiques et égaux à 1, cela implique que le poids des événements est identique.

Test de Breslow (Wilcoxon) : $w_i = n_i$

Dans le test de Breslow ou test de Wilcoxon, la pondération est déterminée par la taille de l'échantillon pour chaque t_i. Lorsque les fonctions de risque varient de manière non proportionnelles, le test de Wilcoxon offre des résultats plus fiables que le test du Log-Rank (Cleves et al., 2010).

Test de Taronne-Ware : $w_i = \sqrt{n_i}$

Lorsque la taille de l'échantillon est grande, le test de Taronne-Ware et de Breslow accordent plus de poids aux événements ayant eu lieu en début de période d'observation et peu de poids à ceux ayant eu lieu en fin de période d'observation (Cleves et al., 2010; Le Goff et al., 2013).

Test de Peto-Peto-Prentice : $w_i = \widetilde{S(t_i)}$

Dans ce test, le poids est déterminé par une estimation de la fonction de survie globale qui est similaire mais pas exactement identique à la fonction de survie de Kaplan-Meier (Cleves et al., 2010). Ce test n'est pas sensible aux différences liées aux censures entre les différents groupes.

Test de Fleming-Harrington : $w_i = \left(\widehat{S(t_i)}\right)^p \left(1 - \widehat{S(t_i)}\right)^q$

Dans ce test, $\widehat{S(t_i)}$ représente l'estimateur de Kaplan-Meier, p et q sont des paramètres qui peuvent être fixés par le chercheur. On peut rencontrer les situations suivantes :

1. Lorsque $p > q$, on accorde plus de poids aux événements qui ont eu lieu en début d'observation
2. Lorsque $p < q$, on accorde plus de poids aux événements qui ont eu lieu en fin d'observation
3. Lorsque $p = q = 0$, le test de Fleming-Harrington n'est autre que le test du Log-Rank.

Pour tous ces tests, lorsqu'on compare deux groupes, la statistique de test suit une loi de χ^2 à un degré de liberté (Le Goff et al., 2013) ; lorsque nous comparons plusieurs groupes, la statistique de test suit sous l'hypothèse H_0 une loi de χ^2 à $k-1$ degrés de liberté, k désignant le nombre de groupes à comparer (Cleves et al., 2010; Blossfeld et al., 2009).

Tous ces tests sont implémentés dans les logiciels courants, notamment dans Stata et sont très faciles à mettre en application [2].

3.4 Exemple d'application de l'estimation de Kaplan-Meier et de l'estimation actuarielle

Le but de cet exemple est de faire une estimation de la méthode de Kaplan-Meier et de la méthode actuarielle avec le logiciel Stata tout en montrant la structure de la base de données en vue de faire ce type d'analyse.

3.5 Estimation par la méthode de Kaplan-Meier

La question de recherche à laquelle nous tentons de répondre concerne la durée des études des étudiants internationaux et des étudiants africains au niveau du bachelor dans une université donnée. Cette question de recherche se formule de la manière suivante :

La durée des études de bachelor des étudiants africains est-elle différente de celle des étudiants internationaux ?

Pour répondre à cette question de recherche, considérons une base de données [3] dans laquelle on a enregistré l'année de début du bachelor, l'année d'obtention du diplôme de bachelor, la nationalité, ainsi que l'âge de 555 étudiants internationaux qui ont débuté des études de bachelor dans une université.

3.5.1 Données

La base de données provient de la quatrième édition du livre « An Introduction to Stata for Health Researchers » (Cleves et al., 2010). Des variables ont été recodées et d'autres ont été renommées de manière à obtenir la structure des données présentée dans le tableau 3.3 et de ramener l'exemple à la problématique des étudiants internationaux.

La variable nationalité est une variable qualitative nominale qui prend la valeur 1 si la personne a une nationalité africaine et la valeur 0 sinon. La variable censure est aussi une variable qualitative qui prend la valeur 1 si la personne a obtenu son bachelor et la valeur 0 sinon. La variable « age1 » est une variable quantitative alors que les variables « debut-bach » et « fin-bach » représentent respectivement la date de début du bachelor et la date de fin du bachelor.

2. Dans Stata : sts test nom de la variable, nom du test. Par exemple : sts test sexe, wilcoxon

3. En raison du très grand retard lié à la livraison des données par l'OFS, nous avons décidé de prendre des bases de données soit dans des livres, soit sur internet pour tester les différents modèles. Ceci nous a permis d'avancer dans ce travail en nous offrant la possibilité de tester tous les modèles ciblés avant la livraison très tardive des données. Ces exemples visant uniquement à illustrer les différents modèles, nous les gardons même si les données ne sont pas en lien avec la mobilité des étudiants.

Tableau 3.3 – Exemple de format de données pour une estimation de Kaplan-Meier

id	nationalite	debut-bach	fin-bach	c_i	age1
1	1=afrique	31oct2004	31dec2009	0	23
2	0=sinon	29août2006	31dec2009	1	22
.	.	.	.	.	.
.	.	.	.	.	.

3.5.2 Analyses exploratoires

Avant toutes analyses de données, il est nécessaire procéder à des analyses exploratoires dans le but de vérifier la cohérence des données enregistrées d'une part, d'autre part ces analyses exploratoires permettent d'avoir une idée sur la distribution des variables. La Figure 3.3 représente la boîte à moustaches de la variable « dur-bach » ainsi que le diagramme en barre de la variable nationalité.

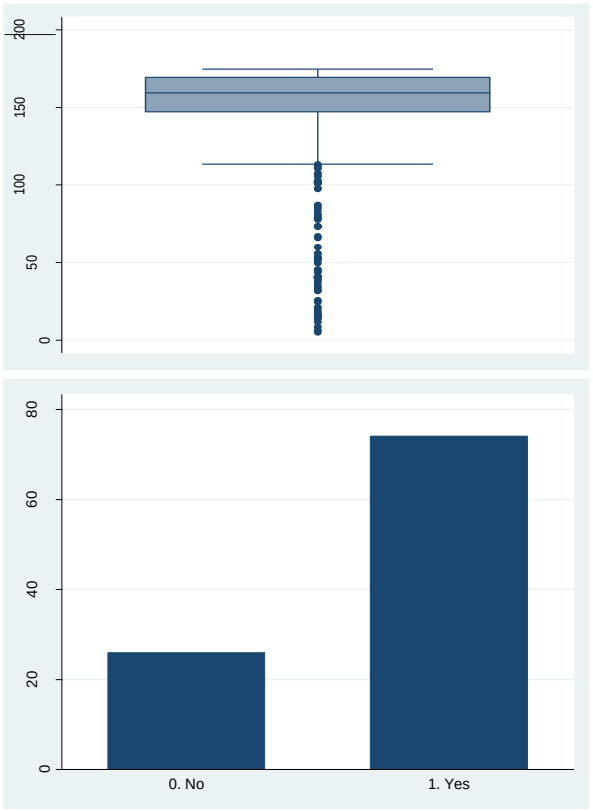


FIGURE 3.3 – Boîte à moustaches de la variable age1 et diagramme en barre de la variable nationalité des données du Tableau 3.3

Sur la boîte à moustaches, on remarque que la variable « dur-bach » qui représente la durée du bachelor (écart entre le début et la fin du bachelor) a des valeurs extrêmes. Le diagramme en barre de la variable nationalité montre qu'on a plus d'étudiants africains que d'étudiants venant d'autres pays. On peut à présent estimer la

distribution de la durée et faire des représentations graphiques de ces différentes fonctions ainsi que les tests statistiques qui nous permettront de répondre à la question de recherche.

3.5.3 Estimation de la fonction de séjour

Le Tableau 3.4 donne un extrait des calculs de l'estimation de la fonction de séjour. Dans ce tableau, la première colonne représente le temps, la deuxième colonne le nombre d'individus soumis au risque de connaître l'événement d'intérêt (obtenir le diplôme de bachelor), la troisième colonne représente le nombre d'individus ayant connu l'événement d'intérêt à chaque instant, dans la quatrième colonne, on a le nombre d'individus sortis de l'étude à chaque instant. Dans la cinquième colonne, nous avons une estimation de la fonction de survie à chaque instant t, dans les sixièmes et septièmes colonnes, nous avons respectivement les erreurs standards et les intervalles de confiance associés à l'estimation de la fonction de séjour. Au temps $t = 61$, on avait 555 étudiants africains et internationaux soumis à observation, un seul a obtenu son bachelor (Evénement=1) à l'instant $t = 61$, personne n'est sorti de l'étude (perdu = 0) et la fonction de survie estimée vaut 0.9982. Dans ce tableau, on observe qu'on a des durées de séjour très élevées et proches de 1 au départ et des durées de séjour faibles et proches de 0 vers la fin.

Tableau 3.4 – Distribution de la fonction de survie de l'estimation de Kaplan-Meier

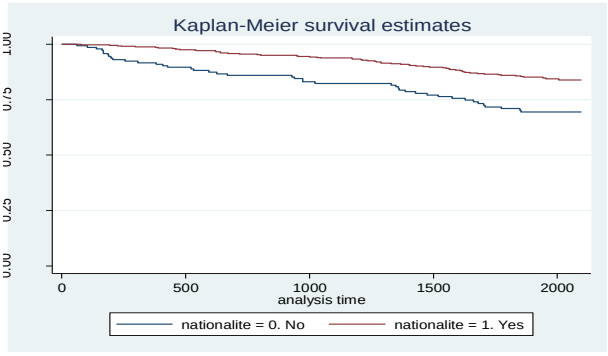
Temps	Nombre d'individus	Evénement	perdu	Survie	Erreur standard	IC
61	555	1	0	0.9982	0.0018	0.987-0.999
77	554	1	0	0.9964	0.0025	0.985-0.999
103	553	1	0	0.9946	0.0031	0.983-0.998
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2093	51	1	0	0.0178	0.7632	0.8332

Sur la Figure 3.4, nous avons la représentation de l'estimateur de Kaplan-Meier et le graphique du risque cumulé selon la nationalité. Sur ce graphique, on peut à chaque instant comparer les probabilités de finir les études de bachelor pour les étudiants africains et pour les étudiants internationaux. Ce graphique a une forme en escalier, ce qui est une particularité de la courbe de Kaplan-Meier. En effet, lorsqu'un individu a connu l'événement d'intérêt à un instant donné, la fonction de séjour reste horizontale quel que soit le temps t jusqu'à ce qu'un autre individu connaisse l'événement d'intérêt. La courbe la plus haute a les probabilités les plus élevées et la courbe la plus basse, les probabilités les plus faibles. Graphiquement, on peut dire que la durée des études de bachelor des étudiants africains est plus longue que celle des autres étudiants internationaux parce que la courbe associée aux étudiants africains est au-dessus de celle associée aux étudiants internationaux.

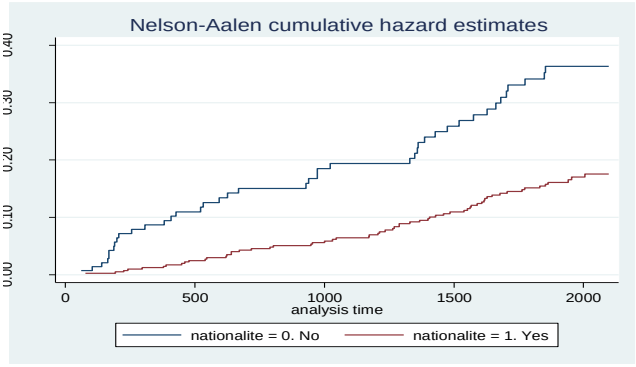
La figure 3.4 représente aussi la courbe de risque cumulé selon la nationalité ; ce graphique nous permet de connaître à chaque instant et selon la nationalité la proportion d'individus ayant terminé les études de bachelor parmi tous les étudiants qui suivent cette formation. L'écart entre les deux courbes nous donne la différence en terme de probabilité de finir les études de bachelor pour les étudiants non africains et les étudiants africains. On remarque ainsi sur ce graphique qu'au temps $t = 10$, cette probabilité est de 0.05 pour les étudiants africains contre 0.19 pour les autres étudiants. La probabilité de finir les études de bachelor avant chaque instant t est plus élevée pour les étudiants internationaux que pour les étudiants africains.

3.5.4 Tests statistiques

L'hypothèse nulle postule l'égalité du nombre d'occurrence de l'événement d'intérêt au sein des deux groupes. Dans cet exemple, l'hypothèse nulle peut se traduire par une égalité entre la proportion d'étudiants africains ayant obtenu le diplôme de bachelor et la proportion d'étudiants internationaux ayant obtenu le même diplôme pour chaque instant. Cette hypothèse nulle peut aussi être formulée par une égalité des fonctions de séjour (ou des fonctions de risque) entre le groupe des étudiants africains et le groupe des étudiants internationaux. Les hypothèses nulle et alternative peuvent être formulées comme suit :



Graphique1 : Fonction de survie de K-M



Graphique2 : Risque cumulé

FIGURE 3.4 – Courbe de l'estimateur de Kaplan-Meier et de la fonction de risque cumulé en fonction de la nationalité

$H_0 : S_1(t) = S_2(t)$ vs $H_1 : S_1(t) \neq S_2(t)$

On peut aussi écrire ces hypothèses à l'aide des fonctions de risque.

$H_0 : h_1(t) = h_2(t)$ vs $H_1 : h_1(t) \neq h_2(t)$

La statistique de test suit une distribution de Chi-carré à $k-1$ degrés de libertés, k désignant le nombre de groupes. On rejette H_0 au seuil de 5% si la statistique de test est supérieure au Chi-Carré à $k-1$ degrés de liberté ou si la p-valeur associée au test est inférieure à 0.05. Le test du logrank est présenté dans le tableau 3.5.

Tableau 3.5 – Test de Log-rank pour l'égalité des fonctions de survie

Nationalité	Evénements Observés	Evénements espérés
Non	42	24.23
Oui	61	78.77
Total	103	103.00
Chi2(1) = 17.06	Pr> Chi2(1) = 0.0001	

Dans cet exemple, la p-valeur associée au test de Log-rank est de $0.0001 < 0.05$, on rejette H_0 au seuil de 5%

ce qui nous permet de dire que la durée des études de bachelor des étudiants africains est différente de celle des autres étudiants internationaux et on conclut que notre hypothèse de recherche est corroborée. Pour comparer les deux groupes d'étudiants, on aurait aussi pu faire le test de Wilcoxon ou d'autres tests comme : le test de Tarone-Ware, le test de Peto-Peto-Prentice ou le test de Fleming-Harrington.

3.6 Estimation par la méthode actuarielle

Dans cette partie, on va répondre à la même question de recherche, avec les mêmes données mais en utilisant une estimation actuarielle. On remarquera que pour cette méthode, la survie est calculée dans des intervalles de temps d'une part, d'autre part, la censure est répartie de manière égale dans chaque intervalle de temps. Cette méthode offre ainsi la possibilité d'avoir un tableau de distribution des durées moins volumineux que celui obtenu dans la méthode de Kaplan-Meier à travers la possibilité qu'elle offre de résumer les données dans des intervalles de temps.

Question de recherche : la durée des études de bachelor des étudiants africains est-elle différente de celle des étudiants internationaux ?

Les analyses exploratoires effectuées précédemment sont valables pour cette analyse et les interprétations restent identiques.

3.6.1 Estimation de la survie actuarielle

Dans le Tableau 3.6 , nous avons une estimation de la fonction de survie actuarielle. Dans ce tableau, on remarque que la durée est exprimée sous forme d'intervalle de temps de largeur identique. Dans cet exemple, la largeur de cet intervalle est d'une année, mais, cette largeur peut être définie comme on le souhaite dans les options de Stata ou de tous les logiciels courants permettant de faire ce type d'analyse. Dans le premier intervalle de temps, la totalité des étudiants (555 étudiants) étaient soumis à observation, cinq étudiants ont obtenu le diplôme de bachelor, douze étudiants ont disparu de l'étude, la survie estimée pour cet intervalle de temps est de 0.9909. Pour calculer cette survie, on calcule d'abord la probabilité conditionnelle d'obtenir son bachelor qui vaut $p_i = \dfrac{5}{555 - \dfrac{12}{2}} = 0.0091$. Le complémentaire à 1 (1-0.0091 = 0.9909) de cette probabilité représente la survie estimée pour cet intervalle. Dans l'intervalle suivant c'est-à-dire [1 ,2[, nous n'avons plus que 538 étudiants qui poursuivent les études de bachelor parce que dans l'intervalle précédent, cinq étudiants avaient obtenu le bachelor et douze avaient disparu de l'étude ; il ne reste plus que 555-5-12 = 538 étudiants soumis à observation. Dans cet intervalle, 26 étudiants ont obtenu le bachelor et 13 sont sortis de l'étude ; la probabilité d'obtenir le bachelor dans cet intervalle vaut $p_i = \dfrac{26}{538 - \dfrac{13}{2}} = 0.0489$. La survie correspondante à cet intervalle est calculée en faisant le produit de la survie au temps un et au temps deux soit $0.9909 * (1 - 0.0489 = 0.9424)$. Les deux dernières colonnes donnent les erreurs standards (S.E.) ainsi que les intervalles de confiance (IC) à 95% pour les survies estimées. Le reste du tableau se lit de la même manière.

Tableau 3.6 – Distribution de la fonction de survie actuarielle

Interval de temps	Effectif total	Evénement	Perdu	Survie	S.E.	IC à 95%
0 1	555	5	12	0.9909	0.0041	0.9783 ; 0.9962
1 2	538	26	13	0.9424	0.0100	0.9191 ; 0.9592
2 3	499	11	6	0.9215	0.0116	0.8953 ; 0.9414
3 4	482	24	9	0.8752	0.0144	0.8439 ; 0.9006
4 5	449	23	12	0.8298	0.0165	0.7947 ; 0.8594
5 6	414	322	92	0.1037	0.0145	0.0776 ; 0.1342

Dans ce tableau, SE = erreur standard et IC = intervalle de confiance.

3.6.2 Représentation graphique de la courbe de survie actuarielle

La représentation graphique de la courbe de survie actuarielle de la figure 3.5 nous montre que celle-ci est une fonction décroissante du temps ; elle a des valeurs très élevées au début puis décroît progressivement pour tendre vers 0. Ce graphique montre qu'au début, la durée des études des étudiants africains est légèrement inférieure à celle des étudiants internationaux pour ensuite converger vers les mêmes valeurs en devenant de plus en plus proche de zéro. Le résultat du test statistique reste identique à celui effectué dans l'estimation de Kaplan-Meier.

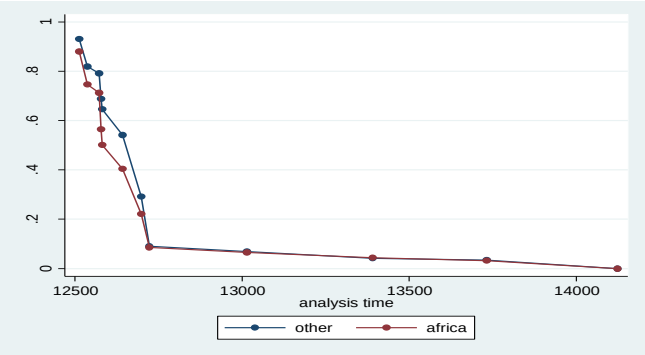


FIGURE 3.5 – Représentation graphique de la courbe de survie actuarielle en fonction de la nationalité

Comme nous venons de le voir, la méthode de Kaplan-Meier et la méthode actuarielle permettent de donner une estimation de la fonction de survie. Cette estimation ne tient compte d'aucune hypothèse ni en ce qui concerne la forme du risque ni sur le rythme d'occurence des événements étudiés.

Il existe d'autres modèles dits explicatifs qui nous permettent de mesurer l'impact des caractéristiques individuelles sur la probabilité ou sur le risque d'occurrence de l'événement étudié. Les modèles paramétriques et le modèle semi-paramétrique de Cox appartiennent à cette famille de modèle. Ces modèles permettent aussi d'évaluer l'importance de l'effet propre de chaque caractéristique individuelle sur la durée de séjour ou sur la fonction de risque. On peut ainsi résumer l'allure de la fonction de risque à partir d'un ou de plusieurs paramètres estimés. Nous commencerons d'abord par présenter les modèles de régression paramétriques qui permettent de donner une forme paramétrique à la distribution des durées non observées. Chaque modèle présenté sera suivi d'une application sur une question de recherche.

Chapitre 4

Introduction aux modèles de régression paramétriques

Dans cette partie, nous allons aborder les modèles dits explicatifs ou prédictifs dans la mesure où ces modèles nous permettent d'expliquer ou de prédire une variable donnée à partir d'une ou de plusieurs autres variables. Ces modèles paramétriques sont des modèles de régression.

Pour introduire ces modèles, nous allons reprendre l'exemple de nos quatre étudiants africains (Abdoul, Michelle, Khaled et Sadio).

Ces modèles permettent de répondre à des questions du genre :

Quel est le risque [1] de terminer les études de master cinq ans après le début des études de bachelor en fonction du sexe, de l'âge, de la nationalité, de la haute école fréquentée et de la filière d'études ?

Pour cette question de recherche, la « variable dépendante [2] » est le risque de terminer le master et les variables indépendantes sont : le sexe, l'âge, la nationalité, la haute école fréquentée et la filière d'études.

Quel est le risque d'abandonner les études selon l'âge, la situation financière de l'étudiant, le sexe, la cohorte d'entrée et la haute école fréquentée ?

Pour cette deuxième question de recherche, la « variable dépendante » est aussi le risque (le risque d'abandon des études) et les variables indépendantes sont : l'âge, la situation financière, le sexe, la cohorte d'entrées et la filière d'étude.

Quelle est la durée des études universitaires selon l'âge, le sexe et le continent d'origine des étudiants ?

Pour cette troisième question de recherche, la variable dépendante est la durée des études qui est en fait la durée de séjour et les variables indépendantes sont : l'âge, le sexe et le continent d'origine des étudiants. Dans ce troisième exemple de question de recherche, la variable dépendante qui est le temps est une caractéristique observée.

Ces exemples de questions de recherche nous montrent que les caratéristiques individuelles peuvent agir soit sur la fonction de risque (cas des deux premiers exemples de questions de recherche) soit sur la fonction de séjour (cas de l'exemple de la troisième question de recherche).

Pour répondre à ces questions de recherche, nous avons besoin de données longitudinales parce que les mesures sur les individus sont répétées d'année en année. L'organisation de la base de données pour répondre à ces questions de recherche est identique. Des exemples d'organisation de bases de données pour la mise en pratique des méthodes seront présentés dans les sections suivantes.

Supposons maintenant que l'on s'intéresse au départ de la Suisse de nos étudiants (Abdoul, Michelle, Khaled et Sadio) après les études. Dans les modèles paramétriques, le fait de quitter la Suisse est exprimé à travers la fonction de risque qui, elle même, dépend des caractéristiques individuelles ; les caractéristiques individuelles

1. Pour rappel, on parle de risque en tant que fonction statistique même si la formulation peut parfois paraître superflue.
2. Nous mettons variable dépendante entre guillemets parce que le risque dans ce cas n'est pas une caractéristique observée.

pouvant être par exemple, le sexe, l'âge, la nationalité, la cohorte d'entrées en Suisse, etc. Le fait que dans les modèles paramétriques, les caratéristiques individuelles puissent agir soit sur la fonction de risque ou sur la fonction de séjour subdivise les modèles paramétriques en deux groupes de modèles. Avant d'introduire ces deux groupes de modèles, on va d'abord rappeler la notion d'hypothèses de base ou de postulats d'application.

La plupart des modèles en statistiques reposent sur des hypothèses de base, c'est-à-dire des conditions sans lesquelles les résultats obtenus à partir de ces modèles conduiraient à fausser les paramètres estimés. Par exemple, pour une régression linéaire, on fait l'hypothèse que la relation est linéaire entre la variable dépendante et les variables indépendantes quantitatives, que les résidus sont normalement distribués et que leur variance est homogène. Pour faire un test t de comparaison de moyennes, on suppose que les variables suivent des lois normales. Avant d'appliquer un modèle quelconque, la première chose à faire est de vérifier que les hypothèses de base ou postulats d'application du modèle concerné sont remplis. Les modèles d'analyses de biographie n'échappent pas à cette règle et reposent sur des hypothèses de base. L'objectif de ce chapitre est de présenter les hypothèses sous-jacentes à l'analyse paramétrique des biographies et de montrer comment on peut tester ces hypothèses par des approches graphiques ou statistiques puis de vulgariser ces méthodes à l'aide d'exemples d'application. La première hypothèse de l'analyse des biographies porte sur la forme de la force d'occurrence de l'événement étudié qui est exprimée à travers la fonction de risque. Les modèles paramétriques et semi-paramétriques supposent que le rapport des fonctions de risque entre deux individus ayant des caractéristiques individuelles différentes est non seulement constant mais aussi indépendant du temps (Courgeau et Lelièvre, 1989; Hill et al., 1991). Les modèles, exponentiel, de Weibull, de Gompertz et de Cox sont de ce type (modèles à risques proportionnels). Les individus soumis au risque connaissent l'événement d'intérêt les uns après les autres s'ils le connaissent (deux personnes ou plus ne peuvent pas connaître l'événement d'intérêt au même moment). Ceci peut poser un problème majeur, car dans la réalité le temps est mesuré de manière discrétisée ; ce qui est compatible avec le fait que plusieurs personnes peuvent connaître l'événement d'intérêt simultanément (on parle alors de groupes d'égalité ou de *ties* en anglais). Les durées d'occurrence des événements peuvent être mesurées en mois, années ou décades (Allison, 1982).

Lorsque l'unité de temps considérée est grande, en l'occurrence l'année, la probabilité d'obtenir des groupes d'égalité devient aussi grande (probabilité pour deux ou plusieurs personnes de connaître l'événement d'intérêt au même moment). Lorsque 5% des individus connaissent l'événement d'intérêt au même moment, les modèles paramétriques et le modèle semi-paramétrique de Cox peuvent fournir une estimation biaisée des paramètres (Yamagushi, 1991; Vermunt, 1997). Nous reviendrons sur cet aspect lorsqu'il s'agira de répondre aux questions de recherche.

Dans les modèles à risques proportionnels, le risque instantané s'écrit comme étant le produit de deux fonctions dont l'une dépend du temps $(h_0(t))$ et l'autre des caractéristiques individuelles $g(Z)$, Z étant la matrice des covariables.

D'où la fonction de risque qui caractérise ces modèles :

$$h(t, Z) = h_0(t)g(Z) \tag{4.1}$$

La deuxième hypothèse vient du fait que tous les modèles d'analyse de survie supposent l'indépendance des temps de censure et des temps de survie. Les autres hypothèses portent sur les modèles eux-mêmes à travers les suppositions qui seront faites sur la distribution de $h_0(t)$.

Les modèles de régression paramétriques se divisent en deux groupes de modèles qui sont : les modèles à risques proportionnels et les modèles à temps de sorties accélérées. Dans les modèles à risques proportionnels, les caractéristiques individuelles agissent sur la fonction de risque alors que dans les modèles à temps de sorties accélérées, elles agissent sur la fonction de séjour. Ces deux groupes de modèles sont présentés dans les sections suivantes.

4.1 Modèles à risques proportionnels (HP)

Dans les modèles à risques proportionnels, on fait l'hypothèse que toutes les caractéristiques individuelles agissent de manière multiplicative tout au long du temps sur une fonction de risque qui est la même pour l'ensemble de la population (Courgeau et Lelièvre, 1989). Les quotients individuels sont proportionnels entre eux indépendamment de la durée écoulée. Le quotient instantané pour un individu ayant les caractéristiques **Z** s'écrira de la manière suivante.

$$h(t; z_j) = h_o(t)e^{(Z_j\beta_z)} \tag{4.2}$$

Z_j représentant la matrice des prédicteurs et β_z le vecteur colonne des paramètres du modèle.

Le produit matriciel $Z_j\beta_z$ s'écrit :

$$Z_j\beta_z = z_1\beta_1 + z_2\beta_2 + ... + z_n\beta_n$$

On peut donc écrire l'équation 4.2 de la manière suivante :

$$ln\left[h(t; z_j)\right] = ln\left[h_o(t)\right] + z_1\beta_1 + z_2\beta_2 + ... + z_n\beta_n \tag{4.3}$$

Dans la plupart des modèles qui s'écrivent comme une combinaison linéaire des variables explicatives, nous avons un terme d'erreur, c'est-à-dire un écart entre ce qu'on a observé et ce qu'on a prédit. Dans l'équation 4.3, on remarque que le risque est une combinaison linéaire des variables explicatives mais il n'y pas de terme d'erreur (Allison, 2010). Cela s'explique par le fait que le risque n'est pas une **caractéristique observée** et cela est absolument nécessaire pour comprendre la suite des modèles. Nous parlerons donc du risque dans les modèles à risques proportionnels comme étant une pseudo variable dépendante et pour rappel, nous l'écrirons entre guillemets.

Lorsque la matrice Z_j est une matrice nulle, c'est-à-dire sans variables explicatives, le quotient instantané s'écrira comme suit :

$$h(t; 0) = h_o(t) \tag{4.4}$$

Si la variable $z_1 = 1$ alors que toutes les autres variables sont nulles, on peut écrire :

$$h(t; z_1) = h_0(t)e^{\beta_1} \tag{4.5}$$

Le rapport des équations 4.3 et 4.4 est constant et indépendant du temps :

$$\frac{h(t; z_1)}{h(t; 0)} = e^{\beta_1}$$

$\Longleftrightarrow$

$$h(t; z_1) = e^{\beta_1}h(t; 0)$$

On peut en déduire la proportionnalité des risques : le risque d'occurrence de l'événement étudié pour l'individu possédant la caractéristique individuelle $z_1 = 1$ est égal à une constante qui vaut e^{β_1} multipliée par le risque d'occurrence de l'événement étudié pour un individu n'ayant pas cette caractéristique individuelle.

Le paramètre β_i mesure l'effet de la variable z_i sur le risque instantané ; il est cependant plus simple d'interpéter e^{β_i} comme un risque relatif (Courgeau et Lelièvre, 1989). L'interprétation du paramètre β_i dépend de la nature de la variable à laquelle ce paramètre est associé. Lorsque la variable associée à β_i est quantitative, on pourra dire par exemple que lorsque z_i augmente d'une unité, le risque instantané pour un individu de connaître l'événement étudié augmente d'environ e^{β_i} fois, toutes choses étant égales par ailleurs. Lorsque $\beta_i > 0$, la

variable exerce un effet positif sur le risque, lorsque $\beta_i < 0$, l'impact sur le risque est négatif, et lorsque $\beta_i = 0$, il n'y a pas d'impact de la variable sur le risque d'occurrence de l'événement étudié. Lorsque la variable z_i associée à β_i est catégorielle, le paramètre β_i mesure l'impact de cette variable sur le risque d'occurrence de l'événement étudié lorsque nous appartenons à cette catégorie par rapport au fait d'appartenir à l'autre catégorie.

Comme nous venons de le voir, dans les modèles à risques proportionnels, la variable dépendante est le risque. On parle donc du risque d'occurrence d'un événement donné en fonction des caractéristiques individuelles. Selon la nature de l'événement d'intérêt et le domaine d'application, la formulation de ces modèles peut paraître un peu « bizarre ».

Pour le montrer, sortons du cadre de cette recherche et prenons comme événement d'intérêt l'occurrence d'un accident de la route au cours du temps en fonction de l'âge, du sexe et de la consommation de substances illicites. Les modèles à risques proportionnels nous imposent de formuler notre question de recherche de la manière suivante : le risque d'occurrence d'un accident de la route au cours du temps dépend de l'âge, du sexe et de la consommation de substances illicites. Il n'y a rien de dérangeant dans cette formulation de la question de recherche car la variable dépendante est le risque d'occurrence d'un accident et les variables indépendantes sont : l'âge, le sexe et la consommation de substances illicites.

Revenons maintenant dans le cadre de la mobilité des étudiants et supposons que l'on s'intéresse à l'obtention du diplôme de bachelor par des étudiants internationaux entre l'année d'obtention du baccalauréat et cinq ans après l'obtention de celui-ci en fonction de l'âge, du sexe et de la nationalité des étudiants. Cette question de recherche pourrait être formulée de la manière suivante en utilisant les modèles à risques proportionnels : le risque d'obtention du diplôme de bachelor par les étudiants internationaux dépend de l'âge, du sexe et de la nationalité de ces étudiants.

Dans cette formulation de notre question de recherche, il n'y a aucun problème au niveau des variables explicatives qui sont clairement définies. Par contre parler de risque dans l'obtention d'un diplôme peut paraître « bizarre ». Parler de probabilité d'obtenir un diplôme en fonction de l'âge, du sexe et de la nationalité peut paraître plus logique ; mais le problème réside dans le fait qu'une probabilité est comprise dans l'intervalle $[0, 1]$ alors que le risque tel que spécifié par ces modèles peut être supérieur à 1 et ne peut donc pas être une probabilité.

Nous allons donc préférer ce que les modèles nous imposent en parlant de risque d'obtenir un diplôme à ce qui nous paraît être plus approprié mais dépourvu de fondement scientifique. On parlera dès lors de risque de se marier, ou de risque de terminer les études ou risque d'abandon des études, etc.

Dans le deuxième groupe de modèles, les caractéristiques individuelles agissent sur la fonction de séjour. Ce groupe de modèles porte le nom de modèle à temps de sorties accélérées.

4.2 Modèles à temps de sorties accélérées (AFT)

Dans les modèles à temps de sorties accélérées (AFT en anglais pour *Accelerated Failure Time model*) ou modèles ln(temps), on suppose que les caractéristiques individuelles agissent sur les fonctions de séjour au lieu d'agir directement sur les fonctions de risque instantané. Dans ces modèles, l'équation estimée porte sur le logarithme du temps de survie.

Ces modèles s'écrivent de la manière suivante :

$$ln(t_j) = Z_j\beta_z + \varepsilon_j$$

On utilise le mot accéléré car dans ces modèles, au lieu de supposer que le temps suit une distribution donnée, nous allons introduire une paramètre τ_j et écrire la forme générale de ces modèles comme suit :

$$\tau_j = exp(-Z_j\beta_z)t_j \tag{4.6}$$

L'expression $exp(-Z_j\beta_z)$ est appelée facteur d'accélération.

Les situations suivantes peuvent se présenter :

- Si $exp(-Z_j\beta_z) = 1$ alors $\tau_j = t_j$ et le temps s'écoule à son rythme normal.
- Si $exp(-Z_j\beta_z) > 1$ alors $\tau_j > t_j$ et le temps s'écoule plus vite pour cet individu car le temps est accéléré [3] et on pourra s'attendre à ce que les événénements étudiés se produisent plus vite.
- Si $exp(-Z_j\beta_z) < 1$ alors $\tau_j < t_j$, le temps passe plus lentement pour cet individu car le temps est décéléré [4] et on peut s'attendre à ce que les événements étudiés se produisent tardivement.

En effet, l'obtention des modèles à temps de sorties accélérées est simple :

Comme $\tau_j = exp(-Z_j\beta_z)t_j$ on en déduit que $t_j = exp(Z_j\beta_z)\tau_j$.

En prenant le lograrithme de t_j, on obtient le modèle suivant :

$$ln(t_j) = Z_j\beta_z + ln(\tau_j) \tag{4.7}$$

C'est l'expression générale des modèles à temps de sorties accélérées, et pour chaque modèle, nous allons nous intéresser à la distribution de la quantité $ln(\tau_j)$; cette quantité étant aléatoire (Cleves et al., 2010). Dans les modèles à temps de sorties accélérées, la variable dépendante qui est t_j ou $ln(t_j)$ est une caractéristique observée d'où l'existence d'un terme d'erreur.

Par soucis de simplification, on s'intéressera à la distribution de τ_j au lieu de s'intéresser directement à celle de $ln(\tau_j)$.

On dira par exemple que τ_j suit une distribution exponentielle pour la loi exponentielle, τ_j suit une loi log-normale pour la loi log-normale ou τ_j suit une loi de Weibull pour la distribution de Weibull. Les modèles log-normal, log-logistique et gamma appartiennent à cette famille de modèles, tandis que le modèle exponentiel et le modèle de Weibull appartiennent à la fois aux modèles à HP et aux modèles à AFT.

Le tableau 4.1 résume les différents modèles selon qu'ils soient à risques proportionnels (HP) ou à temps de sorties accélérées (AFT).

Tableau 4.1 – Les modèles paramétriques à HP et à temps de sorties accélérées (AFT)

Distribution	HP	AFT
Exponentielle	Oui	Oui
Weibull	Oui	Oui
Gompertz	Oui	Non
Lognormale	Non	Oui
Loglogistique	Non	Oui
Gamma généralisée	Non	Oui

Pour les modèles exponentiel et Weibull, le passage d'un modèle à HP à un modèle en AFT se fait simplement en procédant à quelques tranformations sur les coefficients estimés. Pour le modèle exponentiel par exemple, on passe d'un modèle à HP à un modèle en AFT en changeant simplement les signes des coefficients obtenus par la régression dans le modèle à HP et vice versa (Cleves et al., 2010; Allison, 2010).

Par exemple, si on obtient un paramètre $\beta = 1.5$ pour une variable dans un modèle à HP, ce paramètre β sera égal à -1.5 en AFT pour la même variable.

$$\beta_{HP} = -\beta_{AFT} \tag{4.8}$$

3. Lorsque le risque d'occurrence des événenements étudiés est élevé, ceux-ci se produisent vite et la durée de séjour est réduite ; on parle de temps accéléré.

4. Lorsque le risque d'occurrence des événements étudiés est faible, ceux-ci se produisent rarement et la durée de séjour augmente ; on parle de temps décéléré.

Pour le modèle de Weibull, le passage d'un modèle à AFT à un modèle à HP se fait en changeant les signes des paramètres obtenus et en les divisant par la pente (le paramètre) de la loi de Weibull (Cleves et al., 2010; Allison, 2010).

De manière générale, pour la régression de Weibull, les modèles AFT et les modèles PH sont reliés par la relation suivante :

$$\beta_{AFT} = \frac{-\beta_{HP}}{p} \tag{4.9}$$

$\Longleftrightarrow$

$$\beta_{HP} = -p\beta_{AFT} \tag{4.10}$$

p représentant le paramètre de forme de la loi de Weibull.

Pour mettre en application ces modèles, les données doivent être organisées dans des formats qui leur sont compatibles. Les différents modèles seront présentés dans les chapitres à suivants. Nous commencerons d'abord par présenter les modèles à risques proportionnels ensuite nous présenterons les modèles à temps de sorties accélérées. Le premier modèle à risques proportionnels qui sera présenté est le modèle exponentiel.

Chapitre 5

Le modèle de régression exponentielle

Dans ce chapitre, le modèle de régression exponentielle sera présenté ainsi qu'un exemple d'application avec le logiciel Stata. Par soucis de simplification et pour ne pas surcharger les explications de formules mathématiques, les analyses seront axées seulement sur certaines fonctions permettant d'étudier la distribution de la durée telles que la fonction de risque, de risque cumulé ou la fonction de séjour.

Dans une régression paramétrique sur des données biographiques individuelles en général, et, particulièrement dans la régression exponentielle, la « variable dépendante » est le risque ou le logarithme du risque selon la manière dont le modèle est spécifié. Il faut cependant préciser que cette « variable dépendante » n'est pas une caratéristique mesurée. L'estimation des paramètres est faite en maximisant la vraisemblance du modèle. Les coefficients ainsi estimés nous permettent de mesurer l'impact des variables explicatives sur les risques d'occurrence de l'événement d'intérêt ou sur le logarithme du risque. Le rapport des risques pour deux individus ayant des caractéristiques individuelles différentes est constant et indépendant du temps. Ce rapport représente la proportionnalité des risques entre les deux individus. Lorsqu'on augmente une variable explicative d'une unité, la proportionnalité des risques entre deux individus est multipliée par l'exponentielle du coefficient correspondant à cette variable. Les paramètres estimés permettent de mesurer l'effet des variables explicatives sur le risque d'occurrence de l'événement étudié. Lorsqu'un paramètre est positif, cela veut dire que la variable associée à ce paramètre exerce un effet positif sur le risque et inversément : si le paramètre est négatif, cela veut dire que la variable qui lui est associée exerce un effet négatif sur le risque d'occurrence de l'événement d'intérêt. Lorsque nous avons une variable explicative qui est catégorielle, le coefficient qui lui est associé indique le risque d'occurrence de l'événement d'intérêt lorsque nous appartenons à cette catégorie plutôt qu'à l'autre.

Dans la régression exponentielle, le risque d'occurrence de l'événement étudié et les caractéristiques individuelles sont liés de manière multiplicative ; le risque instantané s'écrit donc comme étant le produit de deux fonctions dont l'une dépend du temps $(h_0(t))$ et l'autre des caractéristiques individuelles $g(Z)$, Z étant la matrice des covariables.

D'où la fonction de risque qui caractérise ces modèles :

$$h(t, Z) = h_0(t)g(Z) \tag{5.1}$$

Avec :

$$h_0(t) = e^{\beta_0} \quad et \quad g(Z) = e^{(Z_j\beta_z)}$$

On peut alors écrire :

$$h(t; Z_j) = h_o(t)e^{(Z_j\beta_z)} = e^{(\beta_0)}e^{(Z_j\beta_z)} = e^{(\beta_0 + Z_j\beta_z)}$$

Donc :

$$h(t; Z_j) = e^{\beta_0+\beta_1 Z_1+\beta_2 Z_2+\cdots+\beta_k Z_k}$$

Ce qui équivaut, en prenant le logarithme de la fonction de risque à :

$$log\left[h(t, Z_j)\right] = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_k Z_k \tag{5.2}$$

C'est le modèle paramétrique le plus simple avec une fonction de risque de base ($h_o(t)$) qui est constante et indépendante du temps. La fonction de risque de base pour tout t, lorsque les caractéristiques individuelles sont nulles s'écrit :

$$h(t, 0) = \beta_0 \tag{5.3}$$

Cette fonction de risque est constante quel que soit le temps t auquel on se trouve. On observe donc que dans le modèle exponentiel de base, on a un seul paramètre à estimer qui est β_0.

Pour tester la significativité des paramètres estimés, la méthode du rapport de vraisemblance a été privilégiée parce qu'elle est plus robuste que les autres tests proposés (Wald, score) lorsque nous avons de petits échantillons. Le principe du test de rapport de vraisemblance consiste dans un premier temps à ajuster un modèle avec toutes les variables explicatives de la question de recherche. Ce modèle complet va générer une déviance qu'on va appeler la déviance du modèle complet (DEV_M) ; cette déviance ne changera pas.

Par exemple, pour tester la significativité d'une variable x_1 de la question de recherche, on va encore ajuster un deuxième modèle sans cette variable x_1 dont on veut tester la significativité ; on obtiendra une déviance qui correspond à ce modèle sans la variable x_1 qu'on va appeler par exemple DEV_1.

La variable x_1 est significative si :

$$DEV_1 - DEV_M > \chi_1^2 = 3.84$$

Mais avant d'ajuster le modèle, il est d'abord nécessaire de préparer la base de données dans le format adéquat.

5.1 Format des données

La préparation des données dans le but de faire une régression exponentielle sur des données biographiques individuelles se fait de la même manière que pour faire une analyse non paramétrique. Nous avons besoin d'un identifiant unique pour chaque individu, d'un instant de début d'observation, d'un instant de fin d'observation, d'une variable qui va nous informer si un individu a connu ou pas l'événement d'intérêt et des variables explicatives, parce que nous sommes dans un modèle prédictif. La base de données ressemble à la structure qui est présentée dans le Tableau 5.1.

Tableau 5.1 – Format de la base de données pour faire une régression exponentielle

Identifiant	Début	Fin	Evénement	x_{1i}	x_{2i}
1	date début	date fin	1	x_{11}	x_{21}
2	date début	date fin	1	x_{12}	x_{22}
⋮	⋮	⋮	⋮	⋮	⋮

Dans ce tableau, la première colonne représente l'identifiant des individus, les deux colonnes suivantes représentent respectivement les dates de début et de fin d'observation, la colonne événement est une variable binaire qui prend la valeur 1 si l'individu a connu l'événement d'intérêt et la valeur 0 sinon, les deux dernières colonnes (x_{1i} et x_{2i}) représentent des variables explicatives.

5.2 Approche graphique

Lorsque nous avons des données à disposition, on peut tester de la manière suivante si le modèle exponentiel est adapté ou pas.

1. On représente les données dans un graphique en mettant sur l'axe des ordonnées la fonction de risque et sur l'axe des abscisses le temps. Si le modèle exponentiel est vérifié, on obtiendra une droite horizontale qui est parallèle à l'axe des temps (axe des abscisses). Le graphique ci-dessous (figure 5.1)représente la fonction de risque d'une distribution exponentielle de paramètre $\beta_0 = 0.2$.
2. On représente les données dans un graphique en mettant sur l'axe des ordonnées le logarithme de la fonction de séjour, et sur l'axe des abscisses le temps. Si le modèle exponentiel est vérifié, on obtiendra une courbe décroissante du temps.

Comme on peut le remarquer sur le graphique 5.1, la fonction de risque de paramètre $\beta_0 = 0.2$ est une droite parallèle à l'axe des abscisses quelle que soit la valeur prise par la variable temps.

La courbe de la fonction de séjour ($\beta_0 = 0.4$), elle, est décroissante du temps, ses valeurs sont grandes pour de petites valeurs du temps et petites pour de grandes valeurs du temps.

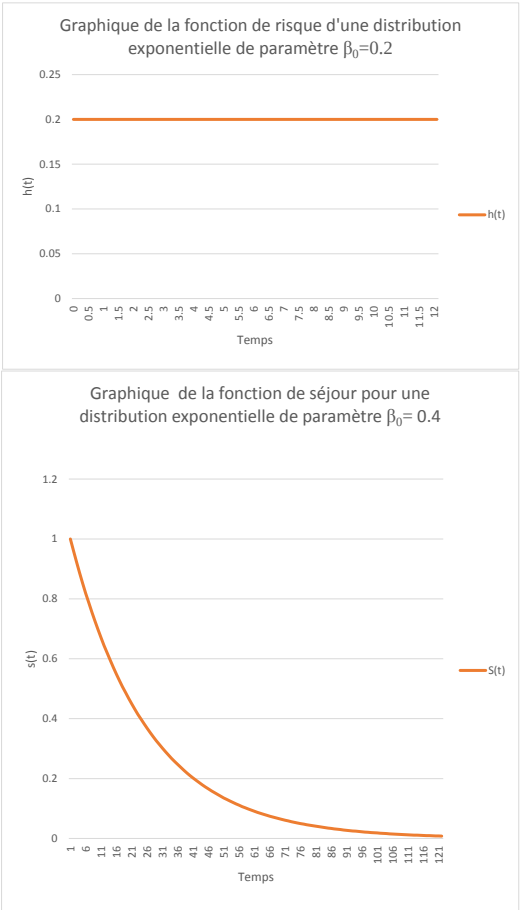


FIGURE 5.1 – Fonction de risque et fonction de séjour d'une distribution exponentielle de paramètres respectifs $\beta_0 = 0.2$ et $\beta_0 = 0.4$

5.3 Exemple d'application

Dans cette section, l'objectif est de mettre en pratique un modèle de régression exponentielle à partir d'une question de recherche. La base de données sur laquelle l'exemple est basé sera présentée et des analyses seront faites dans le but de répondre à la question de recherche.

Question de recherche : Le « risque » d'obtenir le diplôme de bachelor (dans notre cas c'est en fait la chance d'obtenir le diplôme de bachelor) des étudiants internationaux dépend-elle du sexe et de l'âge de ces étudiants ?

5.3.1 Données

Pour répondre à cette question de recherche, nous disposons d'une base de données [1] dans laquelle nous avons l'année de début du bachelor, l'année d'obtention du diplôme de bachelor, d'une variable événement qui prend la valeur 1 si l'individu a obtenu son bachelor, 0 sinon, nous disposons également des variables explicatives âge et sexe (1=femme, 0=hommes) des étudiants. Comme pour les modèles non paramétriques, la raison de ce choix est le retard pris dans la livraison des données par l'OFS.

L'analyse exploratoire de la variable sexe présentée dans la figure 5.2 montre que nous avons plus de femmes que d'hommes avec des parts respectives de 67.98% et 32.08%.

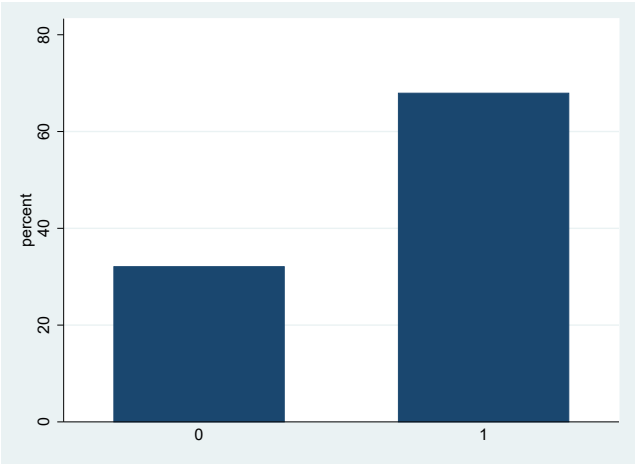


FIGURE 5.2 – Diagramme en barre de la variable sexe

La boîte à moustaches de la variable âge présentée dans le figure 5.3 montre que la distribution de cette variable est plus ou moins symétrique [2] et nous n'avons pas de données extrêmes.

Le tableau ci-dessous montre les résultats de la régression exponentielle et, comme on peut le remarquer, la valeur du paramètre estimé $\widehat{\beta_0} = -7.893$.

Variable	Coeff.	Erreur standard	Z	P-valeur	IC à 95%
sexe	-1.688958	0.3703357	-4.56	0.000	-2.414803 -0.9631137
âge	0.0809663	0.0342787	2.36	0.018	0.0137813 0.1481514
constante	-7.892737	2.458841	-3.21	0.001	-12.71198 -3.073498

Le tableau des coefficients montre que lorsque l'âge augmente d'une unité, le risque d'obtenir le diplôme de bachelor augmente d'environ 1.084 ($e^{0.0809}$) fois ; ce qui correspond aussi à une augmentation du risque de 8.4%. Pour le même âge, être une femme plutôt qu'un homme ferait diminuer [3] le risque d'obtenir le diplôme de bachelor de 81.53%.

1. La base de données provient du livre « An introduction to Survival Analysis Using Stata », troisième édition (Cleves et all., 2010). Des variables ont été renomées en vue de créer une question de recherche en lien avec la mobilité des étudiants.
2. La distribution est symétrique si la médiane est au milieu de la boîte à moustache
3. Cette diminution est obtenue en faisant le calcul suivant : $(odds-1)*100$, qui correspond dans notre cas à $(e^{-1.6889}-1)*100 = -81.53\%$. Le signe négatif implique une diminution.

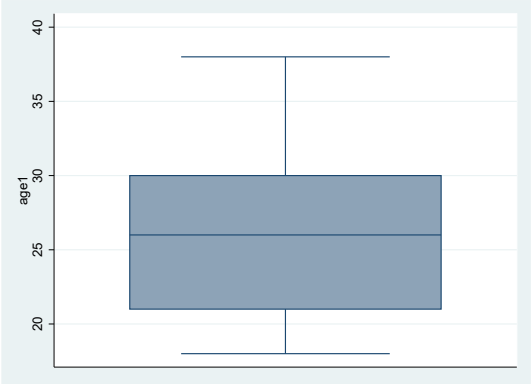


FIGURE 5.3 – Boîte à moustaches de la variable âge

On peut écrire la fonction de risque pour tout t en y intégrant les coefficients estimés du modèle ainsi que les variables explicatives.

La fonction de risque estimée lorsque les caractéristiques individuelles sont nulles s'écrit :

$$\widehat{h_0(t)} = e^{\beta_0} = e^{-7.892} = 0.000373$$

La fonction de risque estimée pour tout t s'écrit :

$$log\left[\widehat{h(t, Z_j)}\right] = -7.892 - 1.689sexe + 0.081age$$

On peut tester la significativité des coefficients du modèle en posant les hypothèses nulle et alternative suivantes :

$H_0 : \beta_j = 0 \ \ vs \ \ H_1 : \beta_j \neq 0, j = 1, 2, \cdots k.$

La statistique de test associée suit une loi de Chi-carré à un degré de liberté sous H_0. Le tableau ci-dessous résume le test statistique par la méthode du rapport de vraisemblance.

Modèles	Déviances	Différence de déviances	Comparaison	Conclusion
Complet	47.534			
Sans l'âge	50.277	50.277 - 47.534 = 2.743	2.743 < 3.84	Non rejet de H_0 au seuil de 5%
Sans le sexe	58.075	58.075 - 47.534 = 10.541	10.541 > 3.84	On rejette H_0 au seuil de 5%

La variable âge s'est avérée non significative au seuil de 5% alors que la variable sexe est siginificative parce que la différence de déviance pour la variable âge est inférieure à 3.84 alors que cette différence de déviance est supérieure à 3.84 pour la variable sexe. On peut conclure que notre question de recherche est partiellement corroborée par le test du rapport de vraisemblance. Comme on peut l'observer dans le tableau des coefficients, toutes les p-valeurs associées aux coefficients du modèle sont inférieures à 0.05, ce qui nous permet de rejeter aussi H_0 au seuil de 5% pour tous les paramètres du modèle et de conclure ainsi que l'hypothèse de recherche est corroborée par le test de Wald. Les deux tests donnent donc des résultats contradictoires pour ce modèle ; le test que nous privilégierons dans cet exemple est le test du rapport de vraisemblance car le test de Wald peut conduire à des résultats erronés lorsque nous avons de petits échantillons comme c'est le cas dans cet exemple. Lorsque nous avons de grands échantillons, nous privilégierons le test de Wald qui est un test asymptotique, c'est-à-dire que plus la taille de l'échantillon est grande plus ce test est précis.

Dans le modèle de régression exponentielle, la fonction de risque de base est constante quel que soit le temps. Ce modèle repose donc sur une hypothèse très forte qui est celle de dire que le risque d'occurrence d'un événement ne dépend pas du temps. Mais dans la réalité, il est possible que le risque évolue avec le temps.

Cela nous pousse à penser par exemple que le risque de décès ne dépend pas de l'âge, c'est-à-dire, en d'autres termes qu'il n'y a pas de vieillissement. Ce qui représente une hypothèse forte dans le sens où nous savons qu'en réalité le risque de mortalité augmente avec l'âge. Les exemples les plus fréquents viennent du domaine de la médecine ; la probabilité de développer un cancer de la prostate par exemple augmente avec l'âge, les troubles de mémoire augmentent aussi avec l'âge. Il est dès lors nécessaire de penser à d'autres types de modèles pour lesquels le risque n'est pas constant au cours du temps. Le modèle de Weibull qui est présenté dans la section suivante fait partie de ce genre de modèle. Comme nous le verrons, dans ce modèle, le risque peut varier de manière monotone croissante ou décroissante.

Chapitre 6

Le modèle de régression de Weibull

Dans le modèle de Weibull, les variables explicatives exercent un effet multiplicatif sur la fonction de risque comme c'est le cas dans les modèles à hasard proportionnel. Le modèle de Weibull appartient à la fois à la famille des modèles à hasard proportionnel et à celle des modèles à temps de sorties accélérées (Zhang, 2016; Cleves et al., 2010).

Le modèle de base, c'est-à-dire le modèle sans variables explicatives s'écrit de la manière suivante :

$$h_0(t) = pt^{p-1}exp(\beta_0)$$

Avec p et β_0 étant les deux paramètres du modèle à estimer.

A l'image du modèle exponentiel, le risque d'occurrence de l'événement d'intérêt et les caractéristiques individuelles sont liés de manière multiplicative ; on parle alors de proportionnalité des risques. Le risque d'occurrence de l'événement étudié en fonction du temps et des variables explicatives s'écrit comme étant le produit de deux fonctions dont l'une $(h_0(t))$ dépend du temps et l'autre $(Z_j\beta_z)$ des caractéristiques individuelles, avec Z_j, la matrice des variables explicatives et β_z, le vecteur colonne des paramètres du modèle.

La fonction de hasard de ce modèle s'écrit :

$$h(t; Z_j) = h_0(t)exp(Z_j\beta_z) = pt^{p-1}exp(\beta_0 + Z_j\beta_z) \tag{6.1}$$

En prenant le logarithme de la fonction de risque de base, on obtient :

$$ln(h_0(t)) = ln(p) + (p-1)ln(t) + \beta_0$$

$ln(p)$ et β_0 étant constants et indépendants du temps, le signe de $ln(h_0(t))$ dépend du signe de $(p-1)$ qui représente la pente.

En posant que la constante c = $ln(p)$ + β_0, on peut écrire le logarithme de la fonction de risque de base comme :

$$ln(h_0(t)) = c + (p-1)ln(t)$$

De la même manière, on peut réécrire le logarithme de la fonction de risque pour tout t comme suit :

$$ln\left(h(t; Z_j)\right) = ln(p) + (p-1)ln(t) + \beta_0 + Z_j\beta_z$$

Dans ce modèle aussi, la « variable dépendante » est le risque ou le logarithme du risque.

En posant $C = ln(p) + \beta_0 + Z_j\beta_z$, on obtient :

$$lnh\left((t; Z_j)\right) = C + (p-1)ln(t) \tag{6.2}$$

C étant une constante quelconque

Les situations suivantes se présentent en fonction de la valeur de p.

1. Si $p = 1$, la fonction de hasard est constante et le modèle de Weibull n'est autre que le modèle exponentiel.
2. Si $p < 1$, la fonction de hasard est une fonction monotone décroissante, car $p - 1$ devient négatif ;
3. Si $p > 1$, la fonction de hasard est une fonction monotone croissante, car $p - 1$ devient positif.

6.1 Approche graphique

Pour mieux comprendre l'aspect graphique de la distribution de Weibull, il est nécessaire, dans un premier temps d'écrire les fonctions de risque cumulé et de séjour. Nous procéderons ensuite à quelques légères transformations logarithmiques dans le but de simplifier ces fonctions le plus possible en vue d'une représentation graphique moins complexe et familière. La fonction de séjour s'écrit de la manière suivante :

$$S(t_j; Z_j) = exp\left[-exp(\beta_0 + Z_j\beta_z)t^p\right] \tag{6.3}$$

Après quelques transformations algébriques, cette fonction de séjour peut s'écrire de la manière suivante :

$$ln\left[-ln(S(t_j; Z_j))\right] = \beta_0 + Z_j\beta_z + pln(t) = constante + pln(t) \tag{6.4}$$

En posant que $constante$ = $\beta_0 + Z_j\beta_z$.

Dès lors, il est possible de discuter de la représentation graphique de ces différentes fonctions selon le signe du paramètre p.

6.1.1 Représentation graphique de la fonction de risque

Nous avons vu précédemment que la fonction de risque de la distribution de Weibull pouvait être soit constante, soit monotone croissante ou encore monotone décroissante selon la valeur et le signe de p.

Cas où $p = 1$

Dans le cas où $p = 1$, en faisant une représentation graphique dans laquelle on met sur l'axe des ordonnées $ln\left(h(t; Z_j)\right)$ et sur l'axe des abscisses le temps, on obtiendra une droite constante qui est parallèle à l'axe des abscisses comme dans le cas de la régression exponentielle. Dans ce cas, la distribution de Weibull n'est autre que la distribution exponentielle.

Cas où $p > 1$

Dans le cas où $p > 1$, en faisant une représentation graphique dans laquelle on met sur l'axe des ordonnées $ln\left(h(t; Z_j)\right)$ et sur l'axe des abscisses le temps, on obtiendra une courbe monotone [1] croissante dans la mesure où $p - 1$ sera positif.

Cas où $p < 1$

Dans le cas où $p < 1$, en faisant une représentation graphique dans laquelle on met sur l'axe des ordonnées $ln\left(h(t; Z_j)\right)$ et sur l'axe des abscisses le temps, on obtiendra une courbe monotone décroissante dans la mesure où $p - 1$ sera négatif.

1. Un courbe est dite monotone croissante ou décroissante sur un intervalle donné lorsqu'elle est strictement croissante ou décroissante sur cet intervalle

6.1.2 Représentation graphique de la fonction de séjour

L'allure du graphique de la fonction de séjour dépend aussi du signe du paramètre p. En effet, on a aussi vu précédemment qu'en procédant à quelques transformations logarithmiques de la fonction de séjour, on obtenait :

$$ln\left[-ln(S(t_j; Z_j))\right] = constante + pln(t)$$

En faisant une représentation graphique dans laquelle on met sur l'axe des ordonnées $ln\left[-ln(S(t_j; Z_j))\right]$ et sur l'axe des abscisses le temps ou $ln(t)$, on obtiendra soit une droite constante, soit une droite croissante ou une droite décroissante.

Cas où $p = 0$

Dans ce cas, le graphique de $ln\left[-ln(S(t_j; Z_j))\right]$ est une droite parallèle à l'axe du temps, et la distribution obtenue n'est autre que la distribution exponentielle.

Cas où $p > 0$

Dans ce cas, en mettant $ln\left[-ln(S(t_j; Z_j))\right]$ sur l'axe des ordonnées et le logarithme du temps ou le temps sur l'axe des abscisses, on obtient une courbe croissante comme le montre le graphique 6.1.

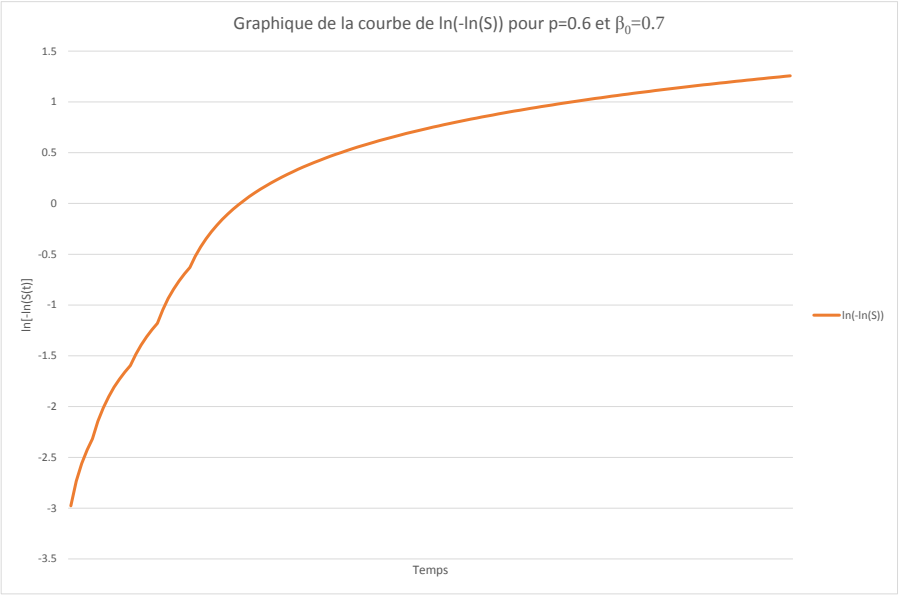


FIGURE 6.1 – Graphique de $ln[-ln(S(t)]$ d'une distribution de Weibull de paramètres $p = 0.6$ et $\beta_0 = 0.7$

Cas où $p < 0$

Dans cette situation, en mettant $ln\left[-ln(S(t_j; Z_j))\right]$ sur l'axe des ordonnées et le logarithme du temps ou le temps sur l'axe des abscisses, on obtient une courbe décroissante comme le montre le graphique 6.2.

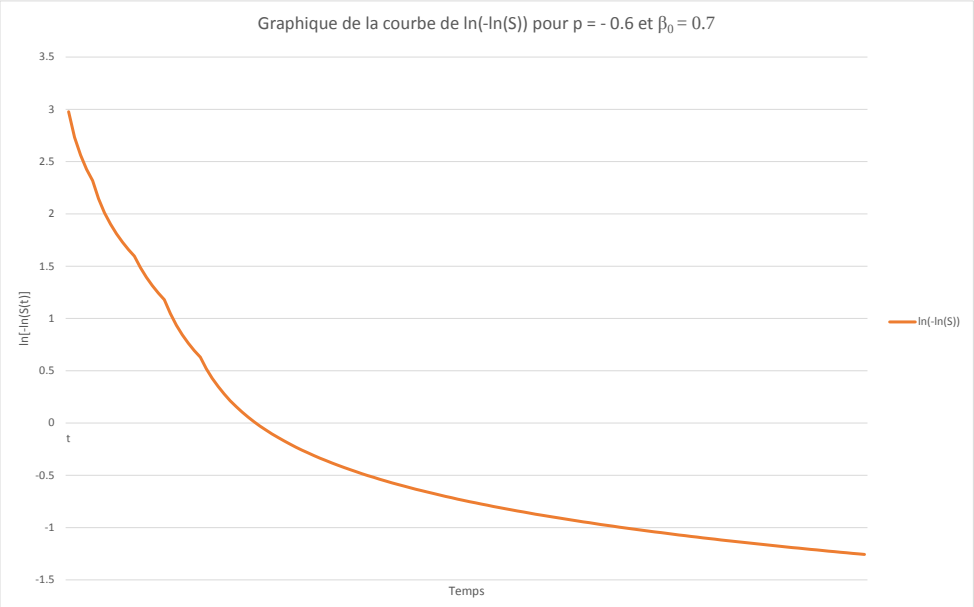


FIGURE 6.2 – Graphique de $ln[-ln(S(t)]$ d'une distribution de Weibull de paramètres $p = -0.6$ et $\beta_0 = 0.7$

Par ailleurs, la fonction de risque cumulé $H(t_j; Z_j) = -ln(S(t_j; Z_j))$, on peut aussi utiliser cette fonction pour tester graphiquement si les données peuvent être analysées avec une distribution de Weibull.

La distribution de Weibull convient à la modélisation de données ayant une fonction de hasard présentant une allure monotone (Cleves et al., 2010; Blossfeld et al., 2009). Si nous avons des données à disposition, pour tester l'adéquation des données au modèle de Weibull, on prodédera comme suit : on fera un graphique dans lequel on portera sur l'axe des ordonnées soit le logarithme de la fonction de risque, soit la fonction $ln\,[-ln(S(t_j; Z_j))]$ ou H(t) et sur l'axe des abscisses le temps. Si le modèle de Weibull est vérifié, le graphique donnera une allure monotone croissante ou décroissante.

6.2 Format des données en vue d'une régression de Weibull

La préparation des données dans le but de faire une régression de Weibull sur des données biographiques individuelles se fait de la même manière que pour la préparation des données en vue d'une régression exponentielle.

Nous avons besoin d'un identifiant unique pour chaque individu, d'un instant de début d'observation, d'un instant de fin d'observation, d'une variable qui va nous informer si un individu a connu ou non l'événement d'intérêt et des variables explicatives puisque nous sommes aussi dans un modèle prédictif. La base de données présentée dans le graphique 6.1 aura une structure identique à celle de la régression exponentielle ; pour sa description, se référer à la section 5.1 de la régression exponentielle.

6.3 Exemple d'application d'une régression de Weibull

La question de recherche traitée dans cet exemple est la même que celle traitée dans la régression exponentielle. Cependant, on doit garder à l'esprit que le risque n'est pas spécifié de la même manière dans le modèle

Tableau 6.1 – Format de la base de données pour faire une régression de Weibull

Identifiant	Début	Fin	Evénement	x_{1i}	x_{2i}
1	date début	date fin	1	x_{11}	x_{21}
2	date début	date fin	1	x_{12}	x_{22}
⋮	⋮	⋮	⋮	⋮	⋮

exponentiel et dans le modèle de Weibull. Dans le modèle exponentiel, le risque de base est constant alors que dans le modèle de Weibull, il peut être monotone croissant ou monotone décroissant.

6.3.1 Données

Les données qui sont utilisées pour mettre en pratique la régression de Weibull sont les mêmes données que celles qui ont été utilisées dans l'exemple de la régression exponentielle. Le descriptif de la base de données ainsi que les analyses exploratoires sont les mêmes et ne seront pas présentées pour cet exemple (voir section 5.3.1).

En effectuant la régression de Weibull, on a obtenu les résultats qui sont présentés dans le Tableau 6.2.

Tableau 6.2 – Tableau des coefficients de la régression de Weibull

Variable	Hazard Ratio	Erreur standard	Z	P-valeur	IC à 95%
sexe	0.1099	0.4482	-5.42	0.001	.0494629 .2444548
âge	1.1171	0.4231	2.93	0.003	1.03726 1.203271
constante	8.54e-06	.0000248	-4.01	0.001	2.85e-08 .0025565
$/ln\_p$	.5188	.1376	3.77	0.001	.2490831 .7886556
p	1.6801	.2312		0.001	1.282849 2.200436
1/p	.5951	.08192			.4544553 .7795152

Dans le tableau des coefficients, on a les coefficients estimés ainsi que les tests statistiques qui leur sont associés d'une part, d'autre part on remarque aussi trois lignes supplémentaires par rapport au tableau de résultats de la régression exponentielle. Ces lignes supplémentaires sont : $/ln\_p$, p et $1/p$, et testent la même chose mais selon la manière dont le modèle est présenté. Nous discuterons seulement les deux premières présentations qui testent les hypothèses nulle et alternative suivantes :

— $H_0 : ln(p) = 0$
— $H_1 : ln(p) \neq 0$

qui équivaut à :

— $H_0 : p = 1$
— $H_1 : p \neq 1$

Pour rappel, dans la régression de Weibull, lorsque $p = 1$, la régression de Weibull n'est autre que la régression exponentielle qui est un modèle à risque constant. L'hypothèse nulle sur le paramètre p teste donc en termes de mots, H_0 : le risque suit une distribution exponentielle versus H_1, le risque ne suit pas une distribution exponentielle. La p-valeur associée à ce paramètre vaut 0.001 < 0.05, ce qui nous permet de rejetter H_0 au seuil de 5% et de dire que le modèle n'est pas à hasard constant et par conséquent le risque ne suit pas une distribution exponentielle.

Nous avons également effectué un test du rapport de vraisemblance dont les résultats sont présentés dans le Tableau 6.3. Les variables sexe et âge sont significatives car les différences de déviances qui leur sont associées sont toutes supérieures à 3.84 ; ce qui nous conduit à rejeter H_0 (que les coefficients associés aux variables âge et sexe sont nuls) au seuil de 5% pour les deux variables et de conclure que l'hypothèse de recherche est corroborée.

Tableau 6.3 – Tableau résumant le test du rapport de vraisemblance

Modèles	Déviances	Différence de déviances	Comparaison	Conclusion
Complet	41.992			
Sans l'âge	50.277	50.277 - 45.53 = 4.947	4.947 > 3.84	On rejette H_0 au seuil de 5%
Sans le sexe	58.075	58.075 - 45.53 = 12.545	12.545 >3.84	On rejette H_0 au seuil de 5%

Comme nous venons de le voir, dans le modèle exponentiel, le risque d'occurence de l'événement étudié est constant ; dans le modèle de Weibull, la fonction de risque de base peut aussi être constante, monotone croissante ou monotone décroissante [2].

Nous avons pu observer que le modèle exponentiel est un cas particulier du modèle de Weibull. Pour choisir un modèle entre les deux, on commencera par ajuster un modèle de Weibull, ensuite on procédera à un test d'hypothèses sur le paramètre de forme p de la distribution de Weibull. On pourra ainsi choisir le modèle adéquat entre la distribution de Weibull et la distribution exponentielle selon les résultats du test.

Le modèle de Weibul est utile pour analyser des phenomènes dont l'occurrennce augmente ou diminue de manière non linéaire au cours du temps. Nous savons ainsi par exemple que malgré les disparités entre les pays, le risque de mortalité infantile dans le monde diminue avec le temps tout comme le nombre de femmes qui décèdent en couche diminue aussi au cours du temps. Le modèle de Weibull est utile pour analyser ce genre de phénomènes.

On pourrait aussi imaginer que le risque puisse évoluer de manière linéaire avec le temps. Le modèle de Gompertz qui sera présenté dans la section suivante fait partie de ce genre de modèle. Une étude a montré par exemple qu'il existe une relation linéaire entre le risque de disparition des animaux et la destruction de leur habitat (Olivieri et Vitalis, 2001). Le modèle de Gompertz sera utile pour analyser par exemple ce genre de phénomènes. Dans ce modèle, le logarithme du risque de base est une fonction linéaire du temps, cette fonction pouvant être croissante, décroissante ou constante.

2. Une fonction est dite monotone croissante ou monotone décroissante sur un intervalle, si elle est strictement croissante ou strictement décroissnte sur cet intervalle

Chapitre 7

Le modèle de régression de Gompertz

La distribution de Gompertz est une distribution dont la fonction de risque de base s'écrit :

$$h_0(t) = exp(\gamma t)exp(\beta_0)$$

La fonction de hasard est de la forme suivante :

$$h(t; Z_j) = h_0(t)exp(Z_j\beta_z) = exp(\gamma t)exp(\beta_0 + Z_j\beta_z) \tag{7.1}$$

En prenant le logarithme de la fonction de risque de base $h_0(t)$, on obtient :

$$ln\left[h_0(t)\right] = \gamma t + \beta_0$$

Le signe de cette fonction dépend du signe de γ qui représente la pente.

En prenant le logarithme de la fonction de risque pour tout t en fonction des caractéristiques individuelles on obtient :

$$ln\left[h(t; Z_j)\right] = \gamma t + \beta_0 + Z_j\beta_z$$

En posant $c = \beta_0 + Z_j\beta_z$, on peut écrire l'équation ci-dessus comme :

$$ln\left[h(t; Z_j)\right] = c + \gamma t \tag{7.2}$$

Dans la distribution de Gompertz, on remarque que le signe de la fonction de risque pour tout t dépend du signe du paramètre de forme γ ; les situations suivantes se présentent :

1. Lorsque $\gamma > 0$ (positif), la fonction de risque est une fonction croissante du temps (droite croissante de pente γ). Cela veut dire que lorsqu'on la représente dans un graphique, en mettant le logarithme de la fonction de risque sur l'axe des ordonnées et le temps sur l'axe des abscisses, on obtient une droite croissante du temps et de pente γ. Cela impliquera que la distribution de Gompertz est appropriée pour les données à disposition.
2. Lorsque $\gamma < 0$ (négatif), la fonction de hasard est une fonction décroissante du temps (droite décroissante de pente γ). Cela implique que lorsque l'on représente dans un graphique en mettant le logarithme de la fonction de risque sur l'axe des ordonnées et le temps sur l'axe des abscisses, on obtient une droite décroissante du temps et de pente γ. Cela implique que la distribution de Gompertz est appropriée pour analyser les données à disposition.
3. Lorsque $\gamma = 0$, la fonction de hasard est constante (indépendante du temps) et la distribution de Gompertz n'est autre que la distribution exponentielle de paramètre $exp(\beta_0)$ (Cleves et al., 2010).

7.1 Approche graphique

L'approche graphique nous permet de tester si le modèle est adapté aux données. Pour ce faire, on va représenter soit le logarithme de la fonction de risque soit la fonction de séjour en fonction du temps et observer l'allure de la courbe qu'on obtient.

Cas où $\gamma > 0$

Dans le cas où $\gamma > 0$, lorsqu'on représente un graphique dans lequel on met sur l'axe des ordonnées le logarithme du risque et sur l'axe des abscisses le temps, on obtient une droite croissante du temps (pente positive). Le premier graphique de la figure 7.1 illustre cette situation.

Cas où $\gamma < 0$

Dans le cas où $\gamma < 0$, lorsqu'on représente un graphique dans lequel on met sur l'axe des ordonnées le logarithme du risque et sur l'axe des abscisses le temps, on obtient une droite décroissante du temps (pente négative) comme le montre le deuxième graphique de la figure 7.1.

Cas où $\gamma = 0$

Dans le cas où $\gamma = 0$, lorsqu'on représente un graphique dans lequel on met sur l'axe des ordonnées le logarithme du risque et sur l'axe des abscisses le temps, on obtient une droite parallèle à l'axe du temps. Dans ce cas, la fonction de risque est constante et indépendante du temps, la distribution de Gompertz n'est alors autre que la distribution exponentielle.

7.1.1 Format des données

Le format des données pour faire une régression de Gompertz est identique au format de données pour faire une régression exponentielle ou une régression de Weibull. Nous avons besoin des dates de début et de fin d'observation, de la variable censure et des caractéristiques individuelles.

7.1.2 Exemple d'application

Dans cet exemple d'application, on va reprendre la même base de données pour répondre à la même question de recherche que celle du chapitre dédié à la régression exponentielle. Dans la régression de Gompertz, la « variable dépendante » c'est le risque, mais celui-ci évolue de manière différente que dans le modèle de régression exponentiel et dans le modèle de Weibull. Dans la régression de Gompertz, le risque de base évolue de manière linéaire (croissant ou décroissant) ou constante selon la pente.

Question de recherche : Le « risque » d'obtenir le diplôme de bachelor (dans notre cas c'est en fait la chance d'obtenir le diplôme de bachelor) des étudiants internationaux dépend-elle du sexe et de l'âge de ces étudiants ?

Pour les analyses exploratoires, se référer à la régression exponentielle (section 5.3.1). Le Tableau 7.1 représente le tableau des coefficients de la régression de Gompertz. On remarque que le coefficient $\widehat{\gamma}$ est positif ce qui implique que le logarithme du risque est une fonction croissante du temps. Le coefficient de la variable sexe (codée 1=femme et 0=homme) est négatif, cela veut dire qu'être une femme par rapport à être un homme pour le même âge ferait dimuner le risque d'obtenir le diplôme de bachelor de 2.311 fois. Cela veut dire que le risque d'obtenir le diplôme de bachelor est plus faible pour les femmes que pour les hommes. Le coefficient de la variable âge est positif, ce qui se traduit par un effet positif de l'âge sur le risque d'obtenir le diplôme de bachelor. En effet, lorsque l'âge augmente d'une unité, toutes choses étant égales par ailleurs, le risque d'obtenir le diplôme de bachelor augmente de 11.23%.

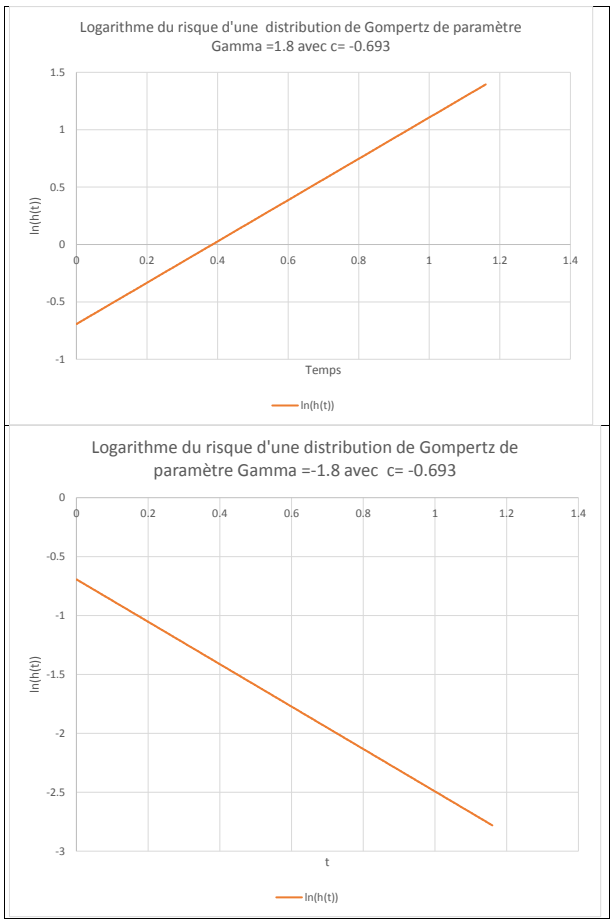


FIGURE 7.1 – Fonction de risque d'une distribution de Gompertz de paramètres respectifs $\gamma = 1.8$ et $\gamma = -1.8$

Tableau 7.1 – Tableau des coefficients d'une régression de Gompertz

Variable	Coeff.	Erreur standard	Z	P-valeur	IC à 95%
sexe	-2.311	0.4436	-5.21	0.001	-3.181 - 1.442
âge	0.1055	0.0371	2.84	0.005	0.0326 - 0.1784
constante	-10.1927	2.7217	-3.74	0.001	-15.5272 - 4.8582
gamma	0.0752	0.0233	3.22	0.001	0.0295 - 0.1211

La fonction de risque estimée pour tout instant t s'écrit de la manière suivante :

$$ln\left[h(\widehat{t;Z_j})\right] = -10.1927 + 0.1055age - 2.311sexe + 0.0752t$$

Cette fonction de risque nous permet de prédire le risque de voir la durée des études s'allonger selon le sexe et l'âge des étudiants pour une durée donnée.

On peut aussi tester la significativité des coefficients du modèle en posant les hypothèses nulle et alternative suivantes :

$H_0 : \beta_j = 0 \ \ vs \ \ H_1 : \beta_j \neq 0, j = 1, 2, \cdots k.$

Le tableau des coefficients montre que les p-valeurs associées au test de Wald sont toutes inférieures au seuil α de 5%, ce qui nous permet de rejeter H_0 et de conclure que l'hypothèse de recherche est corroborée. Mais le test de Wald étant moins performant que le test du rapport de vraisemblance, on va refaire le test statistique par la méthode du rapport de vraisemblance. Par ce test, on rejettera H_0 au seuil de 5% si la différence de déviance est supérieure à 3.84 soit au χ^2_1. Le Tableau 7.2 résume le calcul des différences de déviances.

Tableau 7.2 – Tableau résumant le test du rapport de vraisemblance pour la régression de Gompertz

Modèles	Déviances	Différence de déviances	Comparaison	Conclusion
Complet	42.609			
Sans l'âge	46.584	46.584 - 42.609 = 3.975	3.975 > 3.84	On rejette H_0 au seuil de 5%
Sans le sexe	57.497	57.497 - 42.6069 = 14.88	14.88 >3.84	On rejette H_0 au seuil de 5%

Toutes les différences de déviances se sont avérées supérieures à 3.84, ce qui nous permet de rejetter H_0 au seuil de 5% et de conclure que notre hypothèse de recherche est corroborée.

Lorsque nous voulons comparer le modèle exponentiel et le modèle de Gompertz, la procédure de choix se fera exactement comme celle décrite entre le modèle de Weibull et le modèle exponentiel. Le modèle exponentiel étant un cas particulier du modèle de Gompertz, on commencera par ajuster un modèle de Gompertz puis on fera des tests statistiques sur le paramètre de forme γ de la distribution de Gompertz pour faire le choix entre les deux modèles. On testera $H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$. Si on rejette H_0, cela impliquera que le modèle exponentiel ne convient pas pour analyser les données à disposition. Nous remarquons cependant que les modèles de Weibull et de Gompertz ne sont pas comparables selon la même procédure que pour les modèles de Weibull vs exponentiel ou Gompertz vs exponentiel dans le sens où les deux modèles n'ont pas de liens directs. En effet, le modèle de Weibull n'est pas un cas particulier du modèle de Gompertz et vice versa. Pour choisir un modèle entre les deux, on passera par le critère AIC ; ce dernier sera présenté ultérieurement.

7.2 Diagramme récapitulatif des modèles à hasard proportionnel

Le diagramme ci-dessous donne un récapitulatif des modèles de régression paramétriques à hasard proportionnel selon l'allure de la fonction de risque de base $h_0(t)$.

Le diagramme récapitulatif de la figure 7.2 donne un aspect visuel des liens entre les différents modèles à hasard proportionnel. Le tableau récapitulatif 7.3 ci-dessous complète cet aspect visuel par un aspect plus technique à travers quelques fonctions qui caractérisent ces différents modèles à hasard proportionnel.

Dans les modèles que nous venons de présenter, les caractéristiques individuelles agissent sur la fonction de risque. Cette dernière pouvant être constante, monotone ou linéaire en fonction du temps. Ce groupe de modèles porte le nom de modèles à hasard proportionnel.

Il existe d'autres modèles pour lesquels, les caratéristiques individuelles agissent sur la durée de séjour. Ces modèles portent le nom de modèles à temps de sorties accélérées. Les modèles que nous présenterons dans les sections suivantes appartiennent à ce groupe de modèles. Ces modèles sont utiles pour analyser des phénomènes dont l'allure de la fonction de risque peut être unimodale, c'est-à-dire que l'allure de la fonction de risque peut croître, atteindre un maximum, puis décroître. Par exemple, le risque de mort subite est très élevé chez le nouveau-né avant un an mais ce risque diminue avec l'âge. Les modèles, log-normale, log-logistique et gamma appartiennent à ce groupe de modèles. Nous commencerons d'abord par présenter le modèle de régression log-normale.

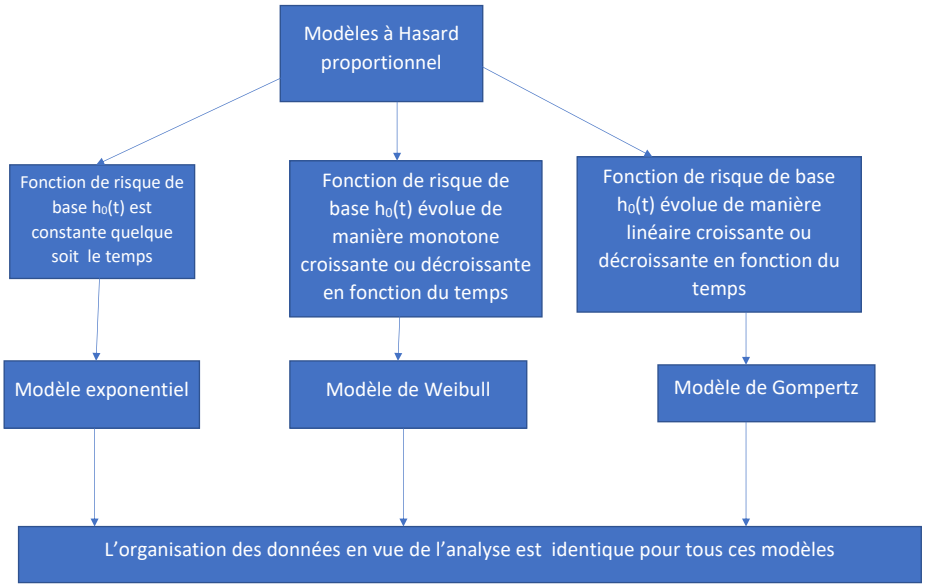


FIGURE 7.2 – Diagramme récapitulatif des modèles de régression paramétriques à hasard proportionnel

Tableau 7.3 – Tableau récapitulatif des modèles à hasard proportionnel

Modèles	Caractéristiques
Exponentiel	$h_0(t) = \beta_0$ $h(t, Z_j)$ est constant et est indépendant de t. $H(t, Z_j)$ est une droite qui passe par $(0, 0)$.
Weibull	$ln[h_0(t)] = c + (p-1)ln(t)$ $h(t, Z_j)$ est monotone croissant si $p > 1$. $h(t, Z_j)$ est monotone décroissant si $p < 1$. $h(t, Z_j)$ est constant si $p = 1$ (modèle exponentiel) $ln[-lnS(t)]$ est une fonction linéaire de $ln(t)$ Weibull vs Exponentiel : le choix se fera par simple test d'hypothèses sur le paramètre de forme p de la loi de Weibull.
Gompertz	$ln\left[h_0(t)\right] = \gamma t + \beta_0$ $h(t, Z_j)$ est une droite de pente γ $h(t, Z_j)$ est une droite croissante du temps si $\gamma > 0$ $h(t, Z_j)$ est une droite décroissante du temps si $\gamma < 0$ $h(t, Z_j)$ est une droite constante du temps si $\gamma = 0$ (modèle exponentiel) $ln[lnS(t)]$ est une fonction linéaire de t Gompertz vs Exponentiel : le choix se fera par simple test d'hypothèses sur le paramètre de forme γ de la loi de Gompertz. Gompertz vs Weibull : le choix se fera avec le critère AIC car ces deux modèles ne sont pas directement liés.

Chapitre 8

Le modèle de régression Log-normale

La régression log-normale est un modèle à temps de sorties accélérées. La formulation pour les modèles à temps de sorties accélérées comme pour les autres modèles du même type est la suivante :

$$\tau_j = exp(-Z_j\beta_z)t_j, \tag{8.1}$$

τ_j étant une variable aléatoire.

Pour la distribution log-normale, on fait l'hypothèse que :

$$\tau_j \sim lognormale(\beta_0, \sigma)$$

où β_0 et σ sont deux paramètres à estimer.

On lit cela comme : τ_j suit une loi log-normale de paramètres (β_0, σ).

En prenant le logarithme de l'équation (8.1) et après quelques transformations algébriques, on obtient le modèle log-normal pour un individu j ayant les caractéristiques individuelles Z_j. Ce modèle s'écrit :

$$ln(t_j) = \beta_0 + Z_j\beta_z + u_j$$

u_j suivant une loi normale standard de moyenne 0 et d'écart-type σ.

Dans le modèle de régression log-normale, en transformant le temps par son logarithme $(ln(t_j))$, le modèle devient un modèle de régression linéaire. En l'absence de censures, les paramètres du modèle peuvent ainsi être estimés par la méthode des moindres carrés ordinaires (Lollivier, 2000).

Un exemple d'application du modèle log-normal est présenté dans la section qui va suivre.

8.1 Exemple d'application

Nous allons considérer la même base de données que pour les modèles précédents ; l'objectif étant de répondre à une question de recherche qui sera formulée à partir des mêmes données en utilisant la régression log-normale. Dans la régression lognormale, il est important de rappeler que la variable dépendante est la durée de séjour.

La question de recherche est la suivante :

Question de recherche : La durée des études de bachelor des étudiants internationaux dépend-elle du sexe et de l'âge de ces étudiants ?

Le Tableau 8.1 qui représente le tableau des coefficients montre que tous les paramètres sont significativement différents de zéro selon le test de Wald car les p-valeurs qui leurs sont associées sont toutes inférieures à 0.05.

Tableau 8.1 – Tableau des coefficients de la régression log-normale

Variable	Coeff.	Erreur standard	z	P-valeur	IC à 95%
sexe	1.4595	0.2469	5.91	0.001	-0.9755 1.9436
âge	-0.07856	-0.02222	-3.54	0.001	-0.1221 -0.03501
constante	7.4580	1.5903	4.69	0.001	4.3411 10.5750
$/ln\_sig$	-0.2962	0.1268	-2.33	0.020	-0.5438 -0.0467
sigma	0.7443	0.09439			0.5805 0.9544

Le tableau des coefficients montre que lorsque l'âge augmente d'une unité, toutes choses étant égales par ailleurs, la durée moyenne des études est multipliée par $e^{-0.07856} = 0.9444$, ce qui correspond à une diminution de la durée des études de 7.55%. Par ailleurs, être une femme par rapport à être un homme, toutes choses étant égales par ailleurs, la durée moyenne des études est multipliée par $e^{1.4595} = 4.304$.

En faisant le test du rapport de vraisemblance pour tester la significativité des paramètres estimés, on observe que les variables sexe et âge sont significatives car les différences de déviances qui leurs sont associées sont toutes supérieures au $\chi^2_1 = 3.84$. Ce qui nous permet de rejetter H_0 au seuil de 5% pour les paramètres des deux variables et de conclure que l'hypothèse de recherche est corroborée. Le Tableau 8.2 présente le test du rapport vraisemblance effectué.

Tableau 8.2 – Tableau résumant le test du rapport de vraisemblance pour la régression log-normale

Modèles	Déviances	Différence de déviances	Comparaison à $\chi^2_1 = 3.84$	Conclusion
Complet	41.8453			
Sans l'âge	47.5113	47.8656 - 41.8453 = 6.0203	6.0203 > 3.84	On rejette H_0
Sans le sexe	56.4024	56.4024 - 41.8453 = 14.5571	14.5571 >3.84	On rejette H_0

A partir du tableau des coefficients, on peut également écrire le modèle pour tout t en fonction des caractéristiques individuelles.

Ce modèle s'écrit :

$$\widehat{ln(t_j)} = 7.4580 + 1.4595sexe - 0.07856age$$

Cette équation nous permet de prédire la durée des études de bachelor en fonction du sexe et de l'âge de l'étudiant à chaque instant t. On remarque un effet négatif de l'âge sur la durée moyenne des études. On remarque également un effet fort et significatif du genre sur la durée moyenne des études ; la durée moyenne des études est plus longue pour les femmes que pour les hommes pour le même âge.

Dans le modèle log-normale, nous avons fait l'hypothèse que τ_j suit une distribution log-normale de paramètres β_0 et σ, ainsi $ln(\tau_j)$ suit une ditribution normale standard de moyenne 0 et d'écart-type σ. En prenant le logarithme du temps, le modèle de régression log-normale devient un modèle de régression linéaire. Il est aussi possible de faire l'hypothèse que u_j suit une distribution log-logistique. On obtient ainsi le modèle de régression log-logistique qui est présenté dans la section suivante.

Chapitre 9

Le modèle de régression Log-logistique

Le modèle de régression log-logistique est un modèle à temps de sorties accélérées (AFT). La formulation générale des modèles à AFT est la suivante (Allison, 2010) :

$$\tau_j = exp(-Z_j\beta_z)t_j \tag{9.1}$$

Pour le modèle log-logistique, on fait l'hypothèse que :

$$\tau_j \sim log - logistique(\beta_0, \gamma)$$

où β_0 et γ sont des paramètres à estimer et τ_j une variable aléatoire.

On lit cette expression comme τ_j suit une loi log-logistique de paramètres (β_0, γ).

En prenant le logarithme de l'équation (9.1), et après quelques calculs algébriques, on peut l'écrire comme :

$$ln(t_j) = Z_j\beta_z + ln(\tau_j) = \beta_0 + Z_j\beta_z + u_j$$

Où u_j suit une distribution logistique.

Dans la distribution log-logistique, les fonctions de risque et de séjour s'écrivent de la manière suivante :

$$h(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1 + (\lambda t)^{\gamma}} \tag{9.2}$$

Et

$$S(t) = \frac{1}{1 + (\lambda t)^{\gamma}} \tag{9.3}$$

Avec $\gamma = \dfrac{1}{\sigma}$ et $\lambda = e^{-\beta Z} = e^{-[\beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k]}$.

En procédant à une transformation mathématique de la fonction de séjour, on peut obtenir une équation qui met en lien le logarithme du rapport de la fonction de survie à son complémentaire à 1 en fonction des variables explicatives, du paramètre γ et du logarithme du temps.

$$log\left[\frac{S(t)}{1 - S(t)}\right] = \gamma\left[\beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k\right] - \gamma log(t)$$

Comme $\gamma = \dfrac{1}{\sigma}$, cette équation peut s'écrire comme :

$$log\left[\frac{S(t)}{1 - S(t)}\right] = \frac{1}{\sigma}\beta_0 + \frac{1}{\sigma}\beta_1 z_1 + \cdots + \frac{1}{\sigma}\beta_k z_k - \gamma log(t)$$

En posant $\beta^* = \dfrac{\beta_i}{\sigma}$, on peut écrire l'équation ci-dessus comme suit :

$$log\left[\frac{S(t)}{1-S(t)}\right] = \beta_0^* + \beta_1^* z_1 + \cdots + \beta_k^* z_k - \gamma log(t) \tag{9.4}$$

Cette équation présente un double avantage. Le premier avantage réside dans le fait que la partie $\gamma log(t)$ peut s'ajouter à la constante β_0^* du modèle de manière à écrire la partie de droite de cette équation comme une combinaison linéaire des variables explicatives (Allison, 2010). Dans ce cas, en posant $\beta_0^{**} = \beta_0^* - \gamma log(t)$, l'équation s'écrit comme :

$$log\left[\frac{S(t)}{1-S(t)}\right] = \beta_0^{**} + \beta_1^* z_1 + \cdots + \beta_k^* z_k \tag{9.5}$$

Le deuxième avantage réside dans l'utilisation d'une approche graphique en observant l'allure de la courbe qu'on obtient lorsqu'on représente un graphique dans lequel on met
$log\left[S(t)/(1-S(t)\right]$ sur l'axe des ordonnées et $log(t)$ sur l'axe des abscisses. Si le modèle log-logistique est vérifié on obtient une droite.

Un des avantages que présente le modèle log-logistique par rapport au modèle lognormal réside dans le fait que les expressions mathématiques de la fonction de hasard et de la fonction de survie du modèle log-logistique sont simples. Dans la régression log-logistique, on peut facilement déduire l'allure de la fonction de hasard à travers le signe du paramètre γ selon les configurations suivantes :

Si $\gamma < 1$, la fonction de hasard de la distribution log-logistique part de zéro, croît, atteint un maximum puis décroît. Dans ce cas, la courbe de la fonction de risque de la distribution log-logistique est similaire à celle de la distribution log-normale (Courgeau et Lelièvre, 1989; Allison, 2010).

Si $\gamma \geq 1$, la fonction de hasard de la distribution log-logistique est une fonction monotone et décroissante. Elle commence à l'infini et décroît en tendant vers zéro. Dans ce cas, la distribution log-logistique est très similaire à celle de la distribution de Weibull.

Si $\gamma = 1$, la fonction de risque de la distribution log-logistique vaut λ à l'instant $t = 0$ et tend vers zéro au fur et à mesure que t augmente.

9.1 Exemple d'application de la distribution log-logistique

Dans cet exemple, on s'intéresse à l'impact des variables âge et sexe sur la fonction de survie ou fonction de séjour en utilisant les mêmes données que dans les exemples précédents. Dans notre cas, la fonction de survie représente, pour rappel, la probabilité que l'étudiant n'obtienne pas son diplôme après une certaine durée t. En d'autres termes, on aurait pu formuler cette question de recherche comme suit : « la probabilité que les étudiants internationaux n'obtiennent pas leurs diplômes après une certaine durée t dépend-elle de l'âge et du sexe des étudiants ? »

En réalisant le modèle log-logistique, on obtient le tableau des coefficients du Tableau 9.1 :

Tableau 9.1 – Tableau des coefficients de la régression log-logistique

Variable	Coeff.	Erreur standard	z	P-valeur	IC à 95%
sexe	1.4344	0.2483	5.78	0.001	0.9477 1.9211
âge	-0.07558	0.0221	-3.42	0.001	-0.1189 -0.0322
constante	7.2844	1.5620	4.66	0.000	4.2229 10.3460
$/ln\_gam$	-0.8565	0.1476	-5.8	0.001	-1.1459 -0.5671
gamma	0.4246	0.6269			0.3179 0.5671

Le tableau des coefficients montre que les paramètres associés aux variables sexe et âge selon le test de Wald ont un impact sur la fonction de survie (sur la probabilité de ne pas obtenir son diplôme après une certaine durée t) car les p-valeurs qui leur sont associées sont inférieures à 0.05. On peut donc rejeter l'hypothèse de nullité des coefficients du modèle.

Le modèle estimé peut s'écrire de la manière suivante :

$$log\left[\frac{\widehat{S(t)}}{1-\widehat{S(t)}}\right] = 7.2844 + 1.4344sexe - 0.075age$$

A partir de ce modèle, on peut mesurer l'impact des variables âge et sexe sur la fonction de survie. Le coefficient de la variable sexe est positif et celui de la variable âge est négatif, ce qui nous permet de dire que l'âge a un impact négatif sur la fonction de survie alors que la variable sexe a un impact positif sur cette dernière. Le test de rapport de vraisemblance effectué pour ce modèle est présenté dans le Tableau 9.2.

Tableau 9.2 – Tableau résumant le test du rapport de vraisemblance pour la régression log-logistique

Modèles	Déviances	Différence de déviances	Comparaison à $\chi^2_1 = 3.84$	Conclusion
Complet	42.2405			
Sans l'âge	47.8956	47.8656 - 42.2405 = 5.6251	5.6251 > 3.84	On rejette H_0
Sans le sexe	56.8101	56.8101 - 42.2405 = 14.5696	14.5696 >3.84	On rejette H_0

Ce test montre aussi que les deux variables sont significatives, car les différences de déviances pour les deux variables sont supérieures à 3.84.

Par ailleurs, le tableau des coefficients montre aussi que $\gamma = 0.4246 < 1$, ce qui nous permet de dire que la fonction de risque de cette distribution croît, atteint un maximum (un sommet), puis décroît.

A l'image des modèles log-normale et log-logistique, on peut aussi faire l'hypothèse que τ_j suit une distribution gamma dont les paramètres sont β_0, k, et σ ; on obtient ainsi le modèle de régression gamma. Ce modèle offre l'avantage d'être un cas particulier des modèles exponentiel, de Weibull et log-normale. Le modèle gamma sera présenté dans la section qui suit.

Chapitre 10

Le modèle Gamma généralisé

Dans la régression gamma généralisée qui est aussi un modèle en AFT, nous partons aussi de l'équation $\tau_j = exp(-Z_j\beta_z)t_j$ en supposant que :

$$\tau_j \sim Gamma(\beta_0, k, \sigma)$$

On lit cela comme τ_j suit une distribution gamma de paramètres (β_0, k, σ) où ces paramètres sont à estimer. Il s'agit donc d'une distribution qui a trois paramètres à estimer.

On obtient, à partir de l'équation générale des modèles AFT, que :

$$ln(t_j) = \beta_0 + Z_j\beta_z + u_j$$

u_j étant la partie aléatoire du modèle.

En raison de la très grande complexité mathématique des fonctions à la base de ce modèle (voir encadré), ces dernières ne seront pas présentées. Nous allons cependant réaliser un exemple d'application du modèle gamma dans le but de montrer comment on retrouve certaines distributions paramétriques à partir de cette distribution.

La fonction de répartition de la distribution gamma s'écrit :

$$F(\tau) = \begin{cases} I(\gamma, u), & \text{si } k > 0 \\ \phi(z), & \text{si } k = 0 \\ 1 - I(\gamma, u), & \text{si } k < 0 \end{cases}$$

Avec $\gamma = |k|^{-2}$, $z_0 = sign(k)\dfrac{\{ln(\tau) - \beta_0\}}{\sigma}$, $u = \gamma exp(\sqrt{\gamma}z_0)$,
$\phi()$ représente la fonction de répartition de la loi normale centrée réduite,
$I(a, x)$ représente la fonction de densité de la distribution gamma
La fonction de survie de la distribution gamma peut alors s'écrire :

$$S(t_j|Z_j) = 1 - F^*(t_j)$$

$F^*()$ représente la fonction de répartition de la distribution gamma dans laquelle z_0 est remplacé par $z = sign(k)\dfrac{ln(\tau) - (\beta_0 + Z_j\beta_z)}{\sigma}$

La distribution gamma généralisée est un cas particulier des distributions de Weibull, exponentielle et lognormale. Ainsi, la distribution gamma suit une distribution de Weibull si $k = 1$ et $p = 1/\sigma$; elle suit une distribution exponentielle si $k = \sigma = 1$ et une distribution lognormale si $k = 0$.

La distribution gamma généralisée est très souvent utilisée dans le but d'évaluer et de sélectionner le modèle paramétrique approprié pour l'analyse des données à disposition (Cleves et al., 2010).

En effet, lorsque les paramètres de la distribution gamma généralisée satisfont certaines conditions, il est possible de retrouver trois autres distributions : la ditribution lognormale, la distribution de Weibull et la distribution exponentielle. Pour choisir quelle distribution est adéquate pour analyser les données à disposition, on va d'abord ajuster un modèle de régression gamma généralisée, ensuite on va tester les hypothèses suivantes :

1. Le modèle suit une distribution lognormale ; on testera :

$$H_0 : k = 0$$

$$H_1 : k \neq 0$$

   Si H_0 est vrai, le modèle est lognormal ; autrement il ne l'est pas.

2. Le modèle suit une distribution de Weibull ; on testera :

$$H_0 : k = 1$$

$$H_1 : k \neq 1$$

   Si H_0 est vrai, le modèle est de type Weibull, autrement il ne l'est pas.

3. Le modèle suit une distribution exponentielle, on testera :

$$H_0 : k = 1, \sigma = 1$$

$$H_1 : k \neq 1, \sigma \neq 1$$

   Si H_0 est vrai, le modèle est exponentiel, autrement il ne l'est pas.

10.1 Exemple d'application

Considérons la même question de recherche que dans les modèles précédents. Notre objectif est de déterminer le modèle paramétrique approprié pour analyser les données et répondre à notre question de recherche. Pour ce faire, on a réalisé un modèle gamma généralisé et on a obtenu le Tableau 10.1.

Tableau 10.1 – Tableau des coefficients de la régression gamma généralisée

Variable	Coeff.	Erreur standard	z	P-valeur	IC à 95%
sexe	1.4066	0.2551	5.51	0.001	0.9066 1.9067
âge	-0.0728	0.0231	-3.15	0.002	-0.1180 -0.0275
$constante$	7.2233	1.5977	4.52	0.000	4.0918 10.3548
$/ln\_sig$	-0.3836	0.1768	-2.17	0.030	-0.7302 -0.0370
/kappa	0.4644	0.5288	0.88	0.380	-5719 1.5008
sigma	0.6814	0.1205			0.4818 0.9636

Le tableau des coefficients de la régression gamma généralisée montre que les p-valeurs associées au test de Wald sont toutes inférieures à 0.05, ce qui nous permet de rejeter H_0 au seuil de 5% pour les paramètres associés aux variables sexe et âge et de conclure que l'hypothèse de recherche est corroborée.

Nous avons également effectué un test du rapport de vraisemblance qui est présenté dans le Tableau 10.2. Ce tableau montre que toutes les différences de déviance sont supérieures à 3.84, ce qui nous permet de rejeter H_0 au seuil de 5% et de conclure que les deux paramètres sont significativement différents de 0, ce qui corrobore l'hypothèse de recherche.

Tableau 10.2 – Tableau résumant le test du rapport de vraisemblance pour la régression gamma généralisée

Modèles	Déviances	Différence de déviances	Comparaison à $\chi^2_1 = 3.84$	Conclusion
Complet	41.4834			
Sans l'âge	46.2432	46.2432 - 41.4834 = 4.7598	4.7598 > 3.84	On rejette H_0
Sans le sexe	56.3461	56.3461 - 41.4834 = 14.8627	14.8627 >3.84	On rejette H_0

10.2 Choix du modèle paramétrique adapté à partir de la régression gamma généralisée

A partir des résultats de la régression gamma généralisée, on va à présent faire un test d'hypothèse pour choisir le modèle paramétrique adapté pour analyser les données.

1. Le modèle suit une distribution lognormale ; on testera :

$$H_0 : k = 0$$

$$H_1 : k \neq 0$$

   La statistique de test du $\chi^2 = 0.77 < \chi^2_1 = 3.84$; ce qui ne nous permet pas de rejeter H_0 au seuil de 5%. On peut donc conclure que le modèle lognormal est adapté pour analyser les données et pour répondre à la question de recherche. La p-valeur associée à ce test vaut $0.3798 > 0.05$ (on ne rejette pas H_0 au seuil de 5%), ce qui nous permet d'obtenir la même conclusion.

2. Le modèle suit une distribution de Weibull ; on testera :

$$H_0 : k = 1$$

$$H_1 : k \neq 1$$

   La statistique du test du χ^2 associée à ce test vaut 1.03 < 3.84, on ne rejette pas H_0 au seuil de 5%. La modèle de Weibull est aussi adapté pour analyser ces données.

3. Le modèle suit une distribution exponentielle, on testera :

$$H_0 : k = 1, \sigma = 1$$

$$H_1 : k \neq 1, \sigma \neq 1$$

   La statistique de test associée au test de $\chi^2 = 15.86 > \chi^2_1 = 3.84$, ce qui permet de rejeter H_0 au seuil de 5%. En rejetant H_0, on sous-entend par là que le modèle exponentiel n'est pas adapté pour répondre à notre question de recherche.

   A partir du modèle gamma généralisé, les résultats montrent que pour répondre à notre question de recherche, on aurait pu choisir soit le modèle de Weibull soit le modèle log-normal.

10.3 Diagramme récapitulatif des modèles à temps de sorties accélérées

Le diagramme de la figure 10.1 ci-dessous donne un récapitulatif des modèles de régression paramétriques à temps de sorties accélérées selon les hypothèses faites sur la variable aléatoire u_j.

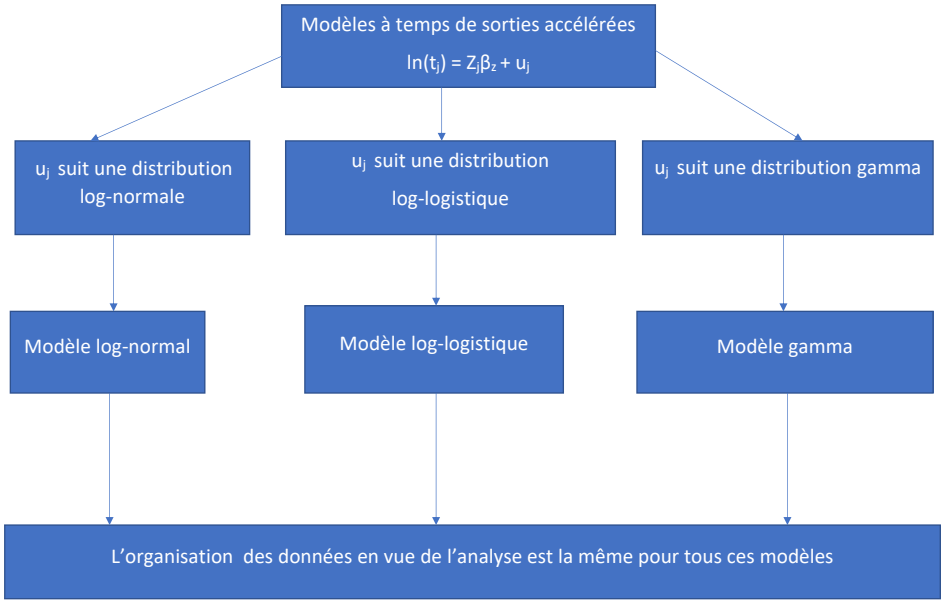


FIGURE 10.1 – Diagramme récapitulatif des modèles de régression paramétriques à temps de sorties accélérées

10.3.1 Autres méthodes de sélection de modèles

A l'image du coefficient de détermination R^2 dans la régression linéaire, au fur et à mesure que l'on introduit des variables explicatives dans un modèle, la vraisemblance associée à ce modèle augmente progressivement même si les variables introduites ne sont pas importantes pour le modèle. Pour freiner l'augmentation non contrôlée de la vraisemblance suite à l'introduction de nouvelles variables explicatives dans le modèle, le statisticien japonais Hirotugu Akaike a introduit le critère AIC (Akaike Information Criterion) (Collett, 2015; Klein et Moeschberger, 2003) qui porte son nom et qui tient compte du nombre de paramètres introduits dans le modèle. Le but de l'introduction du nombre de paramètres étant de minimiser le critère AIC (Dodge et Rousson, 2004).

Ce critère est défini comme suit dans les modèles paramétriques d'analyse des parcours de vie.

$$AIC = -2ln(L) + 2(k + c)$$

k étant le nombre de variables explicatives dans le modèle et c, le nombre de paramètres de base dans chaque modèle ; par exemple $c = 3$ pour le modèle gamma généralisé car on a trois paramètres à estimer dans ce modèle (β_0, k, σ). Ainsi, lorsque nous avons plusieurs modèles à comparer, le meilleur modèle sera celui qui aura le critère AIC le plus petit. Pour faire un lien avec tous les modèles paramétriques qui ont été exposés dans ce chapitre et pour choisir un modèle, on va calculer l'AIC pour tous les modèles. Le meilleur modèle sera celui qui aura la plus petite valeur de l'AIC (Montaseri et al., 2016).

Le Tableau 10.3 résume les valeurs de l'AIC pour l'exemple développé dans les différents chapitres. Le paramètre $k = 2$ pour tous les modèles parce que dans l'exemple qui a été développé, on avait deux variables explicatives qui sont l'âge et le sexe. Le paramètre c lui varie selon le nombre de paramètres à estimer dans chaque modèle.

Tableau 10.3 – Tableau résumant les valeurs de l'AIC pour l'exemple développé dans tous les modèles paramétriques

Modèles	-ln(L)	k	c	AIC
Exponentiel	47.534	2	1	101.069
Weibull	41.992	2	2	91.985
Gompertz	42.609	2	2	93.218
Lognormal	41.8453	2	2	91.690
Loglogistic	42.2405	2	2	92.481
Generalized gamma	41.4834	2	3	92.967

Pour calculer l'AIC pour le modèle exponentiel par exemple, on procède de la manière suivante :$AIC = 2 * 47.534 + 2(2 + 1) = 101.069$. Le calcul se fait de la même manière pour les autres modèles.

Par le critère de l'AIC, le modèle de régression lognormale est le meilleur modèle parce que la valeur de l'AIC est plus faible pour ce modèle que pour les autres modèles. Il existe aussi le critère BIC (Bayesian Information Criterion) qui s'interprète de la même manière que l'AIC.

Le tableau 10.4 ci-dessous donne un récapitulatif des différents modèles à temps de sorties accélérées.

A ce stade, nous avons présenté deux groupes de modèles qui sont : les modèles non paramétriques et les modèles paramétriques. Il existe une autre approche dite semi-paramétrique. Le modèle semi-paramétrique le plus connu est le modèle de Cox dont la particularité réside dans le fait qu'aucune hypothèse n'est faite en ce qui concerne la distribution du risque au cours du temps. Ce modèle est présenté dans la section suivante.

Tableau 10.4 – Tableau récapitulatif des modèles à temps de sorties accélérées

Modèles	Caractéristiques
Log-normale	u_j suit une loi normale de moyenne 0 et d'écart-type σ. $h(t, Z_j)$ est une fonction croissante puis décroissante vers 0. En l'absence de censures, ce modèle peut être estimé avec une régression linéaire. Pour choisir un modèle parmi les modèles : exponentiel, Weibull, Gompertz et log-normale, on se basera sur le critère AIC.
Log-logistique	u_j suit une loi logistique. $ln[\frac{S(t)}{1-S(t)}]$ est une droite qui est en fonction de $ln(t)$. Si $\gamma < 1$, $h(t, Z_j)$ croît, atteint un maximum puis décroît Si $\gamma \geq 1$, $h(t, Z_j)$ est une fonction monotone décroissante Si $\gamma = 1$, $h(t, Z_j)$ vaut λ pour $t = 0$ et tend vers 0 au fur et à mesure que t augmente. Pour choisir un modèle parmi les modèles : exponentiel, Weibull, Gompertz, log-normale et log-logistique, on se basera aussi sur le critère AIC.
Gamma généralisé	Trois paramètres à estimer :β_0, k et σ. Si $k = 0$, on retrouve la distribution log-normale. Si $k = 1$, on retrouve le modèle de Weibull. Si $k = 1$ et $\sigma = 1$, on retrouve la distribution exponentielle. Si $\gamma = 1$, $h(t, Z_j)$ vaut λ pour $t = 0$ et tend vers 0 au fur et à mesure que t augmente. Pour choisir un modèle parmi : les modèles exponentiel, Weibull et log-normale, on fera un test d'hypothèses sur les paramètres estimés de la distribution gamma. Pour choisir un modèles parmi tous les modèles paramétriques présentés on se basera sur le critère AIC.

Chapitre 11

Le modèle de régression semi-paramétrique de Cox

Dans les chapitres précédents, nous avons présenté les modèles non paramétriques et les modèles paramétriques. Dans ce chapitre, on va présenter un modèle qui combine les approches non paramétrique et paramétrique, d'où l'appellation de semi-paramétrique (Evans, 2017).

L'avantage principal du modèle semi-paramétrique réside dans le fait qu'on ne fait aucune hypothèse en ce qui concerne la distribution du risque avec le temps, par opposition aux modèles paramétriques, tels que les modèles de Gompertz ou de Weibull par exemple. Le plus célèbre des modèles semi-paramètriques est le modèle de Cox (Cox, 1972) en temps continu aussi appelé modèle semi-paramétrique à risque proportionnel (Ritschard, 2004). Le modèle de Cox stipule que la fonction de hasard est de la forme :

$$h(t, Z_j) = h_0(t) exp(Z_j \beta_z) \tag{11.1}$$

Ce modèle peut aussi s'écrire comme une fonction linéaire des caractéristiques individuelles en prenant le logarithme du risque.

$$ln\left[h(t, Z_j)\right] = ln(h_0(t)) + \beta_1 Z_1 + \beta_2 Z_2 + \cdots \beta_k Z_k \tag{11.2}$$

Avec Z_j, la matrice des variables explicatives et β_z, le vecteur des paramètres du modèle qui est à estimer à partir des données.

Dans le modèle semi-paramétrique de Cox, la « variable dépendante » est le risque ou le logarithme du risque et les variables indépendantes peuvent être n'importe quelles caractéristiques individuelles (âge, sexe, nationalité,...). Mais à l'image des modèles paramétriques, le risque n'est pas une caractéristique mesurée. Ce modèle permet aussi de répondre à des questions de recherche qui s'intéressent à l'impact d'une ou de plusieurs variables explicatives sur le risque d'occurrence d'un événement d'intérêt.

Les paramètres estimés $\widehat{\beta_z}$ mesurent l'impact des variables explicatives correspondantes sur le risque (comme dans l'équation 11.1) ou sur le logarithme du risque (comme dans l'équation 11.2) selon la manière dont le modèle est spécifié. En calculant l'exponentielle d'un coefficient (e^{β}), on obtient le rapport des cotes (odds ratios). L'appellation de semi-paramétrique réside dans le fait que la fonction de risque se décompose en deux fonctions :

- La première fonction est $h_0(t)$; cette fonction dépend du temps mais est indépendante des caractéristiques individuelles. C'est la partie non paramétrique du modèle parce que sa forme n'est pas spécifiée.
- La deuxième fonction est $exp(Z_j \beta_z)$; cette fonction dépend des caractéristiques individuelles et est indépendante du temps (Cleves et al., 2010; Kleinbaum, 1996).
  La fonction de risque pour tout t est donc le produit de ces deux fonctions. C'est un modèle qui appartient à la catégorie des modèles explicatifs (Le Goff, 2003).

La particularité de ce modèle réside dans le fait qu'on ne fait aucune hypothèse en ce qui concerne la distribution de $h_0(t)$. Pour rappel, $h_0(t)$ représente la fonction de hasard lorsque la matrice Z est nulle ou en d'autres termes la fonction de hasard pour les individus de référence.

C'est un modèle à hasard proportionnel parce que le rapport des fonctions de risque entre deux individus ayant des caractéristiques individuelles différentes est constant et indépendant du temps et du risque de base $h_0(t)$:

$$\frac{h(t, z_i)}{h(t, z_j)} = \frac{h_0(t)exp(Z_i\beta)}{h_0(t)exp(Z_j\beta)} = exp\left[(z_i - z_j)\beta\right]$$

Le modèle de Cox peut approximer des modèles paramétriques comme les modèles exponentiel, de Weibull ou de Gompertz. Dans le modèle semi-paramétrique de Cox, lorsque l'on spécifie la forme de $h_0(t)$, on retombe sur les modèles paramétriques selon la forme donnée à $h_0(t)$.

1. Si $h_0(t) = e^{\beta_0}$, on retrouve le modèle exponentiel.
2. Si $h_0(t) = exp(\gamma t)exp(\beta_0)$, on retrouve le modèle de Gompertz.
3. Si $h_0(t) = pt^{p-1}exp(\beta_0)$, on retrouve le modèle de Weibull.

A l'image des autres modèles d'analyse des parcours de vie, pour réaliser un modèle de Cox, on a besoin des variables temps et censure. La variable temps nous indique le temps écoulé depuis le début de l'observation jusqu'à l'occurrence de l'événement étudié, si celui-ci a eu lieu, ou jusqu'à la sortie d'observation, s'il n'a pas eu lieu. Par exemple, si l'on s'intéresse à l'obtention du diplôme de master par des étudiants internationaux en Suisse depuis le début du bachelor jusqu'à la fin du master, le début de l'observation correspond au début des études de bachelor et la fin de l'exposition correspond à l'obtention du diplôme de master. L'écart entre la date de début du bachelor et la date de fin du master correspond à la durée jusqu'à l'obtention du master.

La variable événement prendra la valeur 1 si l'événement d'intérêt a lieu et la valeur 0 sinon. En plus de ces deux variables, on a ensuite besoin des variables explicatives. La structure des données est donc identique à celle des modèles paramétriques et se présente comme dans le tableau 11.1.

Tableau 11.1 – Format de la base de données pour faire une régression de Cox

Identifiant	Début	Fin	Evénement	x_{1i}	x_{2i}
1	date début	date fin	1	x_{11}	x_{21}
2	date début	date fin	1	x_{12}	x_{22}
⋮	⋮	⋮	⋮	⋮	⋮

11.1 Tests statistiques

On peut tester la significativité des coefficients à l'aide de plusieurs tests statistiques dont entre autres, le test de Wald, le test du rapport de vraisemblance ou le test du log-rank. De manière générale, l'hypothèse nulle est formulée de la manière suivante :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

On rejettera H_0 si au moins un des paramètres est significativement différent de 0. Le rejet de H_0 implique que les variables explicatives associées à ces paramètres ont un impact significatif sur le risque ou sur le logarithme du risque.

11.2 Exemple d'application

Nous allons considérer la même question de recherche que pour les modèles paramétriques, l'objectif étant de répondre à cette dernière en utilisant le modèle de régression semi-paramétrique de Cox.

Question de recherche : Le risque d'obtenir le diplôme de bachelor (dans notre cas c'est en fait la chance d'obtenir le diplôme de bachelor) des étudiants internationaux dépend-elle du sexe et de l'âge de ces étudiants ?

Le tableau des coefficients présenté dans le tableau 11.2 montre que les p-valeurs associées aux variables sexe et nationalité sont toutes deux inférieures à 0.05 et on peut conclure que ces paramètres sont significativement différents de 0, ce qui corroborre notre hypothèse de recherche.

Tableau 11.2 – Tableau des coefficients du modèle de Cox

Variable	Coeff.	Erreur standard	z	P-valeur	IC à 95%
sexe	0.1047	0.0475	- 4.97	0.001	0.040 0.2548
âge	1.111	0.0420	2.78	0.005	1.0316 1.1964

La première remarque qu'on peut faire sur le tableau des coefficients est que le modèle de Cox n'a pas de constante. En effet, la constante du modèle est absorbée par la partie non paramétrique du modèle, c'est-à-dire par $h_0(t)$; Ce qui est équivalent à définir une nouvelle constante du modèle que l'on pourrait écrire comme $h_0(t)\beta_0$ (Cleves et al., 2010).

On peut écrire la fonction de risque pour tout t de la manière suivante :

$$log\left[\widehat{h(t, Z_j)}\right] = 0.1047sexe + 1.111age$$

Cette équation nous permet de faire des prédictions selon le sexe et l'âge des étudiants. Les coefficients des variables âge et sexe sont tous positifs, cela veut dire que ces deux variables ont un impact positif sur le risque d'obtenir le diplôme de bachelor. Ansi, quel que soit t, lorsque l'âge augmente d'une unité, le risque d'obtenir le diplôme de bachelor est environ $e^{1.111} = 3.037$ fois plus susceptible de se produire. Par ailleurs, être une femme par rapport à être un homme, pour le même âge, augmente le risque d'obtenir le diplôme de bachelor d'environ $e^{0.1047} = 1.1103$ fois ; ce qui correspond à une augmentation de ce risque d'environ 11.04%. Le tableau 11.3 résume le test du rapport de vraisemblance et comme on peut le remarquer, toutes les différences de déviances sont supérieures à 3.84, ce qui nous permet de rejeter H_0 au seuil de 5% et de conclure que l'hypothèse de recherche est corroborée.

Tableau 11.3 – Tableau résumant le test du rapport de vraisemblance pour le modèle de Cox

Modèles	Déviances	Différence de déviances	Comparaison	Conclusion
Complet	82.4702			
Sans l'âge	86.3690	3.8988	3.8988 > 3.84	On rejette H_0 au seuil de 5%
Sans le sexe	96.5366	14.0664	14.0664 >3.84	On rejette H_0 au seuil de 5%

Les modèles paramétriques et le modèle semi-paramétrique de Cox ayant aussi leurs limites, on se penchera dans le chapitre suivant sur un autre type de modèle qui s'appelle le modèle logit à temps discret. En effet, les modèles paramétriques et le modèle semi-paramétrique de Cox sont sensibles aux groupes d'égalité dès lors que 5% des individus sous observation expérimentent au même moment l'événement étudié. Pour aider à mieux comprendre cette notion de groupes d'égalité, prenons un exemple de cent étudiants qui commencent une formation de master dans une faculté quelconque. Si l'on s'intéresse au passage en classe supérieure de ces étudiants, les modèles paramétriques et le modèle semi-paramétrique de Cox peuvent conduire à l'estimation de paramètres biaisés dès lors que cinq étudiants (100*0.05) passent en classe supérieure la même année. En réalité, sur cent étudiants, il est plus que probable que plus de cinq étudiants passeront en classe supérieure la même année. Il nous faut donc penser à d'autres modèles qui ne sont pas sensibles aux groupes d'égalité. Le modèle de régression logit à temps discret qui fait partie des modèles de régression logistiques à temps discret appartient à ce groupe de modèles.

Ces modèles, ainsi que leurs avantages par rapport aux modèles paramétriques et au modèle semi-paramétrique de Cox seront présentés dans le chapitre suivant.

Chapitre 12

Les modèles de régression logistiques à temps discret

Nous avons vu dans les chapitres précédents que les modèles paramétriques et le modèle semi-paramétrique de Cox (Cox, 1972) sont représentés en temps continu ; cela implique que deux personnes ou plus ne peuvent pas connaître l'événement d'intérêt au même moment. Le fait que deux ou plusieurs personnes connaissent l'événement d'intérêt au même moment s'appelle groupes d'égalité ou ties en anglais. Lorsque 5% des individus connaissent l'événement d'intérêt au même moment, cela peut déjà entraîner un biais dans l'estimation des paramètres par les modèles paramétriques et semi-paramétrique (Yamagushi, 1991; Vermunt, 1997).

En effet, dans la réalité, les événements sont mesurés de manière discrétisée (Allison, 1982), cela implique qu'un ou plusieurs événements peuvent être observés au même moment pour un ou plusieurs individus. On peut citer, comme exemples, le nombre de mariages célébrés en un mois dans une commune ou dans une région donnée (si l'unité de temps considéré est le mois, il est fort probable qu'il y ait plusieurs mariages célébrés en un mois dans une commune ou dans une région), ou le nombre d'étudiants ayant obtenu le diplôme de bachelor dans une université en une année.

En reprenant l'exemple sur l'obtention du diplôme de bachelor dans une université, pour des étudiants qui ont commencé les études au même moment, plusieurs étudiants achèveront leur formation la même année. On parlera de groupe d'égalité pour désigner ceux qui ont terminé les études au cours de la même année. Lorsqu'il y a des groupes d'égalités, les modèles dans lesquels le temps est discrétisé doivent être préférés aux modèles paramétriques et au modèle de Cox ; modèles dans lesquels l'intérêt porte sur le risque de connaître l'événement d'intérêt au cours du temps en tenant compte des caractéristiques individuelles.

Pour pallier les insuffisances des modèles paramétriques et de Cox, il existe un autre groupe de modèles qui sont représentés en temps discret et qui peuvent être ajustés sans aucun problème, même en présence de groupes d'égalités ; ce sont les modèles à temps discret. Ces modèles ont été introduits dans les sciences sociales par Allison (Allison, 1982, 1984). Parmi ces modèles, le plus utilisé est le modèle logit à temps discret ou « discrete time logit model » en anglais (Le Goff et al., 2013).

12.1 Le modèle

Dans les modèles paramétriques et le modèle de Cox le risque d'occurrence de l'événement d'intérêt est une fonction du temps et des caractéristiques individuelles. En temps discret, ce risque est une probabilité qui est aussi fonction du temps et des caractéristiques individuelles (âge, nationalité, sexe, état civil,...) et peut être formulée comme suit (Le Goff et al., 2013).

$$p(t) = f(t, Z_t) \tag{12.1}$$

Z_t représente la matrice des caractéristiques individuelles et peut s'écrire comme
$Z_t = (z_{t1}, z_{t2}, z_{t3}, \ldots, z_{tj})$; ces caractéristiques individuelles pouvant être quantitatives ou qualitatives et

dépendantes ou non du temps. Lorsqu'une variable est accompagnée de l'indice t, cela veut dire que cette variable est susceptible de changer au cours du temps (état civil, niveau de formation, nombre d'enfants, etc.). Ce risque représente la probabilité conditionnelle de connaître l'événement d'intérêt à chaque instant t sachant qu'on ne l'avait pas connu à l'instant précédent.

12.2 L'hypothèse de proportionnalité des risques en temps continu et en temps discret

En temps continu, le rapport des fonctions de risque pour deux individus ayant des caractéristiques individuelles différentes est constant et indépendant du temps ; on parle alors de proportionnalité des risques. La proportionnalité des risques implique qu'il est possible d'écrire par exemple que le risque pour l'individu un est le double du risque pour l'individu deux quel que soit le temps auquel on se situe (indépendance du temps). Pour mieux illustrer la notion de proportionnalité, on peut citer comme exemple le rapport entre l'âge d'un chien et l'âge d'un être humain (Allison, 2010). De manière générale, le rapport entre l'âge d'un chien et l'âge d'un être humain est de 7 ; un chien âgé de 5 ans correspond ainsi en termes d'âge humain à une personne âgée de 35 ans. On peut écrire cela de la manière suivante : âge chien/ âge humain =7 ou âge chien =7 fois âge humain. La proportionnalité des risques fonctionne exactement de cette manière. Comme déjà mentionné dans la section 4.1 ou dans les chapitres 5, 6 et 7, l'hypothèse de proportionnalité implique que la fonction de risque peut s'écrire de la manière suivante (Cleves et al., 2010; Le Goff et al., 2013).

$$h(t) = h_o(t)e^{b_j Z_j} \tag{12.2}$$

Z_j représentant la matrice des prédicteurs ou matrice des cartéristiques individuelles et b_j le vecteur colonne des paramètres du modèle.

Comme pour les modèles paramétriques ou pour le modèle de Cox, le produit matriciel $Z_j b_j$ s'écrit :

$$Z_j b_j = z_1 b_1 + z_2 b_2 + ... + z_n b_n$$

En termes de logarithme, la fonction de risque s'écrit de la manière suivante :

$$log\left[h(t)\right] = log\left[h_o(t)\right] + \sum_{j=1}^{k} Z_j b_j \tag{12.3}$$

En temps discret, le risque représente la probabilité conditionnelle de connaître l'événement d'intérêt à un instant t sachant qu'on ne l'avait pas connu à l'instant précédent. Cette probabilité est notée ici par $p(t, Z_t)$. Une probabilité prend valeur dans l'intervalle $[0, 1]$ et de ce fait ne peut pas être supérieure à 1. Dans un modèle à risque proportionnel, il est possible d'estimer des paramètres qui peuvent conduire à l'obtention d'une probabilité conditionnelle de connaître l'événement d'intérêt qui soit supérieure à 1. Le premier problème que pose l'hypothèse de proportionnalité en temps discret est le fait qu'il soit possible d'obtenir une probabilité conditionnelle de connaître l'événement d'intérêt à un instant donné qui soit supérieure à 1. Dès lors, le temps discret n'est plus compatible avec la proportionnalité des risques.

En temps continu, le risque peut être supérieur à 1 dans la mesure où ce n'est plus une probabilité qui est calculée, mais un risque. Il nous faut donc une fonction lien qui puisse contraindre ce risque dans l'intervalle $[0, 1]$, de manière à ce qu'on obtienne une probabilité, d'où la fonction logit. La fonction logit représente le logarithme du rapport entre $p(t, Z_t)$ et $(1 - p(t, Z_t))$ qui est une fonction linéaire des caractéristiques individuelles.

Le modèle logit s'écrit de la manière suivante (Le Goff et al., 2013) :

$$log\left(\frac{p(t, z_t)}{1 - p(t, z_t)}\right) = log\left(\frac{p_0}{1 - p_0}\right) + \sum_{j=1}^{k} Z_j b_j \tag{12.4}$$

p_0 est la probabilité pour les individus à caractéristiques individuelles nulles, b_j, est le vecteur des paramètres à estimer et Z_j, la matrice des caractéristiques individuelles.

L'exponentielle des coefficients est appelée odds ratio. L'interprétation du odds ratio se fait selon la nature de la variable ; si la variable est quantitative, il nous informe de combien augmente ou diminue la chance de connaître l'événement d'intérêt au moment t lorsque la variable explicative concernée augmente d'une unité, toutes choses étant égales par ailleurs. Lorsque la variable est qualitative, l'odds ratio nous informe de combien augmente ou diminue le risque ou la chance de connaître l'événement d'intérêt lorsqu'on appartient à une catégorie plutôt qu'à l'autre. Sa valeur est comprise entre 0 et plus infini (∞). Lorsque l'odds ratio est > 1, cela veut dire que la variable a un impact positif sur la probabilité d'occurrence de l'événement d'intérêt, lorsqu'il est < 1, l'impact est négatif et lorsqu'il vaut 1, il n'y a pas d'impact.

Le modèle logit tel que présenté dans l'équation 12.4 ci-dessus se compose de deux parties : une première partie qui est constante $log(p_0/1 - p_0)$ et une deuxième partie qui dépend des caractéristiques individuelles $(\sum_{j=1}^{k} Z_j b_j)$. En posant que $log(p_0/1 - p_0) = a$ et $\sum_{j=1}^{k} Z_j b_j = bZ_j$, le modèle logit peut s'écrire comme suit :

$$log\left(\frac{p(t, z_t)}{1 - p(t, z_t)}\right) = a + bZ_j \tag{12.5}$$

Le modèle logit tel que spécifié implique que les caractéristiques individuelles des individus de références sont non seulement constantes mais aussi indépendentes du temps d'une part, d'autre part que les caractéristiques individuelles des autres individus ne dépendent pas du temps. Les individus de référence sont un groupe d'individus auquel tous les autres groupes d'individus sont comparés. Mais, en réalité, il n'y a aucune raison de supposer que les caractéristiques individuelles des individus de références sont constantes et indépendantes du temps parce que cela reviendrait à dire que la probabilité conditionnelle de connaître l'événement d'intérêt est la même quel que soit t. On peut dès lors écrire le modèle logit en considérant que les caractéristiques individuelles des individus de références peuvent varier dans le temps et que les caractéristiques individuelles de manière générale peuvent aussi dépendre du temps (Allison, 1982).

$$log\left(\frac{p(t, Z_t)}{1 - p(t, Z_t)}\right) = a(t) + bZ_t \tag{12.6}$$

Puisque $a(t)$ peut varier dans le temps, on peut observer plusieurs situations.

1. Première situation : $a(t) = c + bt$
   Dans cette première situation, $a(t)$ est une fonction linéaire du temps et le modèle logit n'est autre que le modèle de Gompertz. Le modèle de Gompertz étant un modèle en temps continu, nous n'explorerons pas davantage ce modèle. En effet, $a(t)$ est l'expression d'une droite de pente b (pente de la fonction de Gompertz) (Le Goff et al., 2013) ; lorsque la pente est positive, la fonction de risque est croissante et lorsque la pente est négative, la fonction de risque est décroissante et lorsque $b = 0$, la fonction de risque est constante (modèle exponentiel). Dans la réalité, le risque n'est pas uniformément réparti de manière croisssante ou décroissante sur un intervalle de temps, il peut être croissant puis décroissant et inversément.
2. Deuxième situation : $a(t) = c + blog(t)$
   Dans cette deuxième situation, le modèle logit n'est autre que le modèle de Weibull qui est aussi un modèle en temps continu.

Dans le modèle de Gompertz tout comme dans le modèle de Weibull, les odds ratios sont une fonction monotone du temps ; une fonction est dite monotone si elle est croissante ou décroissante sur un intervalle donné. Cela revient à dire que le risque d'occurrence de l'événement d'intérêt au cours du temps est croissant ou décroissant, ce qui peut poser problème dans la mesure où il a été observé que pour certains événements du parcours de vie, le risque pouvait être croissant puis décroissant (Le Goff et al., 2013; Diekmann, 1990).

Le modèle Piecewise constant en temps continu (Blossfeld et Rohwer, 2002) est un modèle qui permet de décomposer le risque d'occurrence de l'événement d'intérêt sur de petits intervalles de temps. Ce modèle

suppose que le logit de la probabilité conditionnelle de connaître l'événement d'intérêt à un instant t sachant qu'on ne l'avait pas connu à l'instant précédent peut être constant sur un intervalle puis être croissant ou décroissant sur un autre intervalle. En d'autres termes, le logit de cette probabilité varie d'un intervalle à un autre en prenant diverses formes. Dans ce modèle, chaque intervalle de temps créé est une variable binaire qui prend la valeur 1 si on appartient à cet intervalle et la valeur 0 sinon ; il y aura donc autant de paramètres à estimer qu'il y a d'intervalles créés.
Ce modèle s'écrit comme suit :

$$log\left(\frac{p(t,z_t)}{1-p(t,z_t)}\right) = c + \sum_{j=1}^{k} d_j a_j + bZ_t \tag{12.7}$$

Où :

c est une constante,
a_j est un paramètre qui sera constant sur un intervalle $(t, t+t_j)$ donné puis pourra prendre une autre forme (selon sa valeur) sur un autre intervalle , d_j est une variable binaire qui prendra la valeur 1 si on se situe dans l'intervalle en question ou la valeur 0 sinon. Par exemple, supposons que l'on divise la variable temps en trois intervalles (t_0, t_{10}), (t_{10}, t_{20}), (t_{20}, t_{30}), la variable d_j prendra la valeur 1 si on se situe dans le premier intervalle, 0 sinon, la valeur 1 si on se situe dans le deuxième intervalle, 0 sinon, et enfin la valeur 1 si on se situe dans le troisième intervalle, 0 sinon. Il y aura autant de variables dichotomiques que d'intervalles créés. C'est un modèle qui présente une grande facilité d'utilisation dans le sens où aucune hypothèse n'est faite en ce qui concerne la distribution de la probabilité conditionnelle d'occurrence de l'événement d'intérêt au cours du temps (Le Goff et al., 2013).

12.3 Estimation par la méthode du maximum de vraisemblance

La vraisemblance mesure l'écart entre les données observées et le modèle. Plus cet écart est petit, plus précis sera le modèle. A l'image de la régression logistique, les différents paramètres (c, $a(t)$, b) du modèle sont estimés par la méthode du maximum de vraisemblance qui est une méthode qui vise à maximiser la probabilité d'obtenir à partir du modèle, les données observées. Dans l'analyse des parcours de vie, un individu i soumis au risque de connaître un événement quelconque peut connaître cet événement ou ne pas le connaître. La contribution à la vraisemblance pour un individu quelconque se fait à travers $f(t_i)$ si l'individu connaît l'événement d'intérêt ou avec $S(t_i)$ s'il ne le connaît pas, $f(t_i)$ et $S(t_i)$ représentant respectivement les fonctions de densité et de survie. L'équation de la vraisemblance pour une population de taille n soumise au risque de connaître un événement d'intérêt quelconque est égale au produit de la contribution à la vraisemblance pour chaque individu composant cette population (Le Goff et al., 2013) :

$$L = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \tag{12.8}$$

Avec $\delta_i = 1$ si l'individu a connu l'événement d'intérêt , 0 sinon.

En temps discret, $f(t_i) = P(T_i = t_i)$ et $S(t_i) = P(T_i > t_i)$, Allison (Allison, 1982) propose d'écrire l'équation de la vraisemblance comme :

$$L = \prod_{i=1}^{n} [P(T_i = t_i)]^{\delta_i} [P(T_i > t_i)]^{1-\delta_i} \tag{12.9}$$

En procédant à quelques transformations sur $P(T_i = t_i)$ et sur $P(T_i > t_i)$, on peut écrire l'équation de la log vraisemblance comme suit.

$$logL = \sum_{i=1}^{n} \sum_{k=1}^{t_i} y_{it} log\left[\frac{P_{ik}}{1-P_{ik}}\right] + \sum_{i=n}^{n} \sum_{k=1}^{t_i} log(1-P_{ik}) \tag{12.10}$$

y_{it} étant une variable aléatoire qui prendra la valeur 1 si l'individu connaît l'événement d'intérêt au temps t et 0 s'il ne le connaît pas.

L'estimation des paramètres à partir de cette équation de vraisemblance nécessite l'organisation des données sous la forme de personnes-période (Le Goff et al., 2013).

12.4 Organisation des données pour mettre en pratique les modèles logit à temps discret

Pour pouvoir mettre en application le modèle logit à temps discret, la base de données doit être organisée sous la forme de personne-période. Si l'unité de temps considérée est le mois, on parle de personne-mois et si l'unité de temps considérée est l'année, on parlera de personne-année (Le Goff et al., 2013). En considérant que l'unité de temps est l'année, chaque ligne représentera une année d'observation pour les individus de la base de données. Cette base de données doit être organisée de manière à ce qu'un individu apparaisse autant de fois qu'il a été soumis à observation et le nombre de lignes pour cet individu correspondra au nombre d'années de présence dans l'étude avant de connaître l'événement d'intérêt s'il l'a connu ou de sortir de l'étude s'il ne l'a pas connu. Supposons que l'on s'intéresse à la survenue du mariage pour un étudiant international cinq ans après son arrivée en Suisse. L'événement d'intérêt est donc le mariage ; cette variable « mariage » prendra la valeur 1 si l'étudiant s'est marié, 0 sinon. Si cet étudiant ne s'est pas marié durant les cinq ans, la variable dichotomique « mariage » prendra la valeur 0 sur les cinq ans de présence en Suisse pour cet étudiant, c'est-à-dire, l'identifiant de cet étudiant va se répéter sur cinq lignes et la variable mariage prendra la valeur 0 sur ces cinq lignes. Si l'étudiant s'est marié la cinquième année, la variable dichotomique « mariage » prendra la valeur 0 sur les quatre premières lignes puis la valeur 1 sur la cinquième ligne parce que l'étudiant a connu l'événement d'intérêt à la cinquième année avec à chaque fois ses autres caractéristiques individuelles variant ou non avec le temps.

Tableau 12.1 – Exemple de données préparées sous le format de personne-période (personne-année)

id	temps	mariage	sexe	nationalité	age
10	1	0	1	Guinée	18
10	2	0	1	Guinée	19
10	3	0	1	Guinée	20
10	4	0	1	Guinée	21
10	5	1	1	Guinée	22
11	1	0	1	Cameroun	25
11	2	0	0	Cameroun	26
11	3	0	0	Cameroun	27
11	4	0	0	Cameroun	28
12	1	0	0	Sénégal	22
12	2	0	1	Sénégal	23
12	3	1	1	Sénégal	24
⋮	⋮	⋮	⋮	⋮	⋮

Dans cet exemple fictif présenté dans le tableau 12.1, nous nous intéressons à l'occurrence du mariage pour des étudiants africains en Suisse mais nous mettrons en évidence seulement le cas de trois étudiants de nationalité guinéenne, camerounaise et sénégalaise. On remarque que pour l'étudiant ayant l'identifiant « 10 », le mariage est survenu à la cinquième année ce qui explique le fait que les quatre premières lignes pour la variable mariage sont des 0 et la cinquième prend la valeur 1. L'étudiant ayant l'identifiant « 11 » ne s'est pas marié durant sa période de présence en Suisse et de ce fait la variable mariage prend la valeur 0 pour toutes les lignes correspondant à cet étudiant. Le troisième étudiant ayant l'identifiant « 12 » s'est marié à sa troisième année et par conséquent la variable mariage prend la valeur 0 pour les deux premières années.

Pour mettre en pratique ces différents modèles, il nous faut des données. Les données qui font l'objet de cette thèse sont des données longitudinales administratives. Ces données administratives, ainsi que les différents travaux de préparation de ces dernières seront présentées dans la partie qui va suivante.

Troisième partie

Données et potentialités d'application des méthodes présentées aux données administratives de l'OFS

Chapitre 13

Données : Présentation, préparation et fusion des données de l'OFS

Ce chapitre présente les différentes sources de données longitudinales, leurs avantages et leurs inconvénients, la préparation et la fusion des données de l'OFS, le traitement des doublons, ainsi que la sélection de l'échantillon des étudiants africains.

13.1 Les sources de données longitudinales

Les données longitudinales sont des données recueillies à plusieurs reprises dans le temps sur les mêmes individus qui composent la population d'étude (Lynn, 2009). Une cohorte est un ensemble de personnes soumises à un même événement sur un même intervalle de temps (par exemple, l'ensemble des étudiants internationaux qui ont commencé le bachelor en 2016 en Suisse). Les données longitudinales se distinguent des données transversales par la possibilité qu'elles offrent d'étudier l'évolution des individus dans le temps. Les données transversales, elles, ne permettent d'étudier la population d'intérêt qu'à un moment précis et n'offrent pas la possibilité de suivre ces individus dans le temps (Tourangeau, 2003). Les données longitudinales sont recueillies de manière rétrospective, prospective ou de manière administrative (données de registre) comme ce sera le cas dans cette thèse.

13.2 Les enquêtes rétrospectives

Les données issues d'enquêtes rétrospectives sont des données récoltées en faisant appel à la mémoire des personnes faisant l'objet de l'enquête. Les enquêtes rétrospectives permettent d'observer l'évolution d'un individu dans le temps à partir d'un questionnaire dans le but de reconstruire les trajectoires de cet individu. On demandera par exemple à une personne de raconter son histoire de vie à partir d'une période précise de sa vie jusqu'au moment de l'enquête. En faisant référence à la mémoire pour récolter ce type de données, on fait face à un risque élevé que ces données contiennent des erreurs ; plus la mémoire doit faire appel à des souvenirs lointains, plus le risque que les données contiennent des erreurs est élevé (Courgeau et Lelièvre, 1993). Les oublis de certains événements importants du parcours de vie posent aussi problème dans la collecte de ce type de données. La précision des questions posées par les enquêteurs est très importante pour réduire les erreurs dans les données issues d'enquêtes rétrospectives. Pour illustrer les incohérences des réponses dans les enquêtes rétrospectives, on peut citer comme exemple une étude qui a été réalisée par l'INED et l'Université Catholique de Louvain (Poulain et al., 1991) auprès de cinq cents couples mariés de nationalité belge à la naissance et vivant toujours avec le premier conjoint. Le questionnaire a été soumis simultanément au couple dans deux pièces différentes par deux enquêteurs distincts dans le but de s'assurer de l'indépendance dans la collecte et dans les réponses données. A la fin du questionnaire, le couple est confronté aux réponses données cette fois-ci en face à face dans le but de déceler les divergences de réponses pour la même histoire de vie. Les

deux réponses individuelles et les réponses issues de la confrontation sont ensuite comparées aux données de registre [1] qui enregistrent de manière précise tous les événements d'état civil et les mouvements à l'intérieur de la Belgique avec les datations précises. Les résultats obtenus montrent des divergences dans les réponses données par certains couples et ce qui est réellement enregistré par les autorités d'une part ; d'autre part, les réponses données par les femmes étaient plus précises que celles de hommes. S'agissant de la date du mariage, certaines personnes ont mentionné la date du mariage religieux au lieu de la date du mariage civil. Des erreurs sur les dates de naissances des enfants ont aussi été constatées. Certes, le taux de mortalité infantile est très bas en Europe, notamment en Belgique mais les enquêteurs n'ont pas exclu que des couples aient oublié de mentionner des enfants décédés en bas âge car le registre de l'état civil indiquait que les couples interrogés ont eu 1078 enfants au total dont 14 enfants étaient décédés à la date de l'enquête. Parmi ces enfants décédés, certains ont été cités par les deux parents, d'autres par l'un des deux parents, mais, il y a tout de même eu neuf omissions d'enfants morts en bas âge. Des erreurs ont aussi été constatées dans la datation des migrations des couples.

Cet exemple a permis de mettre en évidence deux problèmes liés aux enquêtes rétrospectives : soit on omet des informations par oubli soit on cite les événements mais avec une mauvaise datation. On peut aussi mentionner un impact du genre sur la fiabilité des résultats dans l'histoire de vie familiale, les femmes (Courgeau, 1991; Auriat, 1991) donnant souvent des réponses plus précises que les hommes. Pour améliorer la qualité des données issues d'enquêtes rétrospectives, il est nécessaire de procéder à une confrontation entre époux ou membres de la famille si l'étude concerne les familles et ensuite de confronter les résultats finaux obtenus au registre de population si celui-ci existe et est bien tenu. Si l'enquête ne concerne pas la famille, il serait très prudent de confronter les résultats issus de l'enquête à d'autres sources de données pour s'assurer de la chronologie de certains événements si l'on doit se fier au récit d'une seule personne.

13.2.1 Avantages et inconvénients des enquêtes rétrospectives

Les avantages des enquêtes rétrospectives résident dans le fait que ce type d'enquêtes coûte moins cher en ressources financières et en temps. En effet, dans ce type d'enquête, un seul passage suffit pour retracer le parcours de vie de la personne et cela permet un gain de temps pour les enquêteurs et pour les enquêtés. Le risque de voir des personnes disparaître de l'étude est très faible et les changements d'adresse ou les départs de la zone géographique de la recherche n'impactent pas ce type d'enquête en raison du passage unique des enquêteurs (Courgeau et Lelièvre, 1989). Les inconvénients des enquêtes rétrospectives résultent du fait que l'on se fie à la mémoire des personnes enquêtées, ce qui peut entraîner des risques d'oublis de certains événements et des erreurs de datation (Courgeau, 1991).

13.3 Les enquêtes prospectives

Les enquêtes prospectives consistent à suivre une cohorte sur une période donnée qui peut être exprimée en mois ou en année, e but de ce suivi étant de comprendre l'évolution de cette cohorte entre le début et la fin de la période d'observation. On peut citer comme exemples : le Panel Suisse des Ménages (PSM) qui étudie l'évolution du revenu des ménages en procédant à des enquêtes répétées dans le temps ou l'enquête auprès des diplomés des hautes écoles (EHA) qui s'intéresse à la situation professionnelle des personnes diplômées des hautes écoles suisses un an et cinq ans après l'obtention du diplôme. Dans les enquêtes prospectives, on sous-entend par exemple les enquêtes de panel, les essais cliniques et les études de cohortes (Lindsey, 1999).

Les premières enquêtes prospectives de cohortes prennent leurs origines dans les études de Framingham dont l'objectif était d'établir un lien entre la consommation de cigarettes et le cancer des poumons entre 1947 et 1949 (Aronowitz, 2011) d'une part, d'autre part, par les études menées simultanément par Doll et Hill (Doll et al., 2004) en Grande-Bretagne et par Cuyler Hammond et Daniel Horn en 1950 concernant le cancer des poumons

1. La Belgique est dotée d'un registre de population depuis 1847. Ce registre est tenu de manière décentralisé par chaque unité administrative et enregistre tous les événements d'état civil et leurs datations, tous les changements de domicile, ainsi que les entrées et les sorties de la Belgique. Ce registre a été informatisé en 1970

et des maladies cardio-vasculaires. Dans l'étude sur le cancer des poumons, l'idée de faire une étude prospective est née suite à la publication de plusieurs études qui ont montré l'existence d'une corrélation positive entre le cancer des poumons et le tabagisme. Mais ces résultats provenaient d'une base de données construite en interrogeant de personnes malades dans le but de reconstituer leurs histoires de vie et leurs habitudes de consommation. Ces études ont été critiquées car on ne pouvait pas généraliser les conclusions obtenues par ces études, pointant du doigt non seulement la qualité des données mais aussi le fait que le cancer des poumons serait généré par d'autres facteurs non pris en compte par l'étude. Les critiques étaient aussi orientées vers le fait que la connaissance du diagnostic pouvait aussi entraîner un biais dans les réponses en ce qui concerne les habitudes de consommation du tabac (Giroux, 2011). La première grande étude prospective réalisée en 1950 avait donc pour objectif d'apporter des éléments de preuves plus concrètes permettant de généraliser les résultats obtenus antérieurement par enquête rétrospectives. Dans cette étude prospective, un questionnaire a été envoyé à $60'000$ médecins du registre médical concernant leurs habitudes tabagiques (Doll et Hill, 2004) . Ces médecins ont été suivis dans le temps en s'intéressant à la survenue d'un cancer des poumons et au décès des médecins en utilisant les registres de naissance et de mortalité. Les auteurs ont obtenu $40'000$ réponses à partir desquelles ils constituèrent des groupes d'exposition : un groupe de non-fumeurs et trois groupes de fumeurs selon le nombre de cigarettes consommées sur la même unité de temps. En 1954, les premiers résultats qui ont été publiés montraient que sur 789 décès, 35 étaient liés au cancer des poumons. Malgré ce nombre faible de décès observés, les taux de mortalités des différents groupes selon l'âge a mis en évidence une augmentation significative des décès par cancer des poumons en fonction de l'augmentation de la consommation de tabac. Cette étude prospective a continué jusqu'en 2002 avec de nombreuses publications autour d'elle.

De cet exemple sur l'origine des études prospectives, il ressort que malgré le fait que les enquêtes prospectives soient de nos jours plus répandues que les enquêtes rétrospectives, ces dernières ont été à l'origine des enquêtes prospectives. Il permet aussi de mettre en évidence le fait que dans les enquêtes prospectives on fait des mesures répétées des mêmes variables sur les mêmes individus au cours du temps (Gravlee et al., 2009) à l'exception des enquêtes prospectives de type panels rotatifs. Dans les enquêtes de types panels rotatifs, les premières vagues des personnes enquêtées sont supprimées et remplacées par des personnes à caractéristiques plus ou moins similaires après un certain temps dans le but de reposer les personnes enquêtées et de s'assurer de la continuité de l'enquête dans le temps (Menard, 2002). A travers les avantages qu'elles procurent, les enquêtes prospectives ont connu, au cours de ces 50 dernières années, une ascension rapide dans le domaine des sciences sociales, de la médecine, de l'économie ainsi que dans le domaine de l'éducation (Gravlee et al., 2009). Si les enquêtes prospectives offrent de nombreux avantages, elles présentent aussi quelques inconvénients ; ils sont présentés dans les sections suivantes.

13.3.1 Avantages des enquêtes prospectives

Parmi les avantages des enquêtes prospectives, on peut mentionner :

Possibilité de suivre l'évolution de la population étudiée dans le temps

Le premier avantage des enquêtes prospectives réside dans la possibilité qu'elles offrent d'étudier et de suivre l'évolution des individus au cours du temps ainsi que les différentes transitions qu'ils connaissent. Singer et Willet (Carolyn J., 2005) résument l'importance de suivre les transitions au niveau des individus soumis à observation par la possibilité qu'offre ce suivi d'observer d'une part, les changements au niveau des individus, d'autre part, les changements entre les groupes d'individus.

Par exemple, supposons que l'on s'intéresse à l'ensemble des étudiants africains qui ont commencé un bachelor en 2010 en Suisse et qui ont obtenu le diplôme de bachelor en 2013 ou en 2014. Il sera possible d'observer l'évolution de chaque étudiant par année d'une part, d'autre part, il sera aussi possible de comparer les caractéristiques des étudiants qui ont terminé le bachelor en trois ans et les caractéristiques de ceux qui l'ont terminé en quatre ans.

Connaissance de la chronologie de l'occurrence des événements

Les enquêtes prospectives permettent aussi de connaître la chronologie des événements survenus pendant la période d'observation (Gravlee et al., 2009). Au cours de cette période d'observation, il est possible d'étudier simultanément plusieurs événements d'intérêts.

Considérons comme exemple, l'ensemble des étudiants internationaux arrivés en Suisse en 2005 pour faire un master ; on s'intéresse à l'évolution de cette cohorte dix ans après l'arrivée en Suisse. Durant cette période d'observation, on peut étudier simultanément plusieurs événements comme l'obtention du bachelor, l'obtention du master ou d'un doctorat, le mariage, le premier emploi ou la naissance du premier enfant.

La fiabilité et la précision de la mesure

Les enquêtes prospectives améliorent la qualité des données recueillies et la précision des paramètres estimés de deux manières ; la première est le fait que les enquêtes prospectives permettent de contrôler les caractéristiques individuelles invariables dans le temps et la deuxième réside dans le fait qu'elles permettent de réduire les biais de mémoire (un défaut des enquêtes rétrospectives) (Gravlee et al., 2009).

Le suivi des individus au cours du temps entraîne un très grand volume de données à traiter ; le volume des données engendré par les enquêtes prospectives est proportionnel à la taille de l'échantillon et à la durée de suivi de la population d'intérêt (Dormont, 1989). Les enquêtes prospectives reposent sur des fondements statistiques solides comme la théorie des sondages ou sur des démarches probabilistes qui font que les enquêtes prospectives facilitent l'harmonisation du système de collecte et simplifie la comparaison internationale (Coulibaly, 2011).

13.3.2 Inconvénients des enquêtes prospectives

Les enquêtes prospectives présentent aussi quelques inconvénients qui sont entre autres :

Coûts et temps

Les enquêtes prospectives demandent un important investissement en temps (Coulibaly, 2011; Desrosières, 2011) notamment pour les personnes enquêtées et pour les enquêteurs. Ces enquêtes coûtent aussi très cher et ces coûts dépendent de la taille de l'échantillon, de la taille de la région géographique concernée par l'enquête et de la durée de l'enquête.

Taille de l'échantillon et durée de l'enquête

La taille de l'échantillon doit être grande pour obtenir des résultats fiables et la durée d'observation doit aussi être suffisamment grande pour pouvoir obtenir plusieurs vagues. Cependant, plus la durée d'observation est longue, plus le risque de non réponse est élevé du fait de la linéarité de répondre de manière répétitive au même questionnaire (Desrosières, 2011).

L'attrition

L'attrition représente le plus grand problème que l'on rencontre dans les enquêtes prospectives (Hillygus et Snell, 2015). On parle d'attrition lorsque des personnes disparaissent de l'étude avant la fin de celle-ci et ne répondent pas à certaines vagues de l'enquête, ce qui aura pour conséquence une diminution de la taille de l'échantillon initial. L'attrition peut être complètement aléatoire ou sélective (non aléatoire) ; l'attrition complètement aléatoire étant moins problématique pour les analyses que l'attrition sélective. Lorsque l'attrition est complètement aléatoire dans les enquêtes prospectives, cela entraîne des paramètres estimés moins précis mais non biaisés (Iglesias et al., 2017; Groves, 2006). On parle d'attrition sélective ou d'attrition non aléatoire lorsqu'il y a une non-réponse qui est corrélée à une variable d'intérêt de l'enquête (par exemple l'âge), ce qui

peut entraîner un biais de non-réponse qui aura pour conséquence de fausser les paramètres estimés et engendrer des conclusions erronées (Iglesias et al., 2017; Dormont, 1989; Mundlak, 1978). Le biais peut être corrigé par pondération ou par imputation des valeurs manquantes (Hillygus et Snell, 2015). Plusieurs méthodes de traitement de l'attrition sélective ont été proposées ; ces méthodes sont basées sur la pondération (Stoop, 2005) ou sur l'imputation. Dans les deux cas, ces méthodes supposent de connaître le processus qui a conduit à cette attrition ou de faire des hypothèses très fortes sur le processus qui a engendré cette forme d'attrition (Diggle et G. Kenward, 1994).

13.4 Les données administratives ou données de registre

Les Nations Unies (Nations-Unies, 2003) définissent les registres de population comme étant : « *un mécanisme d'enregistrement continu d'informations particulières concernant chaque membre de la population résidente d'un pays ou d'une région, ce qui permet de déterminer et d'actualiser les informations disponibles sur les caractéristiques de la population à certaines périodes données* ».

Les données administratives sont donc des données qui enregistrent l'ensemble de la population résidente d'une ville, d'une commune, d'une région ou d'un pays donné de manière individuelle. Les unités statistiques sont les individus et les mises à jour consistent à actualiser de manière continue les informations sur ces individus, notamment en termes de naissances, de décès, de mariages, de divorces d'une part, et d'autre part, de tous les changements de résidence ou les migrations (Poulain et Hern, 2013). De nos jours, le traitement informatisé de données a permis de centraliser les registres de population dans certains pays (Poulain et Hern, 2013) et de créer divers registres au sein d'un même pays.

En Suisse, on peut citer par exemple le registre [2] central des étrangers (ZAR), le registre de l'asile (AUPER), le relevé structurel, la statistique de la population et des ménages (STATPOP), la Statistique du mouvement naturel de la population (BEVNAT), etc.

13.4.1 Origine des données de registre

Selon l'OSCE (OSCE, 2009), le premier registre de population a été découvert en Chine à l'époque de la dynastie Han au deuxième siècle avant J.-C. En Europe, l'existence du premier registre de population a été identifiée en Suède. Dès 1665, , le processus d'enregistrement de toutes les familles fréquentant les paroisses avait déjà débutéen Suède (Poulain et Hern, 2013). En 1886, un décret a rendu obligatoire pour les pasteurs d'enregistrer toutes les familles fréquentant leurs paroisses ainsi que les couples mariés, les enfants légitimes et illégitimes avec leurs dates et lieux de naissance, leurs dates de baptêmes, les noms de leurs parents, les décès, les inhumations ainsi que toutes les entrées et sorties ayant lieu dans les paroisses (Hofsten et Lundstrom, 1976; Poulain et Hern, 2013). Ces données paroissiales collectées ont été centralisées de manière à être utilisées comme principale source de statistiques démographiques et de mouvements de la population pour tout le pays (Poulain et Hern, 2013). Malgré la centralisation du registre de la population, ce dernier a longtemps été sous la responsabilité de l'église luthérienne (Hofsten et Lundstrom, 1976). C'est seulement en 1991 que le système d'enregistrement de la population est passé sous la responsabilité de l'administration fiscale (Poulain et Hern, 2013; Hofsten et Lundstrom, 1976).

Cet exemple illustre le fait que les données administratives ne sont pas des données qui sont à la base destinées à la recherche même si elles offrent un grand potentiel en matière d'analyses statistiques. Les données administratives, à l'image des autres types de données présentent des avantages et des inconvénients.

2. Il existe en Suisse plusieurs autres registres qui ne sont pas mentionnés ici.

13.5 Avantages des données administratives

Les données administratives sont une source importante de données qui peuvent être valorisées dans le domaine de la recherche. Ces données sont produites à moindres coût [3] (Coulibaly, 2011; Rowebottom, 1978; Desrosières, 2011), de manière continue et offrent la possibilité de pouvoir suivre la tolalité des individus dans le temps (données longitudinales). Les données administratives sont mises à jour de manière régulière et à tout moment il y a la possibilité d'avoir des informations sur les personnes enregistrées dans les bases de données. Ces données offrent un grand potentiel en matière de production de resultats statistiques dans divers domaines (par exemple, la santé, l'éducation, la sécurité) en baissant considérablement les coûts liés à la production de données par enquêtes et sont généralement d'une bonne qualité. Lorsque les registres sont bien tenus et régulièrement mis à jours, les données admistratives reflètent à tout moment la réalité sur la population concernée et offrent une grande fiabilité (Coulibaly, 2011).

Les données administratives offrent aussi la possibilité d'être couplées (fusionnées) avec d'autres bases de données de sources administratives ou même des données issues d'enquêtes prospectives (Rowebottom, 1978). Les données administratives sont aussi indispensables pour augmenter la fiabilité des données recueillies de manière rétrospective lorsque celles-ci contiennent des incertitudes au niveau de la datation ou de la chronologie de certains événements (Poulain et al., 1991).

13.6 Inconvénients des données administratives

Le premier inconvénient des données administratives est lié au fait que ce ne sont pas des données qui sont à la base destinées à la recherche (Rowebottom, 1978) et peuvent par conséquent présenter quelques limites en termes d'utilisation statistique. Les données administratives, une fois récoltées et livrées, offrent peu de flexibilité par rapport à leurs contenus parce qu'il n'y a aucune possibilité de contacter les personnes concernées pour avoir des précisions contrairement à certaines données d'enquêtes. L'obtention de certaines variables jugées sensibles par les autorités est quasiment impossible à cause de la protection des données, ce qui peut avoir pour conséquence la modification des questions de recherche ou la réalisation de modèles statistiques sans certaines variables jugées importantes. Un autre inconvénient des données administratives réside dans le fait que tout changement de système d'enregistrement des données ou dans le système de récoltes des données peut également modifier la qualité de ces dernières en les sur-estimant ou en les sous-estimant (Jabine et Scheuren, 1985).

Les données administratives pouvant être recueillies de manières diverses au niveau d'un même pays, une comparaison internationale est plus difficile à faire avec ce type de données d'une part, d'autre part, il est possible qu'il y ait des biais lors de la chaine de transmission des données entre les différents services (Coulibaly, 2011). La production des données administratives peut nécessiter la collaboration entre différents services et de ce fait entraîner un délai d'attente très long pour leur production.

Les données qui seront utilisées dans cette recherche sont des données administratives qui proviennent de l'Office Fédéral de la Statistique (OFS). Le choix de l'utilisation de données administratives pour étudier les parcours de vie des étudiants internationaux en Suisse réside dans le fait que les registres suisses sont correctement tenus et mis à jour de manière régulière. Ces données retracent le parcours migratoire et académique des étudiants depuis leur arrivée en Suisse jusqu'à la fin des études ou de la disparition des registres. Ces données enregistrent également les changements survenus dans l'état civil de ces étudiants ainsi que la date d'occurrence de certains événements comme l'obtention d'un diplôme par exemple. Ce sont des données longitudinales quantitatives déjà disponibles dans les registres et dont l'extraction et la mise à disposition des chercheurs sont moins onéreuses en ressources financières et en temps par comparaison aux enquêtes prospectives par exemple. Ces données ainsi que les différentes phases de préparation sont présentées dans les sections suivantes.

3. Les données administratives sont des données disponibles, les seuls coûts qu'ils engendrent sont ceux liés à l'enregistrement et à l'extraction des données. Ces coûts sont inférieurs à ceux engendrés par les enquêtes prospectives.

13.7 Présentation des données de l'OFS

Les données qui sont utilisées dans cette thèse ont été livrées par l'Office fédéral de la Statistique (OFS). Les bases de données proviennent de trois sources différentes qui sont : ZEMIS [4], ZAR [5] et LABB [6]. Dans les bases de données ZEMIS et ZAR, sont enregistrées toutes les informations sociodémographiques, de nationalité, de migration et socioprofessionnelles des étrangers en Suisse, alors que la base de données du projet Analyses longitudinales dans le domaine de la formation, ci-après LABB, contient des informations sociodémographiques des étudiants suisses, étrangers et internationaux mais aussi leur parcour académique depuis le début d'une formation de niveau diplôme jusqu'au doctorat. La base de données LABB est le résultat d'un appariement et d'une harmonisation entre de nombreuses sources (SIUS [7] , STATPOP [8] , UPI [9] , SE [10]). Ces trois sources de données peuvent être regroupées en deux groupes qui sont : une base de données qui enregistre les étrangers et une base de données qui enregistre les étudiants.

13.8 Registres contenant les données sur les étrangers : ZEMIS et ZAR

Dans les bases de données ZEMIS et ZAR sont enregistrés tous les étrangers vivant en Suisse quel que soit le motif de leur séjour (études, travail, regroupement familial ou autre). Les bases de données de ZEMIS livrées par l'OFS sont des bases de données annuelles qui sont référencées au 31 décembre de chaque année. Ces bases de données contiennent des variables que l'on peut regrouper en : variables administratives, variables démographiques, variables de nationalité, variables socioprofessionnelles et en variables de mouvement. Seules les variables jugées importantes seront présentées tout en mentionnant le fait que plusieurs variables importantes comme la date du mariage, la naissance des enfants, le statut de l'emploi, le poste occupé n'ont pas été livrées. La variable administrative la plus importante est la clé unique (l'identifiant) qui permet d'identifier les individus de manière claire ; cette variables est indispensable pour la future fusion des bases de données. Le tableau 13.1 ci-dessous contient quelques variables de ZAR et ZEMIS.

13.9 Description des bases de données ZEMIS de 2010 à 2015

Comme le montre le Tableau 13.2, la base de données ZEMIS 2010 contenait 2'047'460 individus identifiés de manière unique et 38 doublons [11] pour un total de 2'047'498 lignes. La base de données ZEMIS 2013 (ZEMIS 13) contenait 2'277'823 individus identifiés de manière unique, 24 doublons pour un total de 2'277'847 lignes. Parmi ces individus, il y a des étudiants et des non étudiants car tous les étrangers vivant légalement en Suisse indépendamment de leur profession , sont enregistrés dans ZEMIS. Avant de procéder à la fusion des différentes bases de données et aux analyses, il sera d'abord nécessaire d'identifier les étudiants internatioanux dans ces différentes bases de données ZEMIS pour la période de 2010 à 2014. Ce tableau résume également toutes les bases de données ZEMIS de 2010 à 2015. Ces bases de données sont présentées sous la forme d'un individu par ligne.

4. Zentrale Migrationsinformationssystem
5. Zentralen Ausländerregisters
6. Analyses longitudinales dans le domaine de la formation. www.labb.bfs.admin.ch
7. Système d'information universitaire suisse
8. Statistique de la population et des ménages :« *donne la référence démographique pour les élèves répertoriés, telles que le statut migratoire (lieu de naissance et nationalité)* ». Source, OFS
9. Unique person identification : « *référence démographique utilisée pour des élèves ou étudiants qui ne seraient pas dans STATPOP* ». Un étudiant inscrit dans une école suisse mais résidant dans un pays voisin n'est pas enregistré dans STATPOP. Source, OFS
10. Relevé structurel : « *informations de contexte pour le ménage (niveau de formation des personnes composant le ménage)* » : Source : OFS
11. Ces doublons ont été identifiés avec la clé unique à l'aide du logiciel SPSS. On obtient ainsi les observations principales qui représentent le nombre d'individus dans la base de données et les observations duppliquées qui repésentent les doublons.

Tableau 13.1 – Description de quelques variables des données de l'OFS

Variables	Description
Identifiant	Numéro AVS anonymisé = clé unique
Sexe	1=Masculin, 2= Féminin
Etat civil	1=célibataire, 2= marié, 3= veuf, 4=divorcé, 5=partenariat
Date de naissance	Format jj.mm.aaaa
Changement d'état civil	Date de changement d'état civil
Age	Age de la personne en années
Nationalité	Modalités :8100 = Suisse, 8315 = Guinée,$\cdots$
Types de permis	Modalités : Livret A = travailleur saisonnier Livret B = autorisation de séjour/séjour annuel Livret C = autorisation d'établissement Livret L = autorisation de courte durée Livret F = pour étrangers admis provisoirement Livret N = pour requérants d'asile Livret S = pour personnes à protéger Livret Ci= autorisation de séjour avec activité lucrative Livret G = travailleur frontalier
Pays de naissance	Modalités : 1=Suisse, 0 sinon
Nationalité du conjoint/partenaire	Modalités : 1=Suisse, 0 sinon
Le statut dans l'emploi	Poste occupé
Les catégories socioprofessionnelles	Nomenclature générale des activités économiques (NOGA)
L'activité et la profession	Domaine d'activité
Date d'entrée en Suisse	Format jj.mm.aaaa
La durée du séjour	Ecart entre date d'entrée et date d'extraction des données
Le canton de résidence	Modalités : 1=Zurich,...,26 =Jura

Tableau 13.2 – Données ZEMIS de 2010 à 2015

Données	ZEMIS 10	ZEMIS 11	ZEMIS 12	ZEMIS 13	ZEMIS 14	ZEMIS 15
Effectifs	2'047'460	2'124'866	2'194'714	2'277'823	2'344'560	47'097
Doublons	38	35	33	24	5	0
Part doublons	0.00185%	0.00164%	0.00150%	0.00105%	0.000213%	0
Total	2'047'498	2'124'901	2'194'747	2'277'847	2'344'565	47'097

La base de données ZEMIS 2015 ne contenait que des étudiants internationaux pour un effectif total de 47'097 individus. Cette base de données est présentée sous la forme d'un individu par ligne et elle ne contenait pas de doublons. En effet, ZEMIS 2015 ne contient que des étudiants internationaux parce qu'un filtre a été appliqué par l'OFS pour extraire ces étudiants. Cette base de données étant organisée sous la forme d'une personne par ligne, le nombre total de lignes représente le nombre d'étudiants dans la base de données.

Comme le montre le tableau 13.2, les bases de données ZEMIS de 2010 à 2014 étaient livrées dans un format différent et contenaient des doublons, ainsi que des étrangers qui ne sont pas en Suisse pour des motifs d'études. Il était donc nécessaire d'extraire les étudiants internationaux dans les bases de données de ZEMIS couvrant la période de 2010 à 2014 afin de les ramener dans la même structure que ZEMIS 2015. Les bases de données ZEMIS étant des bases de données annuelles, l'extraction a été faite sur chacune de ces bases de données en cherchant à chaque fois les indentifiants de ZEMIS qui sont présents dans LABB. La part des étudiants internationaux dans chaque base de données ZEMIS de 2010 à 2015 est présentée dans le Tableau 13.3.

Tableau 13.3 – Le nombre d'étudiants internationaux dans chaque base de données ZEMIS de 2010 à 2014

Données	ZEMIS 2010	ZEMIS 2011	ZEMIS 2012	ZEMIS 2013	ZEMIS 2014
Etudiants	41'875	48'356	53'244	56'836	57'878
Pas étudiants	2'005'585	2'076'510	2'141'470	2'220'987	2'286'682
Total	2'047'460	2'124'866	2'194'714	2'277'823	2'344'560

On remarque par exemple dans ce tableau pour ZEMIS 2010 que sur un total de 2'047'460 étrangers, seulement 41'875 individus sont en Suisse pour des motifs d'études. Ce tableau est très important car il permet de distinguer les étudiants des autres étrangers qui séjournent en Suisse pour d'autres motifs que les études. L'intérêt de

cette étude portant sur les étudiants, on prendra soin de supprimer de ces bases de données tous les étrangers qui séjournent pour d'autres motifs que les études avant de faire la fusion à venir des différentes bases de données.

13.10 La base de données ZAR de 1997 à 2009

Cette base de données ne contient que des étudiants internationaux qui sont entrés en Suisse entre 1997 et 2009 dans le but de poursuivre des études supérieures. A la base, ZAR est une base de données dans laquelle sont enregistrées toutes les personnes de nationalités étrangères qui vivent légalement en Suisse. A l'image de ZEMIS 2015, un filtre a été appliqué à cette base de données par l'OFS pour n'extraire que les étudiants internationaux. Les caractéristiques de cette base de données sont présentées dans le tableau 13.4.

Tableau 13.4 – Données ZAR de 1997 à 2009

Données	ZAR
Observations uniques	66'699
Observations dupliquées	359'931
Total	426'630

Cette base de données ne contenait pas de doublons et était présentée sous la forme de personnes-période ; cela veut dire qu'une personne apparaît sur autant de lignes que cette dernière a été présente dans le registre. Par exemple, un étudiant entré en Suisse en 2000 et qui a séjourné en Suisse pendant six ans apparaîtra sur six lignes pour ses six ans de séjour. Comme le montre le tableau ci-dessus, la base de données contenait $66'699$ observations dupliquées sur $359'931$ lignes selon la durée de séjour de chacun pour un total lignes dans la base de données de $426'630$.

13.11 La base de données des étudiants (LABB)

La base de données du projet LABB [12] (Analyses longitudinales dans le domaine de la formation) est le résultat d'un appariement et d'une harmonisation entre de nombreuses sources (SIUS [13] , STATPOP [14] , UPI [15] , SE [16]). Cette base de données contient entre autres, des informations administratives (identifiants), géographiques (canton de résidence, pays de domicile avant les études), démographiques (sexe, nationalité, date de naissance), de migrations (type de permis, statut migratoire) ainsi que le parcours académique des étudiants pour différents niveaux de formation : diplôme, école supérieure, bachelor, master et doctorat.

Dans la base de données LABB , nous avons le parcours académique des étudiants suisses, étrangers et internationaux qui suivent une formation dans une haute école suisse pour la période allant de 1980 à 2014 . Les informations contenues dans cette base de données sont groupées en six niveaux de formation qui sont : baccalauréat, ES (école supérieure), diplôme, bachelor (BA), master(MA) et doctorat. Concernant la maturité (le baccalauréat), nous avons l'année de son obtention et le nom de l'école dans laquelle il a été obtenu. Pour les autres types de formation (ES, diplôme, BA, MA et doctorat), nous avons l'année de début de la formation, le type d'institution dans laquelle l'étudiant est inscrit, la filière d'étude, le diplôme qui a permis le passage d'un niveau de formation à un autre, l'année d'obtention du diplôme ainsi que d'autres variables liées à la formation mais qui ne sont pas déterminantes pour cette recherche. Cette base de données contient aussi des variables

12. www.labb.bfs.admin.ch
13. Système d'information universitaire suisse
14. Statistique de la population et des ménages : « donne la référence démographique pour les élèves répertoriés, telles que le statut migratoire (lieu de naissance et nationalité) ». Source, OFS
15. Unique person identification : « référence démographique utilisée pour des élèves ou étudiants qui ne seraient pas dans STATPOP ». Un étudiant inscrit dans une école suisse mais résidant dans un pays voisin n'est pas enregistré dans STATPOP. Source, OFS
16. Relevé structurel : « informations de contexte pour le ménage (niveau de formation des personnes composant le ménage) » : Source, OFS

administratives comme l'identifiant (clé unique) des étudiants. L'identifiant de LABB, appelé aussi « vn », a été construit à l'image de ZAR et ZEMIS sur la base du numéro AVS [17] à treize chiffres anonymisé. Cette clé unique est identique à celles présentes dans ZAR et ZEMIS et est indispensable pour la future fusion des bases de données.

La base de données LABB était présentée sous la forme d'un individu par ligne pour un total de $1'011'201$ étudiants. Les caractéristiques de cette base de données sont présentées dans le Tableau 13.5.

Tableau 13.5 – Données LABB de 1980 à 2014

Données	LABB
Observations uniques	1'011'201
Doublons	408
Part des doublons	0.04%
Total	1'011609

Un descriptif de quelques variables de LABB est présenté dans le tableau 13.6.

Tableau 13.6 – Description de quelques variables de LABB

Variables	Description
Identifiant	clé unique (Permet la fusion)
Début formation	L'année de début de la formation
Fin formation	L'année d'obtention du diplôme
Filière d'étude	Modalités : 1= SHS, 2=Economie, 3=Droit,...
Haute école fréquentée	Modalités : 105=unil, 107=unine,...
Certificat	Modalités : 8=certificat étranger, 5=matu. gymnasiale,...
Type d'institution	Modalités :1=HEU, 2=HES, 3=HEP

Ce tableau résume les variables jugées importantes pour cette recherche. Dans ce tableau, est désigné par formation, le niveau de formation de type école supérieure, diplôme, bachelor, master et doctorat. Pour chacune de ces formations, nous disposons des informations mentionnées dans le tableau.

Par exemple, pour le niveau de formation bachelor, les données sont présentées comme dans le tableau 13.7. Nous disposons des mêmes informations pour les autres niveaux de formation.

Tableau 13.7 – Description de quelques variables de LABB pour la formation de bachelor

Variables	Description
Identifiant	clé unique (Permet la fusion)
Début bachelor	L'année de début du bachelor
Fin du bachelor	L'année d'obtention du bachelor
Filière d'étude	Modalités : 1= SHS, 2=Economie, 3=Droit,...
Haute école fréquentée	Modalités : 105=unil, 107=unine,...
Certificat	Modalités : 8=certificat étranger, 5=matu. gymnasiale,...
Type d'institution	Modalités :1=HEU, 2=HES, 3=HEP

17. Le numéro AVS (Assurance Vieillesse et Survivants) est propre à chaque individu. L'AVS est régie par la loi fédérale du 20 décembre 1946

13.12 Préparation des données de l'OFS

Les données reçues de la part de l'OFS étaient en format « csv » et « sas ». Pour la préparation et l'exploration des données, le choix s'est porté sur le logiciel SPSS. La première étape a été d'importer les données ZEMIS en format « csv » dans SPSS dans le but de vérifier s'il y avait les étiquettes de valeurs, les libellés des variables ainsi que la définition des valeurs manquantes. Ces informations n'étaient pas fournies dans ces bases de données, mais étaient disponibles dans d'autres fichiers « Excel » livrées par l'OFS avant la mise à disposition des bases de données. Il fallait donc compléter les bases de données de ZEMIS en trouvant un moyen pour importer les libellés des variables, les étiquettes de valeurs ainsi que les étiquettes de valeurs manquantes dans SPSS. Une fonction créée dans Excel a permis de faire ce travail en adaptant le contenu des fichiers Excel à la syntaxe SPSS. Ce travail fut long compte tenu de l'immensité des bases de données et du nombre de modalités très élevé de la plupart des variables, mais le résultat obtenu répondait autant aux attentes qu'aux exigences de SPSS du point de vue « syntaxe ». La base de données LABB au format « sas » a été transformée en format SPSS (sav) à l'aide du logiciel StatTransfer. Le fichier en format SPSS ainsi obtenu contenait les libellés des variables mais pas les étiquettes de valeurs. Il a ainsi fallu refaire la même procédure que celle qui a permis de compléter les bases de données ZEMIS. La première livraison des données était constituée de cinq bases de données provenant de ZEMIS (ZEMIS de 2010 à 2014) et la base de données LABB. La deuxième livraison de données fut ZEMIS pour l'année 2015 et la dernière livraison a été ZAR pour les années allant de 1997 à 2009.

Pour une question de complémentarité des bases de données, il était donc indispensable de les fusionner dans le but d'obtenir dans une seule base de données contenant toutes les informations nécessaires aux analyses. Cette procédure est décrite dans les sections qui suivent.

13.13 Fusion des bases de données de l'OFS

Fusionner des bases de données consiste à mettre ensemble des informations concernant un ou plusieurs individus lorsque ces informations proviennent de deux ou plusieurs sources différentes (Black et Roos, 2005; Brad, 1999). Dans la plupart des cas, la fusion se fait à partir d'une clé unique (identifiant) qui est identique dans toutes les bases de données à fusionner, cette clé unique est en quelque sorte un numéro d'identification personnel.

Dans le cas de ces données, l'identifiant a été construit sur la base du numéro AVS à treize chiffres anonymisé. Lorsqu'il n'existe pas de numéro de sécurité sociale, par exemple le numéro AVS, la fusion des bases de données peut se faire sur la base de certaines caractéristiques individuelles non changeables, comme le sexe, la date de naissance, le pays de naissance (Wanner et Forney, 2007). Dans ce cas, ces caractéristiques individuelles doivent exister dans les différentes bases de données à fusionner.

L'objectif de la fusion est d'obtenir dans une même base de données toutes les informations sociodémographiques, le parcours migratoire et le parcours universitaire des étudiants internationaux. Pour que la fusion de deux ou plusieurs bases de données soit possible, il est indispensable que les individus aient les mêmes identifiants dans les bases de données à fusionner ou des variables identiques dans les deux bases de données. Dans nos bases de données, les individus sont identifiés par une clé unique « vn ». Par exemple, un individu ayant un numéro « vn = 100 » dans ZEMIS aura aussi un numéro « vn = 100 » dans LABB. La fusion permettra donc d'associer dans une seule base de données les caractéristiques individuelles « ZEMIS » de cette personne aux caractéristiques individuelles de la même personne dans « LABB ». L'existence d'identifiants communs étant indispensable pour qu'une fusion soit possible, il a été nécessaire d'explorer les bases de données pour avoir une idée de la correspondance des clés uniques (identifiants) entre les différentes bases de données à disposition.

Le Tableau 13.8 nous donne un aperçu de la correspondance des « vn » de ZEMIS et de LABB, c'est-à-dire le degré de concordance des identifiants des deux bases de données. Cette analyse a été faite après suppression des doublons de toutes les bases de données et après suppression des étudiants suisses de la base de données LABB d'une part, d'autre part, après suppression de ZEMIS de toutes les personnes n'étant pas des étudiants. La suppression des étudiants suisses de LABB tient du fait que ces étudiants ne rentrent pas dans la définition des étudiants internationaux et ne représentent pas la population cible de cette étude. La suppression des personnes

enregistrées dans ZEMIS pour des motifs autres que les études présente l'avantage d'avoir uniquement des bases de données constituées d'étudiants internationaux.

La première colonne « Données » du tableau ci-dessus fait référence aux données ZEMIS pour la période allant de 2010 à 2014. La deuxième colonne nous donne pour chaque année le nombre de concordances des identifiants ZEMIS, ZAR et LABB. Ce tableau montre que 41'875 indentifiants de ZEMIS 2010 ont été indentifiés dans LABB. Pour ZEMIS 2012, on observe dans le tableau que 53'244 identifiants de cette base de données ont été retrouvés dans LABB. Les concordances les plus élevées étant observées entre ZEMIS 2014 et LABB soit 57'878 identifiants communs.

Cette analyse a permis de mettre en évidence l'existence d'identifiants communs entre les différentes bases de données et de ressortir la possibilité de les fusionner. Avant de décrire la procédure de fusion de ces différentes bases de données, il est d'abord nécessaire de parler des doublons présents dans certaines d'entre elles.

Tableau 13.8 – Correspondance des « Identifiants » de LABB, ZAR et de ZEMIS

Données	Identifiants LABB
Identifiants ZEMIS 2010	41'875
Identifiants ZEMIS 2011	48'356
Identifiants ZEMIS 2012	53'244
Identifiants ZEMIS 2013	56'836
Identifiants ZEMIS 2014	57'878
Identifiants ZAR	426'630

13.13.1 Traitement des doublons

On parle de doublons lorsqu'il y a un ou plusieurs individus qui apparaissent plus d'une fois dans une base de données. Ces doublons peuvent apparaître dans des bases de données en raison d'une erreur humaine lors de la saisie. Ils peuvent aussi résulter d'autres facteurs comme par exemple, le fait qu'un étudiant qui achève une formation en Suisse, puis sort du pays, pour revenir une ou quelques années plus tard dans le but de suivre une autre formation. Un étudiant ayant un parcours de ce type aura deux dates d'entrées en Suisse et apparaîtra deux fois dans la base de données ; dans ce cas, c'est la date de la dernière entrée qui sera prise en compte. En effet, un étudiant qui fait son bachelor en Suisse et qui quitte la Suisse pour faire un master à l'étranger puis qui revient en Suisse pour faire un doctorat aura déjà son parcours de bachelor enregistré dans LABB et son parcours migratoire enregistré dans ZEMIS. La deuxième entrée contient donc les traces de la première ; ceci justifie la prise en compte de la dernière entrée en Suisse en cas de dates d'entrées multiples. Pour cet étudiant, le lieu d'obtention du bachelor sera en Suisse et le lieu d'obtention du master sera à l'étranger ; des informations comme l'année de début du bachelor, l'année de son obtention, le nom de la haute école qui a délivré ce bachelor, la filière d'étude, le sexe, la date naissance, la nationalité seront disponibles dans LABB.

Faire des analyses statistiques sans traiter les doublons peut fausser les paramètres estimés et entraîner des résultats erronés. Les doublons ont été détectés sur la base de l'identifiant des individus. Une fois ces doublons identifiés, on doit investiguer pour s'assurer que ce sont bien des doublons en effectuant des contrôles à l'aide de certaines variables comme la date de naissance, le sexe, la nationalité à l'arrivée en Suisse ou la date d'entrée en Suisse. Si l'identifiant se repète sur plusieurs lignes et que les informations sur les variables citées précédemment sont identiques sur toutes les lignes concernées, alors il s'agit de vrais doublons. Si les identifiants sont identiques alors que les caractéristiques individuelles sont différentes, il peut s'agir de faux doublons. Dans le cas de faux doublons, une solution serait de changer d'identifiants pour les individus concernés en créant de nouveaux identifiants bien distincts des autres. Dans nos bases de données, il n'y avait pas de cas de faux doublons. Un extrait de doublons de ZEMIS 2010 est présenté dans le Tableau 13.9 pour illustrer quelques cas de vrais doublons.

Tableau 13.9 – Exemple de doublons de ZEMIS 2010

Identifiant	Nationalité	Date de naissance	Sexe	Pays de naissance
252104	France	19.07.1983	Féminin	Etranger
252104	France	19.07.1983	Féminin	Etranger
289875	Allemagne	18.06.1966	Féminin	Etranger
289875	Allemagne	18.06.1966	Féminin	Etranger
451688	Grèce	30.10.2010	Masculin	Etranger
451688	Grèce	30.10.2010	Masculin	Etranger

Dans ce tableau, nous avons trois étudiants internationaux dont le premier est de nationalité française, le deuxième de nationalité allemande et le troisième de nationalité grecque. Les étudiants français et allemand sont de sexe féminin et l'étudiant grec est de sexe masculin. On remarque que chacun de ces étudiants apparaît deux fois dans la base de données avec les mêmes nationalités, les mêmes dates de naissance, le même sexe et le même pays de naissance. Pour ces trois étudiants, nous avons des cas de vrais doublons car toutes les caractéristiques individuelles ainsi que les identifiants se répètent identiquement pour chaque individu. Le Tableau 13.10 indique le nombre de doublons dans chacune des bases de données ainsi que le nombre de lignes touchées par ces doublons.

Tableau 13.10 – Doublons de ZEMIS de 2010 à 2014 et de LABB

Données	ZEMIS 2010	ZEMIS 2011	ZEMIS 2012	ZEMIS 2013	ZEMIS 2014	LABB
Doublons	38	35	33	24	5	408
Lignes touchées	76	70	66	48	10	624

Nous avons pris contact avec l'OFS au sujet des ces doublons. Pour LABB, une liste des doublons a été livrée par l'OFS avec une recommandation de les supprimer. Les doublons de ZEMIS ont aussi été supprimés de toutes les bases de données concernées parce qu'ils représentaient une part très faible des données. Une extraction de ces doublons a été effectuée dans le but d'avoir une idée sur l'origine des étudiants concernés par ces doublons, cette analyse est présentée dans la section ci-dessous.

13.13.2 Probabilité d'obtenir des doublons selon la nationalité

Pour illustrer les caractéristiques des doublons, nous avons procédé à une analyse descriptive de la variable nationalité de ZEMIS 2010 dans le but d'avoir une idée de la répartition des doublons selon la nationalité. Cette analyse, présentée dans le Tableau 13.11, montre que la probabilité d'obtenir des doublons est liée à la nationalité des étudiants.

Il ressort de cette analyse que la probabilité qu'un doublon de ZEMIS 2010 soit attribuée à un étudiant allemand est de 22.4%. Cette probabilité est de 14.5% pour la France et de 15.8% pour le Portugal. En ajoutant aux effectifs de ces trois pays les effectifs de la Grèce, de l'Italie, de l'Autriche et du Portugal, on obtient 68.4% de l'effectif total des doublons de ZEMIS 2010. Cela montre que la probabilité d'obtenir des boublons est plus élevée pour les pays européens que pour les pays extra-européens. Une autre remarque qui ressort de cette analyse réside dans le fait que des pays limitrophes de la Suisse (France, Allemagne, Autriche et Italie) représentent la moitié des doublons de ZEMIS 2010 soit 50.1%. Pour les autres bases de données de ZEMIS, la tendance est la même, les pays qui sont les plus touchés par ces doublons sont l'Allemagne, la France et le Portugal mais dans des proportions moindres que ZEMIS 2010. La Serbie occupe la quatrième place des pays ayant un nombre élevé de doublons dans ZEMIS 2010 ; la probabilité qu'un doublon de ZEMIS 2010 soit attribuée à un étudiant Serbe est de 10.5%. Cette probabilité est élevée pour un pays non membre de l'UE. Cela pourrait probablement s'expliquer par l'existence, en Suisse, d'une forte communauté d'origine serbe. Les pays géographiquement situés en Europe constituent à eux seuls 88.2% de l'effectif total des doublons de ZEMIS

Tableau 13.11 – Probabilité d'obtenir des doublons dans ZEMIS 2010 selon la nationalité

Pays	Fréquence	Pourcentage	Pourcentage cumulé
Allemagne	17	22.4%	22.4%
France	11	14.5%	36.8%
Grèce	2	2.6%	39.5%
Italie	6	7.94%	47.4%
Autriche	4	5.3%	52.6%
Portugal	12	15.8%	68.4%
Turkie	2	2.6%	71.1%
Serbie	8	10.5%	81.6%
Croatie	2	2.6%	84.2%
Kosovo	1	1.3%	85.5%
Biélorussie	2	2.6%	88.2%
Algerie	1	1.3%	89.5%
Cameroun	1	1.3%	90.8%
Chine	2	2.6%	93.4%
Liban	1	1.3%	94.7%
Malaysie	2	2.6%	97.4%
Thailande	2	2.6%	100%
Total	76	100%	

2010 ; pour les autres régions géographiques, les doublons sont soit très faibles, voire inexistants. Une analyse similaire a été réalisée pour les doublons de LABB avant la suppression des étudiants suisses de cette base de données. Le Tableau 13.12 présente cette analyse et comme on peut l'observer, la probabilité d'obtenir des doublons est aussi liée à la nationalité.

Tableau 13.12 – Répartition des doublons de LABB selon la nationalité

Pays	Fréquence	Pourcentage	Pourcentage cumulé
Suisse	245	60.05%	60.05%
Allemagne	22	5.40%	65.45%
Italie	46	11.27%	76.72%
France	12	2.94%	79.66%
Liechtenstein	7	1.72%	81.38%
Autres (Europe)	44	10.78%	92.16%
Reste du monde	32	7.84%	100%
Total	408	100%	

Dans la base de données LABB, sont enregistrés tous les étudiants qui suivent une formation supérieure en Suisse qu'ils soient de nationalité suisse ou pas. Il ressort de l'analyse des doublons de LABB que les étudiants suisses sont les plus touchés par les doublons. Avec un effectif de 245 doublons, la probabilité qu'un doublon de LABB soit attribué à un étudiant suisse est de 60%. L'Italie arrive en deuxième position avec 46 doublons, l'Allemagne en troisième position avec 22 doublons puis la France avec 12 doublons et le Liechtenstein avec 7 doublons pour des probabilités respectives de 11.3%, 5.4%, 2.94% et 1.72%. On remarque aussi que les pays limitrophes de la Suisse sont les pays les plus concernés par ces doublons par rapport aux autres pays. En effet, la probabilité qu'un doublon proviennent d'un pays limitrophe de la Suisse est de 21.57% (la part de l'Autriche avec un doublon est comprise dans cette probabilité) ; cette probabilité est de 10.78% pour les autres pays européens et 7.84% pour le reste du monde. Pour cette base de données, les pays extra européens les plus touchés par ces doublons sont le Brésil avec 8 doublons soit une probabilité de 1.96%, suivi du Canada avec 4 doublons soit une probabilité de 0.98%, les autres pays n'ayant que peu ou pas de doublons.

Nous remarquons donc que les doublons sont plus nombreux au sein des étudiants suisses, ensuite suivent les étudiants issus des pays limitrophes de la Suisse puis les étudiants européens et enfin le reste du monde.

13.13.3 Logiciel utilisé pour la fusion des bases de données

Le logiciel Stata a été utilisé pour faire la fusion dans la mesure où Stata offre plus de clarté en ce qui concerne les résultats obtenus après la fusion. Ce logiciel mentionne clairement le nombre d'identifiants fusionnés d'une part, d'autre part ; il nous donne aussi le nombre d'identifiants n'ayant pas trouvé de correspondances dans chaque base de données.

13.13.4 Procédure de fusion

Nous disposions de trois sources de données à fusionner qui sont : ZAR, ZEMIS et LABB. La base de données ZAR couvre la période de 1997 à 2009 et ne contient que des étudiants internationaux. La base de données ZEMIS couvre la période allant de 2010 à 2015. La base de données ZEMIS 2015 ne contenait que des étudiants internationaux alors que les bases de données ZEMIS de 2010 à 2014 contenaient des étudiants et des étrangers qui ne sont pas des étudiants. Des bases de données de ZEMIS de 2010 à 2014, ont été extraits les étudiants internationaux de manière à ramener ces bases de données dans le même format que ZAR et ZEMIS 2015. La base de données LABB contient le parcours académique des étudiants suisses, internationaux et étrangers qui suivent une formation dans une haute école en Suisse. Nous disposons à présent de huit [18] bases de données à fusionner dont sept enregistrent le parcours migratoire des étudiants internationaux et une qui enregistre le parcours académique. La procédure de fusion s'est déroulée selon les étapes suivantes :

1. **Première étape** : On ne peut fusionner des bases de données que si ces dernières ont des clés uniques (identifiants) qui identifient de manière unique chaque individu dans les différentes bases de données à fusionner (Andress et al., 2013). Cette clé unique a été construite sur la base du numéro AVS à treize chiffres. Une analyse exploratoire a permis de mettre en évidence l'existence d'identifiants communs entre les différentes bases de données (ZAR et ZEMIS) et la base de données LABB ; cette analyse a été présentée dans le tableau 13.8. Cette analyse nous rassure sur la possibilité de fusionner les différentes bases de données.
2. **Deuxième étape** : La deuxième étape de la fusion a été d'ouvrir la base de données LABB avec le logiciel Stata et de supprimer tous les étudiants de nationalité suisse d'une part, d'autre part de supprimer également tous les doublons. Cette base de données est présentée sous la forme d'un étudiant par ligne.
3. **Troisième étape** : La troixième étape a consisté à supprimer [19] tous les doublons des bases de données de ZEMIS de 2010 à 2014 et de mettre ensemble tous les fichiers (ZAR, ZEMIS) ne contenant que les étudiants à l'aide de la fonction « append » de Stata. Lorsque l'on utilise la fonction « append » de Stata, on ajoute des lignes à la base de donnés indépendamment du nombre d'observations des bases de données. Toutefois, ces dernières doivent contenir les mêmes variables (Andress et al., 2013). Cette fonction permet donc de superposer les bases de données les unes sur les autres. On obtient ainsi une base de données dans laquelle un étudiant peut apparaître sur plusieurs lignes selon la durée de son séjour en Suisse. Par exemple, un étudiant international ayant séjourné six ans en Suisse apparaîtra sur six lignes différentes pour ses six ans de séjour. La structure générale de cette base de données est un étudiant sur plusieurs lignes selon la durée de séjour de l'étudiant.
4. **Quatrième étape** : La quatrième étape a consisté à ouvrir la base de données LABB préparée dans la deuxième étape avec Stata et de la fusionner avec celles de ZAR et ZEMIS préparées dans la troisième étape.

Cette fusion se fera à partir de la clé unique (l'identifiant) qui est la même dans les bases de données à fusionner. Selon l'organisation des bases de données, plusieurs cas de figures peuvent se présenter lorsqu'on fusionne des bases de données ; on peut rencontrer les cas suivants :

18. ZAR, LABB et ZEMIS de 2010 à 2015 (une base de données par année)
19. La suppression des doublons résulte de la part très faible qu'ils représentent (voir tableau 13.2).

Fusion de type 1 : 1 (un à un)

Dans ce type de fusion, les bases de données à fusionner ont une ligne par personne dans les deux bases de données, c'est-à-dire qu'un individu apparaît sur une ligne dans les deux bases de données. L'identifiant de cet individu apparaîtra donc une fois dans chacune des deux bases de données.

Fusion de type 1 : m (un à plusieurs)

Dans ce type de fusion, dans la première base de données, il y a un individu par ligne et dans la deuxième, ce même individu peut apparaître sur plusieurs lignes. L'identifiant de cet individu apparaîtra donc une fois dans la première base de données et plusieurs fois dans la seconde.

Fusion de type m : 1 (plusieurs à un)

Ce type de fusion est l'inverse de la fusion de type 1 : m. Dans la première base de données un individu peut apparaître plusieurs fois sur plusieurs lignes alors que dans la seconde c'est un individu par ligne. De même l'identifiant se répétera plusieurs fois dans la première base de données, mais n'apparaîtra qu'une seule fois dans la seconde.

Fusion de type m : m (plusieurs à plusieurs)

Dans la fusion de type plusieurs à plusieurs, les individus apparaissent sur plusieurs lignes dans les deux bases de données, de même que pour les identifiants.

En faisant le lien avec les données de l'OFS, la fusion de type un à plusieurs a été privilégiée. En effet, dans la base de données LABB, nous avons un individu par ligne, c'est-à-dire qu'il y a autant de lignes que d'individus. Les bases de données ZAR et ZEMIS ont été mises ensemble avec la fonction « append » de Stata, cette fonction superpose les bases de données les unes sur les autres, ce qui fait que nous aurons plusieurs lignes pour un individu. La base de données LABB (un individu par ligne) a été ouverte la première dans Stata, ensuite cette base de données fut fusionnée avec ZAR et ZEMIS mis ensemble (plusieurs lignes pour un individu) d'où une fusion de type 1 : m dont le résultat est présenté dans la section suivante.

13.14 Résultat de la fusion

Comme on peut le voir dans le Tableau 13.13, le résultat de la fusion a permis de montrer que $588'071$ identifiants de LABB ont été mis en communs dans ZEMIS et ZAR ce qui représente 65.34% d'identifiants communs entre les deux bases de données.

Par ailleurs, on remarque aussi dans ce tableau que $168'065$ identifiants de ZAR et ZEMIS et $143'845$ identifiants de LABB n'ont pas trouvé de correspondances dans les deux bases de données. C'est au total, $311'910$ identifiants $(168'065 + 143'845)$ des deux bases de données qui n'ont pas été fusionnés. En effet, en faisant la somme des identifiants fusionnés et des identifiants non fusionnés, on obtient le nombre total de lignes dans la base de données $(311'910 + 588'071 = 899'981)$.

Tableau 13.13 – Résultat de la fusion de LABB, ZEMIS et ZAR

Fusion	Fréquences	Pourcentage	Pourcentage cumulé
ZAR et ZEMIS non fusionnés	168'065	18.67%	18.67%
LABB non fusionnés	143'845	15.98%	34.66%
ZAR, ZEMIS et LABB fusionnés	588'071	65.34%	100%
Total	899'981	100%	

Une fois ce résultat obtenu, la prochaine étape a consisté à extraire seulement la base de données contenant les identifiants fusionnés parce que c'est sur cette dernière que porteront les futures analyses ainsi que l'extraction de la base de données des étudiants africains.

La syntaxe Stata « keep if $\_merge == 3$ » a permis d'extraire la base de données fusionnée avec les identifiants mis en communs, cette base de données contient les caractéristiques migratoires et académiques des étudiants internationaux et étrangers ainsi que leurs caractéristiques démographiques. La sélection des étudiants africains a ensuite été faite en utilisant cette base de données fusionnée, en se basant sur le critère de la nationalité et du certificat (baccalauréat) qui a permis à ces étudiants d'accéder à l'enseignement supérieur suisse.

13.15 Extraction de l'échantillon des étudiants africains

Comme déjà mentionné, l'extraction de la base de données ne contenant que les étudiants africains a été basée sur la nationalité et sur le certificat (baccalauréat) qui a permis à ces étudiants d'accéder à l'enseignement supérieur suisse. Le critère de la nationalité implique que ces étudiants doivent avoir une nationalité d'un pays africain et le critère du certificat implique que le certificat (baccalauréat) qui a permis à ces étudiants africains d'accéder à l'enseignement supérieur suisse ne doit pas être délivré [20] par une école suisse (ils ne doivent pas avoir été scolarisés antérieurement en Suisse). La variable nationalité des étudiants a été recodée en variable nationalité par contient qui prend les modalités : 1 = Europe, 2 = Afrique, 3 = Asie, 4 = Amérique et 5 = Océanie, dans le but de regrouper toutes les modalités de cette variable correspondant à une nationalité africaine en une seule modalité regroupant toutes les nationalités africaines. La variable « admissionqualification1 » renferme tous les types de certificats qui ont permis à ces étudiants d'accéder à l'enseignement supérieur suisse. La modalité « 8 » de la variable « admissionqualification1 » correspond à tous les étudiants qui ont eu accès à l'enseignement supérieur suisse à partir d'un certificat étranger (baccalauréat étranger). Ceci étant, il est dès lors possible d'extraire une base de données ne contenant que les étudiants africains sur la base de ces critères.

L'échantillon de la base de données ne contenant que les étudiants africains a été obtenu en sélectionnant tous les étudiants qui satisfont les conditions « nationalité par contient » = 2 (une nationalité africaine) et « admissionqualification1 » = 8 (un certificat étranger). Nous obtenons ainsi une base de données ne contenant que les étudiants africains, mais, avec la possibilité qu'une personne apparaisse plusieurs fois dans cette base de données. Par exemple, un étudiant qui est arrivé en Suisse en 2010 et qui a terminé son bachelor en 2013 apparaîtra trois fois (sur trois lignes) dans cette base de données pour ses trois années d'études de bachelor. Une fois cette base de données obtenue il sera nécessaire de l'explorer avant les analyses et de recoder des variables dans le but d'avoir une idée sur la qualité des données.

13.16 Qualité des données

Les données à disposition sont des données longitudinales quantitatives de source administrative. De par l'aspect longitudinal des données, leur qualité peut se mesurer par la capacité qu'elles offrent de pouvoir raconter les parcours migratoires et académiques des personnes qui constituent la base de données avec une datation précise des différents événements. Pour tester la qualité des données, nous avons pensé à présenter quelques parcours extraits des bases de données ZAR et ZEMIS.

Dans le tableau 13.14, on remarque qu'il y a des informations qui n'apparaissent pas. En effet, le nom de famille, le prénom, le numéro AVS, la date du mariage, les dates de naissance des enfants sont des informations confidentielles et de ce fait n'apparaissent pas dans les données livrées par l'OFS en raison de la protection des données. Les colonnes « Id » et « Pays » du tableau 13.14 sont également vides pour la même raison, même si ces informations sont disponibles dans les données livrées pas l'OFS.

20. Voir définition des étudiants internationaux dans le chapitre 1 à la page 19

Tableau 13.14 – Exemple de parcours complexe obtenu avec les données de l'OFS

Nu.	Id	Bac	Entree	Présence	Sexe	Naissance	Mariage	Né à	Epoux	Pays
1	···	1998	2003		M	1980	célib	Etranger	pas marié	···
2	···	1998	2003	2003	M	1980	célib	Etranger	pas marié	···
3	···	1998	2003	2004	M	1980	célib	Etranger	pas marié	···
4	···	1998	2003	2005	M	1980	célib	Etranger	pas marié	···
5	···	1998	2003	2006	M	1980	célib	Etranger	pas marié	···
6	···	1998	2003	2007	M	1980	célib	Etranger	pas marié	···
7	···	1998	2003	2013	M	1980	marié	Etranger	marié	···
8	···	1998	2003	2014	M	1980	marié	Etranger	marié	···

L'objectif principal du tableau 13.14 est de montrer que les données administraives ne permettent pas de retracer certains parcours complexes malgré une mise à jours régulière de ces dernières. En effet, dans le tableau 13.14, chaque ligne correspond à une année de séjour en Suisse. En observant le nombre de lignes du tableau, on pourrait croire à un séjour de huit ans pour cet étudiant. Mais, en regardant l'intersection des lignes 6 et 7 de ce tableau avec la colonne présence, on remarque que cette personne n'a pas été enregistrée dans le registre des étrangers entre 2007 et 2013, soit durant six ans. Il est impossible de savoir pourquoi cette personne n'a pas été enregistrée dans ce registre durant cette période sans procéder à des enquêtes rétrospectives, ce qui est impossible avec ces données en raison de la confidentialité. Cependant, la colonne mariage montre qu'au moment de la disparution du registre en 2007, cette personne était célibataire et quand elle a réapparu dans le registre en 2013, elle s'était mariée.

Dans ce deuxième exemple de parcours présenté dans le tableau 13.15, nous avons un étudiant qui est arrivé en Suisse en 2003 et qui y a séjourné jusqu'en 2006. Il est parti de la Suisse en 2006 pour revenir en 2008 ; il séjournera en Suisse jusqu'en 2013. Le séjour de cet étudiant en Suisse a été interronpu pour une année, mais on ne peut pas savoir quelle est la raison de cette intérruption. Dans les bases de données ZAR et ZEMIS, les entrées multiples sont rares chez les étudiants africains mais plus fréquentes chez les étudiants européens.

Tableau 13.15 – Exemple de parcours avec double entrée en Suisse obtenu avec les données de l'OFS

Nu.	Id	Bac	Entree	Présence	Sexe	Naissance	Mariage	Né à	Epoux	Pays
1	···	1998	2003		M	1975	célib	Etranger	pas marié	···
2	···	1998	2003	2003	M	1975	célib	Etranger	pas marié	···
3	···	1998	2003	2004	M	1975	célib	Etranger	pas marié	···
4	···	1998	2003	2005	M	1975	célib	Etranger	pas marié	···
5	···	1998	2003	2006	M	1975	célib	Etranger	pas marié	···
6	···	1998	2008	2009	M	1975	célib	Etranger	pas marié	···
7	···	1998	2008	2010	M	1975	célib	Etranger	pas marié	···
8	···	1998	2008	2011	M	1975	célib	Etranger	pas marié	···
9	···	1998	2008	2012	M	1975	célib	Etranger	pas marié	···
10	···	1998	2008	2013	M	1975	célib	Etranger	pas marié	···

13.17 Discussion sur les méthodes et sur les données

Dans cette deuxième et troisième partie consacrées aux méthodes et aux données, les modèles non paramétriques, paramétriques et semi-paramétrique de Cox ont été présentés ainsi que les différentes sources de données longitudinales, puis les données de l'OFS. Dans les modèles non paramétriques, nous avons pu observer que le choix de la méthode de Kaplan-Meier ou de la méthode actuarielle peut dépendre de la taille de l'échantillon, du rythme d'occurrence de l'événement étudié, de la connaissance ou non de l'instant précis auquel les événements ont lieu ou si les données sont enregistrées dans une unité de temps large en l'occurrence l'année. Il n'y a donc aucune considération technique qui ferait que l'on doit privilégier l'une ou l'autre méthode. Dans la méthode actuarielle, le chercheur peut créer des intervalles de temps dont il a la liberté de fixer la largeur. Cependant, il doit veiller à ce que dans chaque intervalle de temps il y ait eu quelques occurrences de l'événement étudié. Dans la méthode de Kaplan-Meier, les intervalles sont fixés de manière aléatoire parce que la survie

est calculée à chaque fois que l'événement étudié se produit. L'avantage de ces méthodes réside dans le fait qu'aucune hypothèse n'est faite en ce qui concerne la distribution du risque au cours du temps, on se contente seulement de l'ordre d'occurrence des événements étudiés. Ces deux estimateurs sont asymptotiquement sans biais, cela veut dire que plus la taille de l'échantillon augmente plus on obtient des estimateurs non biaisés (Droesbeke et al., 1989). Les estimateurs obtenus par ces deux méthodes convergent vers les mêmes valeurs lorsque nous avons des échantillons de grandes tailles (Estève et al., 1994). Les modèles de Kaplan-Meier et actuariel sont considérés comme des modèles descriptifs. En effet, ces modèles ne permettent pas de mesurer l'impact d'une ou de plusieurs variables explicatives sur le risque d'occurrence d'un événement donné ou sur la durée de séjour. Nous avons alors introduit les modèles paramétriques qui sont des modèles explicatifs.

Nous avons pu observer que les modèles paramétriques, se subdivisent en modèles à hasard proportionnel (HP) et en modèles à temps de sorties accélérées (AFT), puis des exemples d'application ont été développés pour chaque modèle dans le but de les comparer.

A partir de ces exemples, on a pu faire ressortir une limite importante des modèles paramétriques dans la mesure où la manière dont le modèle est spécifié peut avoir des conséquences sur les résultats. En effet, dans la régression exponentielle par exemple, pour passer d'un modèle à HP à un modèle à AFT, il suffit simplement de changer le signe des coefficients estimés. Cela veut dire qu'une variable qui a un effet positif sur le risque d'occurrence d'un événement lorsque le modèle est à HP aura un impact négatif sur le risque d'occurrence de ce même événement lorsque le modèle est à AFT.

Nous avons fait ressortir le fait que pour les modèles à hasard proportionnel, « la variable dépendante » est le risque ; les variables explicatives agissent donc sur cette fonction de risque. La formulation des questions de recherche pour ce groupe de modèles doit tenir compte du fait que le risque dépend de ces caractéristiques individuelles. Il est important de retenir que ce risque n'est pas une caractéristique observée.

Dans les modèles à AFT, la variable dépendante est la durée de séjour ; les variables explicatives agissent donc sur cette dernière.

Le modèle exponentiel est un cas particulier du modèle de Weibull. Ce dernier est généralement considéré comme un modèle flexible et facile à mettre en pratique en raison de l'aspect monotone de sa fonction de risque et au fait qu'il s'adapte à la fois aux modèles à hasard proportionnel et aux modèles à temps de sorties accélérées, ce qui lui confère une grande popularité dans le domaine médical (Habibi et al., 2018). Les modèles log-logistique et log-normale conviennent à la modélisation de phénomènes dont le risque d'occurrence n'évolue pas de manière monotone au cours du temps (Habibi et al., 2018). Les modèles, exponentiel, Weibull et log-normale sont des cas particuliers du modèle gamma généralisé. Pour choisir un modèle adapté à nos données entre ces trois modèles, on va commencer par ajuster un modèle gamma généralisé et on va procéder ensuite à un test d'hypothèse sur les paramètres obtenus avec le modèle gamma dans le but de choisir le meilleur modèle pour analyser les données. Ces différents tests statistiques ont été présenté dans le chapitre consacré au modèle gamma généralisé (chapitre 10). Les modèles paramétriques offrent une meilleure estimation des paramètres par rapport au modèle de Cox lorsque les conditions d'application sont respectées (Kargarian-Marvasti et al., 2017). Lorsque celles-ci ne sont pas respectées, le modèle de Cox est alors une bonne alternative aux modèles paramétriques (Cleves et al., 2010).

Le modèle de Weibull convient à l'analyse d'événements pour lesquels la fonction de risque a une allure monotone, tandis que les modèles log-logistique et log-normale conviennent à l'analyse de phénomènes dont la fonction de risque a une allure unimodale (Khan, 2017). Par allure unimodale, nous entendons une fonction qui croit, atteint un maximum puis décroît. Ces modèles nous permettent de comprendre la relation entre le temps et la durée jusqu'à l'occurrence d'un événement en fonction d'une ou de plusieurs caractéristiques individuelles. Ces dernières pouvant exercer un impact soit sur la fonction de risque soit sur la fonction de séjour. Lorsque les modèles ne sont pas directement comparables, on va utiliser le critère AIC pour choisir le meilleur modèle. Le meilleur modèle selon ce critère est celui qui aura la valeur la plus petite du critère AIC. Lorsqu'un modèle est un cas particulier d'un autre modèle, le choix se fera à partir du modèle qui contient l'autre en procédant à des tests statistiques sur les paramètres estimés.

L'organisation de la base de données pour mettre en pratique les modèles paramétriques est la même que pour les modèles non paramétriques. On va simplement ajouter les variables explicatives d'intérêt à la base de

données préparée pour les modèles non paramétriques.

Les modèles paramétriques et le modèle de Cox s'appuient sur l'idée d'un temps continu. Cela implique que très peu de personnes peuvent expérimenter l'événement étudié au même moment (Le Goff et al., 2013). L'approche en temps continu peut être contraignante selon la nature des données à disposition, surtout lorsque celles-ci présentent un risque élevé que plusieurs personnes expérimentent simultanément l'événement étudié. Ce risque est plus élevé lorsque la période d'observation est large, en l'occurrence l'année. Comme nous l'avons constaté dans la partie consacrée aux données, les données qui seront utilisées dans cette thèse pour répondre à nos questions de recherche sont des données administratives qui sont récoltées au 31 décembre de chaque année. Le caractère annuel des données nous permet de dire que les événements susceptibles de nous intéresser ont lieu à une période fixe de l'année (rentrée universitaire, obtention d'un diplôme, passage en classe supérieure, prolongation d'un titre de séjour). La rentrée universitaire a lieu en septembre même si la possibilité de commencer les cours en février existe. Cette rentrée a donc lieu à des périodes fixes de l'année (septembre ou février).

Dans le cas des étudiants internationaux de manière générale et particulièrement celui des étudiants africains, la rentrée universitaire débute au mois de septembre. Cette rentrée universitaire coïncide généralement avec l'octroi du permis de séjour pour études qui est aussi annuel et prolongeable chaque année. On peut ainsi partir du principe qu'un étudiant international ou africain qui reçoit son permis de séjour en septembre présentera un risque important de quitter la Suisse plus ou moins en septembre de l'année de la fin des études si le statut du séjour ne change pas entretemps. On peut ainsi partir du principe que la probabilité que plusieurs étudiants quittent la Suisse la même année est très élevée, ce qui constitue un problème dans l'application des modèles paramétriques ou le modèle Cox à nos données. Lorsque la période d'observation est large, les modèles paramétriques et de Cox deviennent moins adéquats parce que la probabilité d'obtenir des groupes d'égalité devient grande. Ces modèles sont donc sensibles aux groupes d'égalités et appartiennent aux modèles à hasard proportionnel. Cette sensibilité aux groupes d'égalité et le fait que ces modèles supposent que le temps est continu nous poussent vers d'autres modèles plus adaptés.

Avec nos données qui sont récoltées à une période fixe de l'année, le temps n'est plus continu mais discret ; il nous faut donc un modèle qui convient à l'analyse de ce type de données présentant un risque élevé que plusieurs personnes expérimentent au même moment l'événement étudié. Les modèles de régression logistiques à temps discret dont le modèle logit à temps discret appartient à ce groupe de modèle. Ce modèle qui convient pour analyser des données longitudinales mesurées sur des individus à des périodes fixes de l'année est le modèle le plus adapté à nos données. L'organisation des données en vue de réaliser un modèle logit à temps discret impose que celles-ci soient organisées sous la forme de personne-période.

Pour mettre en application ces méthodes, nous avons besoin de données. Nous avons ainsi commencé par présenter les différentes sources de données longitudinales ainsi que les avantages et les inconvénients de chaque source de données. Ces données sont recoltées de toutes les manières possibles ; rétrospectives, prospectives et administratives. Les enquêtes rétrospectives bien qu'étant moins répandues de nos jours ont précédé les enquêtes prospectives, malgré le fait que ces dernières soient les plus utilisées. Les données administratives, qui sont des données qui ne sont à la base pas destinées à la recherche sont une source importante de données longitudinales à moindres coûts lorsqu'elles sont régulièrement mises à jour. Ces données administratives peuvent être utilisées pour augmenter la fiabilité des données issues d'enquêtes rétrospectives qui, elles, sont très exposées aux biais de mémoire. Les données administratives peuvent aussi être couplées avec des données d'enquêtes et des données issues d'autres sources administratives pour mettre en commun plusieurs informations de sources différentes concernant les mêmes individus. Dans ce cadre, nous avons présenté toutes les possibilités qu'il est possible de rencontrer lorsque nous mettons en commun des bases de données possedant une clé unique.

Nous avons aussi pu observer que les bases de données administratives peuvent parfois contenir des informations qui nécessitent certaines interrogations. En effet, les types de permis que nous avons dans le tableau 13.1 proviennent de la base de données ZAR de 1997 à 2009. Cette base de données ne contient que des étudiants internationaux et étrangers qui sont entrés en Suisse durant cette période dans le but de faire des études universitaires. Les étudiants détenteurs de permis B et C ne posent aucun problème dans le sens où les étudiants africains par exemple qui arrivent en Suisse pour faire des études bénéficient d'un permis B pour études (court

séjour). On peut partir aussi du principe que les étudiants étrangers, qu'ils soient africains ou non, peuvent être détenteurs d'un permis C à cause du fait que ces étudiants peuvent être nés en Suisse ou avoir grandi et fait toute leur scolarité en Suisse. Ce qui peut paraître un peu confus est le fait que des étudiants internationaux étaient parfois détenteurs de permis saisonniers (abolis le 1er juin 2002), de permis frontaliers ou de permis avec activités lucratives. En effet, les permis saisonniers sont octroyés à des personnes qui viennent travailler en Suisse durant une certaine période de l'année (vendanges, récolte de fruits, de légumes,...). Les données montrent que certains étudiants étaient titulaires de ce genre de permis, ce qui peut être incompatible avec le statut d'étudiant. Nous remarquons aussi que certains étudiants internationaux sont détenteurs de permis frontaliers appelés communément permis G. Le permis G est réservé aux travailleurs frontaliers ; ces derniers peuvent séjourner en Suisse durant la semaine de travail et rentrer chez eux toutes les semaines parce que leur résidence principale n'est pas en Suisse. Les étudiants n'ayant droit qu'à quinze heures de travail par semaine durant la période de cours et à plein temps pendant les vacances d'été, il est tout à fait légitime de se poser la question de savoir pourquoi ces étudiants ont ce type de permis. Il est aussi important de préciser le fait qu'un étudiant qui suit une formation en Suisse et qui réside en dehors de la Suisse n'est pas enregistré dans STATPOP.

Nous avons aussi pu observer que la probabilité d'obtenir des doublons est liée à certaines nationalités et à la proximité géographique de la Suisse. Sachant que les étudiants suisses ne sont pas enregistrés dans ZEMIS, nous avons pu observer que les étudiants issus de pays limitrophes de la Suisse sont ceux qui ont les plus grands effectifs de doublons. La probabilité qu'un doublon de ZEMIS 2010 provienne d'un pays limitrophe de la Suisse est de 50%, cette probabilité est respectivement de 38.16% pour les pays européens et de 11.84% pour le reste du monde. Dans la base de données LABB, les étudiants suisses sont ceux qui ont les plus grands effectifs de doublons, puis, viennent les étudiants issus des pays limitrophes de la Suisse ensuite ceux issus de pays européens puis le reste du monde avec de effectifs de doublons de plus en plus faibles voire inexistants. On pourrait ainsi penser à classifier les doublons de LABB par zones en fonction de l'importance de l'effectif des doublons observés. La zone une serait la Suisse avec 60.05% de l'effectif des doublons de LABB, suivie de la zone deux constituée des pays limitrophes de la Suisse avec 21.57% de l'effectif des doublons de LABB ensuite la zone trois avec 10.78% de l'effectif des doublons de LABB et enfin la zone quatre avec 7.84% des doublons.

Les données administratives étant le type de données utilisées dans cette recherche, nous mentionnerons quelques difficultés rencontrées lors de la préparation de ces données. La première difficulté rencontrée a été celle de mettre ensemble les variables, leurs étiquettes de valeur, les libellés des variables ainsi que les étiquettes de valeurs manquantes. Certaines variables avaient un nombre de modalités considérable, il était donc impossible de saisir manuellement toutes ces informations dans SPSS. Par exemple la variable nationalité regroupe toutes les nationalités du monde (autant de modalités que de pays), la variable NOGA regroupe tous les secteurs économiques (autant de modalités que de secteurs économiques). Il a donc fallu créer une feuille de transformation dans Excel pour écrire les libellés des variables, les étiquettes de valeur et les valeurs manquantes en langage SPSS, puis compléter les bases de données avec les informations manquantes dans SPSS. Ce travail fut très long à faire tant les bases de données sont volumineuses et le risque de faire des erreurs élevé. Ces informations étaient manquantes pour toutes les bases de données ZEMIS, ces dernières ont été complétées l'une après l'autre. A l'importation de la base de données LABB, les libellés des variables furent aussi importés mais pas les étiquettes de valeurs et de valeurs manquantes. Il a donc fallu de nouveau faire recours à la feuille Excel de transformation pour importer les informations manquantes dans LABB, ce qui a de nouveau pris beaucoup de temps, parce que le nombre de variables est plus élevé dans LABB que dans ZEMIS. Pour chaque information ajoutée à une variable, il a fallu exécuter la syntaxe pour voir si cette information a été prise en compte et si la syntaxe était correcte. Ces différents travaux ont conduit à la mise à la disposition des futurs chercheurs intéressés par ces données d'un codebook de 78 pages qui fait partie intégrante des apports de cette thèse.

Une autre difficulté liée à la première est l'immensité des bases de données qui a eu pour conséquence de surcharger mon ordinateur de bureau et de ralentir son fonctionnement.

La deuxième difficulté est liée à la protection de données. En effet, les données qui sont utilisées dans cette thèse sont des données réelles. Certaines personnes présentes dans ces bases de données sont encore en Suisse et d'autres probablement pas. Les personnes encore présentes en Suisse ont une adresse, un numéro AVS, une profession et auraient pu être retrouvées par tout utilisateur de ces données si un travail d'anonymisation

des données n'avait pas été fait au préalable. Le but de cette anonymisation des données est de protéger les personnes d'une utilisation inappropriée de celles-ci tout en maintenant la possibilité de les fusionner. Cette protection des données qui est tout à fait légitime et d'une importance capitale n'est pas sans conséquence pour la recherche. En effet, des variables importantes comme la date du mariage (s'il y en a eu un) et la naissance des enfants n'ont pas été livrées au nom de la protection des données. Ceci nous a contraint à renoncer à ces variables indépendantes et revoir nos ambitions à la baisse. Nous avons ainsi renconcé à certaines analyses portant sur la chronologie de certains événements du parcours de vie des étudiants, comme par exemple la chronologie entre l'obtention du diplôme, le mariage, la naissance des enfants ou le changement de statut.

Une chose intéressante qu'on aurait pu faire s'il n'y avait pas eu l'obstacle lié à la protection des données aurait été de confronter deux parcours de vie dont le premier provient des données livrées pas l'OFS et l'autre d'une autobiographie d'un étudiant quelconque figurant dans les bases de données. L'objectif final étant la confrontation des deux récits dans le but de déceler les divergences. Mais en raison de la protection des données, une telle pratique n'est pas possible avec les données administratives. Nous observons donc que si la protection des données est essentielle, celle-ci peut parfois avoir un impact sur les chercheurs à travers la non mise à disposition de certaines variables ou tout simplement en limitant les analyses que l'on peut faire.

Par ailleurs, le premier objectif des données administratives n'est pas de satisfaire les besoins de la recherche, ce qui fait que la plupart des variables importantes pour cette recherche ne sont en soi pas disponibles. Il va falloir procéder à la construction de la plupart des variables dépendantes avant de procéder aux analyses ciblées.

La troisième difficulté est liée à la livraison par vague des données. La livraison par vague des données a eu pour conséquence de refaire plusieurs fois le même travail. En effet, la première livraison des données était constituée de ZEMIS pour les périodes allant de 2010 à 2014 et de LABB. Ces bases de données ont été fusionnées puis des analyses ont été réalisées suite au retard pris pour la livraison du reste des données . La livraison de ZEMIS 2015 a eu pour conséquence de refaire la fusion ainsi que toutes les analyses effectuées auparavant. Avec la livraison de ZAR pour la période allant de 1997 à 2009, il a fallu de nouveau faire la fusion ainsi que les analyses réalisées avec les bases de données qui étaient déjà à notre disposition. Le fait d'avancer avec les analyses et de devoir tout recommencer à chaque livraison de données, combiné avec la durée d'attente très longue pour la livraison de ces dernières ont été les difficultés majeures rencontrées dans cette recherche. Comme le montre le chapitre consacré aux données administratives de l'OFS, ces difficultés ont néanmoins pu être surmontées.

En ayant à disposition toutes les ressources nécessaires (méthodes statistiques et données) pour mettre en application ces méthodes, nous avons testé, dans le chapitre suivant, les potentialités d'application de nos méthodes aux données administratives de l'OFS.

Chapitre 14

Potentialités d'application des méthodes aux données administratives de l'OFS

Dans ce chapitre, nous voulons tester la possibilité d'application des différents modèles à nos données administratives tout en faisant un exemple d'analyse si c'est possible. Dans le cas contraire on mentionnera pourquoi l'analyse n'est pas adaptée. Nous reprendrons les groupes de modèles déjà présentés dans les chapitres précédents et appliquerons ces modèles à nos données.

14.1 Application aux méthodes non paramétriques

Dans les modèles non paramétriques, les méthodes de Kaplan-Meier et actuarielle ont été présentées ; nous allons à présent faire quelques analyses en appliquant ces deux méthodes à nos données tout en donnant le type de questions auxquelles on peut répondre avec ces données.

Pour ce faire, nous allons considérer la base de données ZAR qui contient les étudiants internationaux arrivés en Suisse entre 1997 et 2009 dans le but de poursuivre des études universitaires en Suisse. De cette base de données, nous avons extrait une autre base de données qui ne contient que les étudiants africains. Les exemples qui vont suivre porteront soit sur les étudiants internationaux de manière générale, soit sur les étudiants africains.

14.1.1 Application à la méthode de Kaplan-Meier

Dans cette analyse, nous nous intéressons à l'occurrence du mariage pour les étudiants internationaux arrivés en Suisse entre 1997 et 2009 selon le continent d'origine des étudiants.

Nous formulons cette question de recherche de la manière suivante : Le mariage des étudiants internationaux arrivés en Suisse entre 1997 et 2009 dépend-il du continent d'origine de ces étudiants ?

L'instant de départ c'est-à-dire le début d'observation est l'année 1997 et la fin de la période d'observation est l'année 2009 ; l'écart entre ces deux dates détermine la durée de l'étude soit douze ans. Nous nous intéressons à l'occurrence du mariage pour les étudiants internationaux selon le continent d'origine sur cette période de 12 ans. Notre base de données contient 21'318 étudiants internationaux répartis sur 53'797 lignes pour un total de 75'115 lignes. Le tableau 14.1 montre que nous avons 58'718 étudiants européens dont 17'960 se sont mariés sur cette période et 40'758 étudiants qui ne se sont pas mariés. Nous avons 4'843 étudiants africains, 1'878 se sont mariés et 2'965 qui sont sortis de l'étude sans être mariés. Nous avons respectivement 5'487 et 5'884 étudiants américains et asiatiques dont 2295 étudiants américains se sont mariés et 2'158 étudiants asiatiques aussi mariés.

Les figures 14.1 et 14.2 qui représentent la survie et la survie cumulée montrent les courbes de survie par nationalité ainsi que les observations censurées par nationalité. Mais comme nous pouvons l'observer, ces

Tableau 14.1 – Récapitulatif de traitement des observations

Continent	N total	Nombres d'événements	Censuré
Europe	58718	17960	40758
Africa	4843	1878	2965
America	5487	2295	3192
Asia	5884	2158	3726
Oceania	183	91	92
Global	75115	24382	50733

graphiques sont illisibles à cause de la taille de l'échantillon. Ceci rappelle une des limites de la méthode de Kaplan Meier. Chaque point matérialisé par une croix représente une censure et derrière chaque point se cachent des centaines voire des milliers de censures en raison du caractère annuel des données.

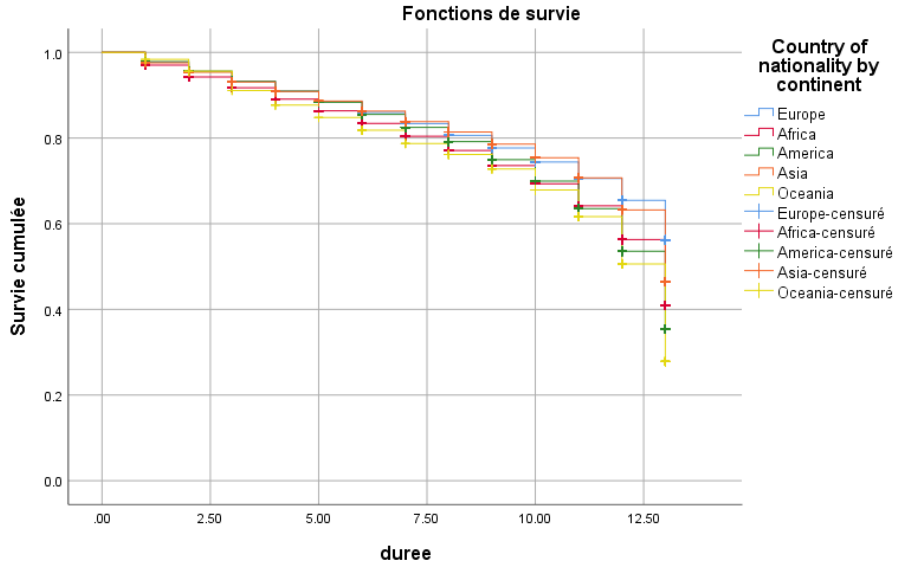


FIGURE 14.1 – Courbe de survie de l'estimateur de Kaplan-Meier

Nous avons également fait les tests du Log Rank et le test de Breslow pour comparer les groupes (les continents). Nous testons l'hypothèse nulle H_0 d'égalité des risques d'occurrence du mariage entre les différents groupes versus H_1, le risque d'occurrence du mariage est différent pour au moins un groupe. On a obtenu des statistiques de test de 404.210 et 166.074 respectivement pout le test du Log Rank et pour le test de Breslow qui sont toutes deux supérieures au $\chi^2_4 = 9.4877$ (p-val = 0.001 < 0.05), ce qui nous permet de rejeter H_0 au seuil de 5% et de conclure que le risque d'occurrence du mariage est différent d'un groupe à l'autre. On ne peut cependant pas savoir ce qui explique cette différence de risque parce que la méthode de Kaplan-Meier est une méthode descriptive.

L'exemple développé ici montre que la méthode de Kaplan-Meier peut s'appliquer à nos données mais à condition d'avoir de petits échantillons. On peut répondre à toutes les questions qui s'intéressent à l'occurrence d'un phénomène entre deux dates.

Avec cette méthode, on peut répondre à des questions de recherche du genre :

Quelle est la durée d'obtention d'un diplôme quelconque (bachelor, master ou doctorat) selon la nationalité ?

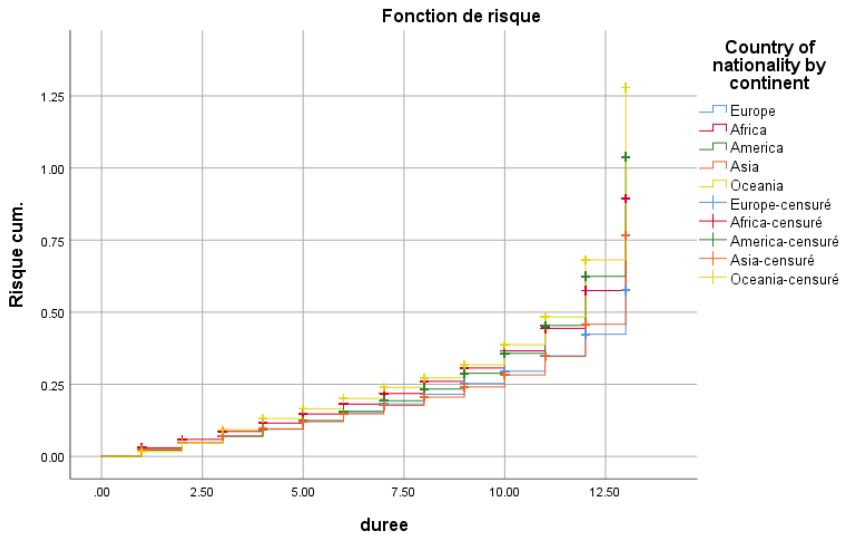


FIGURE 14.2 – Courbe du risque cumulé de l'estimateur de Kaplan-Meier

Le changement du type de permis selon la nationalité, le départ de la Suisse après l'obtention du diplôme selon la nationalité.

La durée des études de bachelor ou de master selon le genre.

Nous allons à présent utiliser la méthode actuarielle pour répondre à la même question de recherche, le but étant de voir si cette deuxième méthode présente des avantages ou des inconvénients par rapport à la méthode de Kaplan-Meier.

14.1.2 Application à la méthode actuarielle

Nous allons répondre à la même question de recherche que celle développée dans la méthode de Kaplan-Meir en utilisant la méthode actuarielle. On peut déjà mentionner un avantage de la méthode actuarielle par rapport à la méthode de Kaplan-Meier : la possibilité de grouper la durée par intervalle de temps ce qui nous permettra d'augmenter la lisibilité des graphiques d'une part et, d'autre part, de résumer les données. On peut librement définir la taille des intervalles de temps. Dans cet exemple on a choisi des intervalles de largeur 3, ce qui nous donne quatre intervalles de trois ans. Cette possibilité de grouper les informations dans des intervalles de temps nous donne des graphiques plus lisibles par rapport à ceux obtenus dans la méthode de Kaplan-Meier. Le graphique 14.3 de la survie actuarielle montre que dans les trois premières années de séjour, les cinq courbes sont confondues. Ceci veut dire que la probabilité d'occurrence du mariage est identique et est nulle ; en d'autre termes, il n'y a pas de mariage dans la première année de séjour. A la troisième année, on remarque une descente de la courbe et la forme en escalier commence à apparaître, ce qui veut dire qu'il y a eu des mariages dans la troisième année de séjour mais les courbes restent confondues. Entre la troisième et la sixième année, on remarque que la probabilité de mariage est plus élevée pour les étudiants africains que pour les autres étudiants parce que la courbe de survie des étudiants africains est plus basse que pour les autres nationalités. Entre six et douze ans, la probabilité de mariage est plus élevée pour les étudiants venant d'Océanie, d'Afrique et d'Amérique que pour les étudiants européens et asiatiques.

Le graphique 14.4 de la fonction de répartition montre l'évolution de la probabilité de mariage au cours du

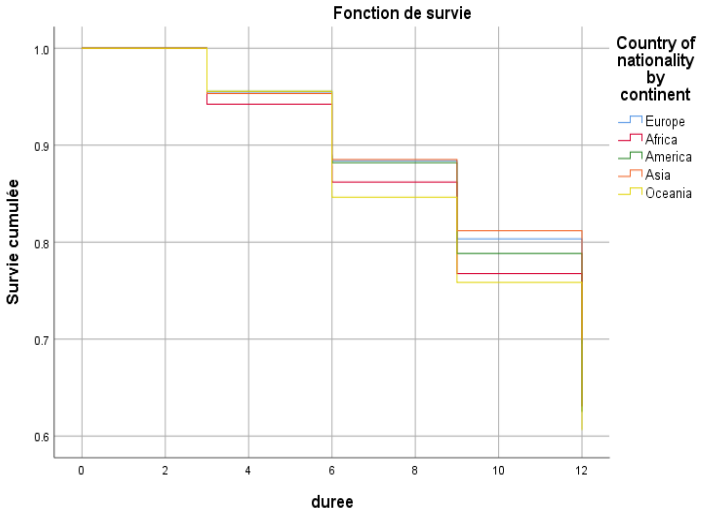


FIGURE 14.3 – Courbe de survie de l'estimateur actuariel

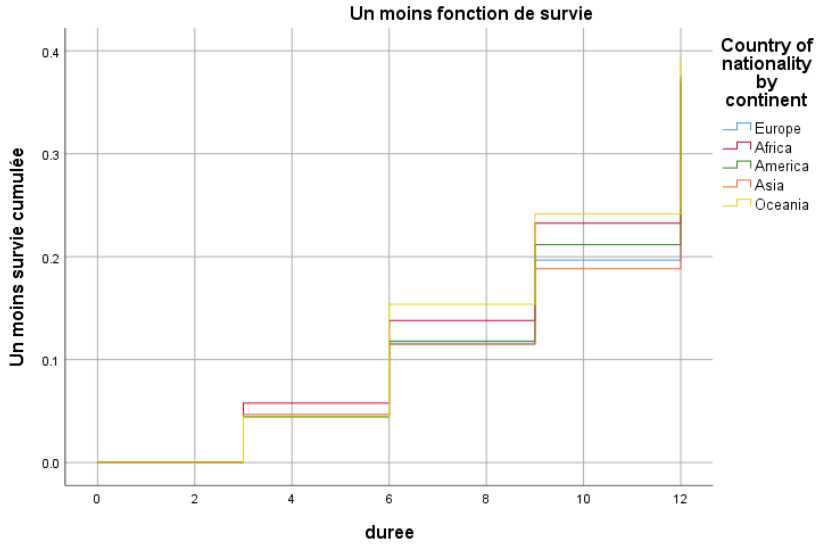


FIGURE 14.4 – Courbe de la fonction de répartition de l'estimateur actuariel

temps. On obtient les mêmes conclusions qu'avec la courbe de survie ; entre trois et six ans la probabilité de mariage est plus élevée pour les étudiants africains, entre six et neuf ans ce sont les étudiants venant d'Océanie

qui ont la plus forte probabilité de mariage et entre six et douze ans ce sont les étudiants venant d'Océanie, d'Afrique et d'Amérique qui ont respectivement les plus fortes probabilités de mariage.

Cet exemple montre que lorsque les échantillons sont grands, la méthode actuarielle est plus adaptée que la méthode de Kaplan-Meier.

14.2 Application aux méthodes paramétriques

Nous n'allons pas montrer un exemple d'application de modèles paramétriques dans la mesure où lorsque le vrai modèle paramétrique n'est pas connu, le modèle de Cox estime correctement les modèles paramétriques. Nous allons donc développer un exemple d'application du modèle de Cox dans la section suivante et cet exemple est généralisable aux modèles paramétriques. Les modèles paramétriques et de Cox sont sensibles aux mêmes problèmes, notamment aux groupes d'égalités qui ont déjà été mentionnés dans les sections précédentes.

Pour mieux illustrer cette notion de groupes d'égalités, considérons un échantillon de 1000 étudiants qui ont commencé un bachelor dans une université quelconque. Supposons qu'on s'intéresse au nombre d'étudiants qui vont obtenir le diplôme de bachelor trois ans après le début des études. Les modèles paramétriques peuvent conduire à une estimation biaisée des paramètres dès lors que 5% des individus expérimentent l'événement étudié. Cela veut dire que dans notre exemple, dès lors que 50 (0.05*1000) étudiants obtiennent le diplôme de bachelor la même année, les modèles paramétriques et le modèle de Cox peuvent conduire à l'estimation de paramètres biaisés. Pour 1000 étudiants qui commencent un cursus, nous nous attendons à ce que beaucoup plus que 50 étudiants obtiennent leur diplôme de bachelor la même année.

Cela est aussi valable pour la question de recherche dévéloppée dans les modèles non paramétriques sur l'occurence du mariage pour les étudiants internationaux arrivés en Suisse entre 1997 et 2009. La sensibilité aux groupes d'égalité explique le choix de ne pas présenter des exemples de modèles paramétriques dans cette section.

14.3 Application à la méthode semi-paramétrique : le modèle de Cox

Nous allons utiliser le modèle de Cox pour déterminer l'impact de quelques variables explicatives sur le risque de rester en Suisse après les études pour les étudiants africains.

La question de recherche sera formulée de la manière suivante :

Le risque pour les étudiants africains de rester en Suisse après les études dépend-il du mariage, des hautes écoles fréquentées (unifr, unige, unil, unine, unizh, HES, EPFL, EPFZ, unibe), de l'âge de ces étudiants, du sexe, de la cohorte d'entrée en Suisse (cohorte 1, cohorte 2, cohorte 3) et de la région d'origine de ces étudiants (Afrique de l'Ouest, Afrique Centrale, Afrique de l'Est, Afrique du Nord et Afrique Australe) ?

La base de données servant aux analyses est ZAR pour la période de 1997 à 2009.

La variable dépendante est le risque de rester en Suisse.

Les variables indépendantes sont : le mariage, les hautes écoles fréquentées (unifr, unige, unil, unine, unizh, HES, EPFL, EPFZ, unibe), l'âge, le sexe, cohorte 1 (étudiants entrés en Suisse entre 1997 et 2000), cohorte 2 (étudiants entrés en Suisse entre 2001 et 2005), cohorte 3 (étudiants entrés en Suisse entre 2006 et 2009), la région d'origine des étudiants (Afrique de l'Ouest, Afrique Centrale, Afrique de l'Est, Afrique du Nord et Afrique Australe).

La description des variables est présentée dans le tableau 14.2.

Pour cette analyse, les catégories de références sont les suivantes :

— Pour la variable mariage, la catégorie de référence qui prend la modalité 0 est non marié. On comparera ainsi le fait d'être marié par rapport à ne pas l'être.

Tableau 14.2 – Description des variables pour l'application du modèle de Cox aux données de l'OFS

Variables	Description	Nature	Valeurs(Modalités)
mariage	Mariée ou pas	Qualitative nominale	1= marié, 0 sinon
unifr	Etudes à l'uni de Fribourg	Qualitative nominale	1 =unifr, 0 sinon
unige	Etudes à l'uni de Genève	Qualitative nominale	1 =unige, 0 sinon
unil	Etudes à l'uni de Lausanne	Qualitative nominale	1 =unil, 0 sinon
unine	Etudes à l'uni de Neuchâtel	Qualitative nominale	1 =unine, 0 sinon
uniZH	Etudes à l'uni de Zurich	Qualitative nominale	1 =uniZH, 0 sinon
HES	Etudes dans une HES	Qualitative nominale	1 = HES, 0 sinon
EPFZ	Etudie à l'EPFZ	Qualitative nominale	1 =EPFZ, 0 sinon
EPFL	Etudie à l'EPFL	Qualitative nominale	1 =EPFL, 0 sinon
age	Age de la personne en années	Quantitative	
sexe	Sexe de la personne	Qualitative nominale	1 = homme, 0 = femme
coho1	Entrée en Suisse entre 1997 et 2000	Qualitative nominale	1=2003 à 2006, 0 sinon
coho2	Entrée en Suisse entre 2001 et 2005	Qualitative nominale	1=2007 à 2009, 0 sinon
coho3	Entrée en Suisse entre 2006 et 2009	Qualitative nominale	1=2010 à 2012, 0 sinon
AO	Vient de l'Afrique de l'Ouest	Qualitative nominale	1 si AO, 0 sinon
AE	Vient de l'Afrique de l'Est	Qualitative nominale	1 si AE, 0 sinon
AN	Vient de l'Afrique du Nord	Qualitative nominale	1 si AN, 0 sinon
AC	Vient de l'Afrique Centrale	Qualitative nominale	1 si AC, 0 sinon
AA	Vient de l'Afrique Australe	Qualitative nominale	1 si AA, 0 sinon

— Pour la variable sexe, la catégorie de référence est femme.
— Pour la variable cohorte, la catégorie de référence est la cohorte 3, c'est-à-dire les étudiants arrivés en Suisse entre 2006 et 2009. On comparera ainsi, pour les étudiants africains, l'impact sur la variable dépendante du fait d'être arrivé en Suisse durant les périodes de 1997 à 2000 et de 2001 à 2005 par rapport à ceux qui sont arrivés en Suisse entre 2006 et 2009.
— Pour la variable région d'origine des étudiants, la catégorie de référence est l'Afrique Australe. On comparera ainsi l'impact sur la variable dépendante du fait de venir d'une autre région d'Afrique par rapport au fait de venir de l'Afrique Australe.
— La catégorie de référence pour les universités est l'Université de Berne. Dans ce dernier cas, on comparera l'impact sur la variable dépendante du fait de fréquenter une autre université par rapport à fréquenter l'Université de Berne.

Nous avons réalisé un modèle de Cox et obtenu le tableau des coefficients qui est présenté dans le tableau 14.3.

Tableau 14.3 – Tableau des coefficients estimés pour le modèle de Cox

Variables	Coeff.	Erreur standard	Test de Wald	P-valeur	Exp(coeff.)
mariage	0.777	0.017	2110.765	0.001	2.174
Université de Fribourg	-0.357	0.081	19.419	0.001	0.700
Université de Genève	-0.984	0.07	195.874	0.001	0.374
Université de Lausanne	-0.747	0.098	57.881	0.001	0.474
Université de Neuchâtel	-0.578	0.126	21.153	0.001	0.561
Université de Zurich	-0.699	0.098	51.243	0.001	0.497
HES	-0.730	0.108	45.496	0.001	0.482
EPFL	-0.998	0.102	95.512	0.001	0.369
EPFZ	-1.812	0.127	204.303	0.001	0.163
age	0.036	0.001	1459.932	0.001	1.036
sexe	0.086	0.016	29.953	0.001	1.090
coho1	20.094	4.173	23.192	0.001	533142704.622
coho2	7.521	0.323	541.627	0.001	1847.136
Afrique-Ouest	-0.027	0.048	0.321	0.571	0.973
Afrique-Est	-0.046	0.172	0.073	0.787	0.955
Afrique-du-Nord	0.073	0.049	2.266	0.132	1.076
Afrique-Centrale	-0.237	0.102	5.401	0.02	0.789

Avec des odds ratios inférieurs à 1, par rapport à fréquenter l'Université de Berne, on observe un effet négatif des hautes écoles fréquentées sur le risque pour les étudiants africains de rester en Suisse après leurs études.

L'âge et le sexe ont des impacts positifs sur le risque pour les étudiants africains de rester en Suisse car lorsque l'âge augmente d'une unité, ce risque augmente de 3.6% alors qu'être un homme par rapport à être une femme toutes choses étant égales par ailleurs, ce risque de rester en Suisse après les études augmente de 9%. L'effet de cohorte montre que les étudiants africains entrés en Suisse entre 1997 et 2000 et ceux entrés en Suisse entre 2001 et 2005 ont plus de chance de rester en Suisse par rapport à ceux entrés en Suisse après 2005. La nationalité n'a pas d'effet significatif sur le risque pour les étudiants africains de rester en Suisse après les études sauf pour les étudiants d'Afrique Centrale où on observe un impact négatif sur le risque. Lorsqu'on vient d'Afrique Centrale par rapport à lorsqu'on vient d'Afrique Australe, le risque de rester en Suisse après les études diminue de 21.1%. Le mariage exerce aussi un effet positif sur le risque pour les étudiants africains de rester en Suisse car être marié par rapport à ne pas l'être toutes choses étant égales par ailleurs, le risque pour les étudiants africains de rester en Suisse après les études est environ 2.174 fois plus susceptible de se produire.

On remarque qu'il est possible d'appliquer le modèle de Cox à nos données et par conséquent on peut appliquer les modèles paramétriques à hasard proportionnel pour répondre à la même question de recherche.

Cependant, le problème avec les modèles paramétriques et le modèle de Cox, est la sensibilité aux groupes d'égalités. Plus la période d'observation est longue, plus la probabilité d'obtenir des groupes d'égalité est grande. Les données à notre disposition sont des données annuelles récoltées au 31 décembre de chaque année ce qui augmente la possibilité d'avoir des groupes d'égalité. Pour analyser les données à disposition, nous privilégierons donc le modèle logit à temps discret qui est plus compatible au type de données à notre disposition. Dans le chapitre suivant, nous allons présenter nos questions de recherche puis nous procéderons aux analyses dans le but de répondre à ces dernières.

Quatrième partie

Application aux étudiants internationaux et aux étudiants africains en Suisse

Chapitre 15

Le parcours de vie des étudiants internationaux et africains en Suisse

Dans ce chapitre, nous allons répondre à nos deux questions de recherche sur les étudiants internationaux et sur les étudiants africains ; ces dernières ont été présentées dans l'introduction (pages sept et huit) et seront rappelées ultérieurement. Nous commencerons d'abord par présenter l'évolution des migrations étudiantes internationales et africaines en Suisse à l'aide de nos données administratives. Nous reprendrons la partie analyse descriptive des données de notre article publié dans la revue Géo-Regards (Barry, 2017).

15.1 Evolution des migrations étudiantes internationales et africaines en Suisse

Les étudiants africains représentent environ 10.5% de l'effectif des étudiants en mobilité dans le monde et la part des étudiants africains en Suisse représentait environ 6% de l'effectif total des étudiants internationaux en 2013 (OFS, 2015). La Suisse, de par la qualité de son enseignement est une destination attractive pour les étudiants africains. Comme on peut le voir sur le graphique 15.1, le nombre d'étudiants internationaux en Suisse n'a cessé d'augmenter entre 1997 et 2014. On est passé d'un effectif de 30'508 étudiants internationaux en 1997 à un effectif de 44'556 étudiants en 2010, soit une augmentation de 46% sur cette période. Depuis 2008, l'augmentation du nombre d'étudiants internationaux en Suisse est très importante car elle est passée d'un effectif de 35'311 étudiants en 2008 à un effectif de 49'178 étudiants en 2011, pour s'établir à un effectif de 55'678 étudiants en 2014.

Le premier graphique de la Figure 15.2 montre l'évolution du nombre d'étudiants africains en Suisse entre 1997 et 2014. On remarque que cet effectif a augmenté de manière considérable durant cette période. En effet, on est passé d'un effectif de 608 étudiants africains en 1997 à un effectif de 1'267 étudiants dix ans plus tard soit une multiplication de l'effectif par deux. Cet effectif est passé à 1'568 étudiants en 2010 puis à 2'357 étudiants en 2013 avant de connaître une légère diminution en 2014 pour se situer à un effectif de 2'338 étudiants. De manière générale, entre 1997 et 2014, le nombre d'étudiants africains en Suisse a été multiplié par environ quatre, passant d'un effectif de 608 étudiants en 1997 à un effectif de 2'338 étudiants en 2014.

Il est dès lors possible de calculer la part [1] des étudiants africains en Suisse par rapport au total des étudiants internationaux (deuxième graphique de la figure 15.2). Cette part a aussi considérablement augmenté entre 1997 et 2014. Elle est passée de 2% en 1997 à 3.1% en 2003 avant de s'établir à 4% en 2009 puis à 4.2% en 2014. Le fléchissement de 2010 s'explique par une augmentation considérable du nombre d'étudiants internationaux en Suisse par rapport au nombre d'étudiants africains.

1. La part est calculée en faisant pour une année donnée, le rapport entre le nombre d'étudiants africains et le nombre total d'étudiants internationaux en Suisse pour la même année.

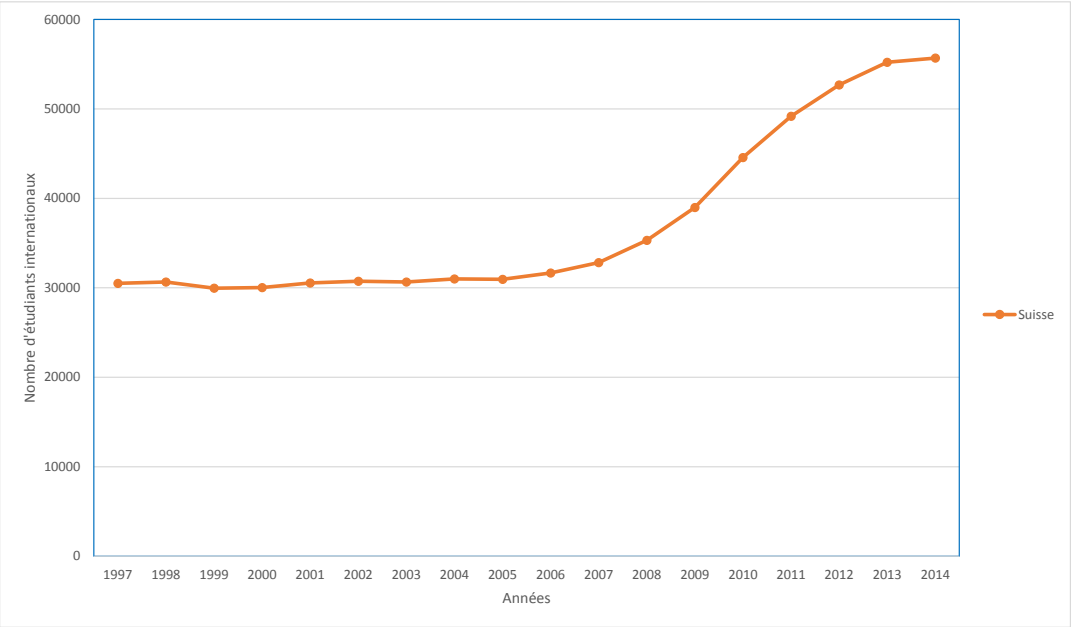


FIGURE 15.1 – Evolution de l'effectif des étudiants internationaux et étrangers en Suisse de 1999 à 2014
Source : OFS (2015), Transitions et parcours dans le degré tertiaire : Edition 2015, Neuchâtel

15.2 Analyse descriptive des données utilisées

Les données que nous utilisons pour les analyses qui vont suivre sont celles qui ont été présentées dans le chapitre 13 à la section 13.15. Dans cette section, nous avons expliqué comment les bases de données des étudiants internationaux et la base de données des étudiants africains ont été extraites.

L'analyse de la variable sexe montre que la majorité des étudiants africains en Suisse sont des hommes, soit 61.9% contre 38.1% de femmes [2] contrairement à la mobilité étudiante [3] européenne Erasmus qui compte une majorité de femmes (61% en 2000) (Cattan, 2004) ou à la mobilité étudiante au sein de l'OCDE qui comptait 54% de femmes en 2016. Les étudiants sont majoritairement célibataires (72.6%), mais 25.2% sont mariés et 1.6% divorcés. Ce taux élevé d'étudiants africains mariés pourrait s'expliquer par l'âge au début des études en Suisse. Il semble donc que la place des études en Suisse dans la trajectoire de vie des étudiants internationaux africains soit différente de celle de la plupart des autres étudiants. En effet, l'âge moyen des étudiants africains en Suisse est de 28 ans, la moitié d'entre-eux a moins de 28 ans, et l'autre moitié a plus de 28 ans, illustrant ainsi la symétrie de cette variable. La catégorie d'âge la plus représentée chez les étudiants africains en Suisse est 29 ans. Le plus jeune des étudiants a 17 ans et le plus âgé 50 ans pour une dispersion [4] de 33 ans. La part des étudiants de moins de 24 ans est de 25% et celle des moins de 32 ans est de 75%. Pour comparaison, l'âge moyen des étudiants suisses au début du bachelor est de 21 ans, la moitié de ces étudiants sont âgés de

2. Il ressort de cette 'étude qu'au Sénégal, pour les femmes, de longues études peuvent diminuer les possibilités de mariage d'une part, d'autre part, leurs intentions migratoires diminuent avec l'âge et le mariage (Efionayi et Piguet, 2014)
3. Mobilité géographique des étudiants au sein de l'OCDE
4. Ecart d'âge entre l'étudiant africain le plus âgé et le plus jeune

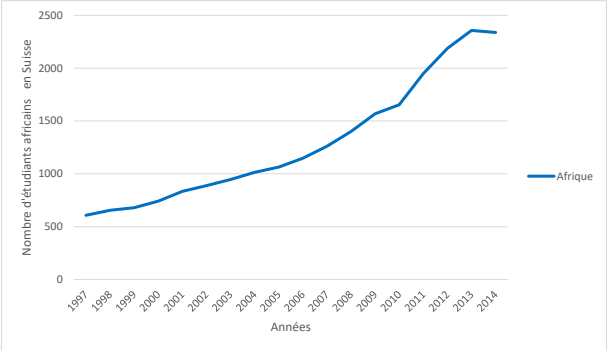


Figure 4

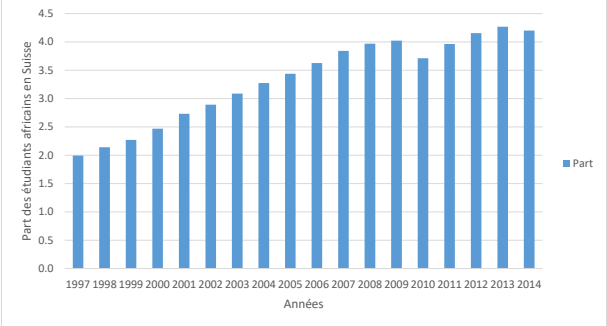


Figure 4a

FIGURE 15.2 – Evolution de l'effectif des étudiants africains en Suisse de 1997 à 2014
Source : OFS (2015), Transitions et parcours dans le degré tertiaire : Edition 2015, Neuchâtel

20 ans et 75% d'entre eux ont 22 ans ou moins. L'âge élevé des étudiants africains par rapport aux étudiants suisses pourrait s'expliquer par le fait que les étudiants africains accomplissent souvent une première formation universitaire dans leurs pays d'origine avant d'avoir des intentions migratoires (Efionayi et Piguet, 2014).

La majorité des étudiants africains en Suisse sont originaires de l'Afrique du Nord. Quatre pays, la Tunisie (19.8%), le Maroc (17.8%), l'Algérie (4.1%) et l'Egypte (5.7%) représentent à eux seuls près de la moitié de l'effectif avec un taux cumulé de 47.4% de l'effectif total des étudiants africains. En Afrique sub-saharienne, le Cameroun est de loin le pays le plus représenté avec un poids de 9.5%, suivi par le Sénégal, la Côte d'Ivoire et Madagascar avec des parts respectives de 4.9%, 3.6% et 3.5%. Le Nigéria (2.7%), le Ghana (2.6%), le Burkina Faso (2.4%) et l'Afrique du Sud (2.4%) complètent le podium des douze pays africains les plus représentés en Suisse. En ajoutant les effectifs des étudiants du Cameroun et du Sénégal à l'effectif des étudiants d'Afrique du Nord, on obtient 60.5% de l'effectif total des étudiants africains en Suisse. Cela signifie que six pays se partagent plus de 60% de la part des étudiants africains en Suisse. Le graphique 15.3 illustre ces résultats.

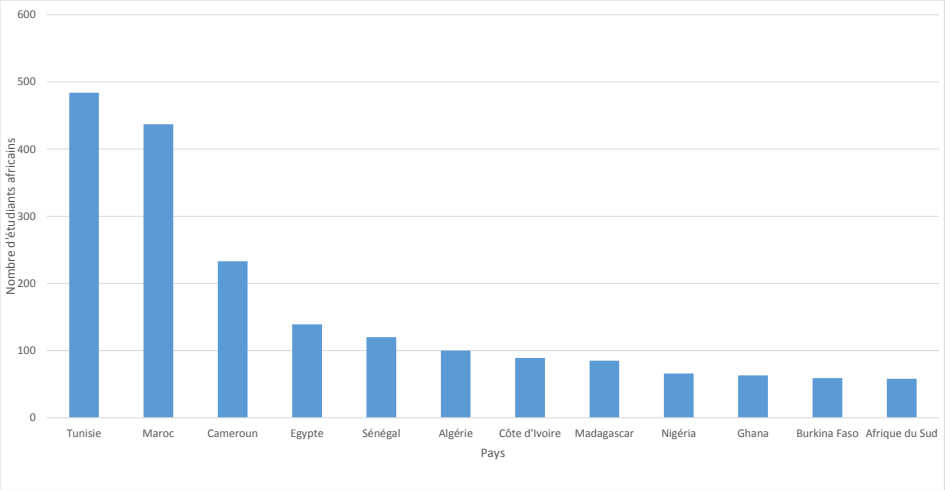


FIGURE 15.3 – : Les pays africains ayant le plus grand nombre d'étudiants en Suisse
Source : OFS (2015), Transitions et parcours dans le degré tertiaire : Edition 2015, Neuchâtel

En regroupant les pays par régions, on remarque que l'Afrique du Nord (48.3%) et l'Afrique de l'Ouest (32.5%) se partagent 80.8 % de l'effectif total des étudiants africains en Suisse. Les autres régions ne sont que faiblement représentées. Au début des études de bachelor, les étudiants africains sont très nombreux à fréquenter les HES (16.9% pour toutes les HES regroupées), suivi de l'Ecole Polytechnique Fédérale de Lausanne (9.9%) et des Universités de Lausanne et de Genève pour des parts respectives de 7.5% pour chacune des deux universités. La part des étudiants africains à l'Université de Neuchâtel est de 1.6%, soit quarante étudiants. Les filières privilégiées par ces étudiants au niveau du bachelor sont les branches techniques et les informations technologiques au niveau des HES (10.4%). Dans les universités et les écoles polytechniques fédérales, les étudiants africains s'orientent davantage vers les sciences économiques (7.6%), les sciences techniques (7.1%), les sciences exactes et naturelles (5.5%), la médecine et la pharmacie (4.1%), les sciences humaines et sociales (3.1%). La proportion d'étudiants africains dans les universités et les écoles polytechniques pour les autres filières est moindre. Si l'on compare le poids des étudiants africains dans les différentes branches avec le poids

qu'y occupent l'ensemble des étudiants, on peut remarquer que les étudiants africains sont surreprésentés [5] dans les branches techniques en bachelor au niveau des HES ; dans les universités et les écoles polytechniques fédérales, ils sont aussi surreprésentés dans les sciences économiques, les sciences techniques et les sciences exactes et naturelles. Ils sont sous-représentés dans les sciences humaines et sociales, en médecine et pharmacie, en chimie et science de la vie. Les formations de type bachelor ont débuté en 2001 et depuis, le nombre d'étudiants africains inscrits à ces formations n'a cessé d'augmenter. En 2003, il y avait 50 étudiants africains inscrits en bachelor, cet effectif est passé à 314 étudiants trois ans plus tard, avant de connaître un recul en 2007 pour se situer à 214 étudiants. Entre 2006 et 2014, l'effectif moyen des étudiants africains inscrits en bachelor dans les universités suisses était de 280 étudiants. Le graphique 15.4 illustre l'évolution en Suisse du nombre d'étudiants africains au niveau du bachelor ainsi que l'évolution du nombre de diplômes de bachelor décernés à des étudiants africains entre 2001 [6] et 2014. On remarque sur ce graphique que 2006 est l'année qui enregistre le plus grand nombre d'étudiants africains inscrits en bachelor dans les hautes écoles suisses. Le nombre de diplômes de bachelor décernés est en augmentation depuis 2004 jusqu'en 2010 avant de connaître un léger recul en 2014. Le nombre moyen de diplômes de bachelor décernés à des étudiants africains est de 109 diplômes sur la période de 2009 à 2014.

Au niveau du master, la tendance s'inverse : si les étudiants africains étaient nombreux à poursuivre leurs études de bachelor dans les HES, au niveau master, par contre, ils s'orientent vers les universités et les écoles polytechniques fédérales. Ils ne sont plus que 2.1% à poursuivre un master dans les HES contre 10.5% inscrits dans ce type d'école au niveau du bachelor. Les Universités de Genève, de Lausanne et l'EPFL sont les institutions qui comportent le plus d'étudiants africains au niveau du master avec des parts respectives de 12.1%, 8% et 6.5%. Avec 4.4% d'étudiants africains au niveau du master, l'Université de Neuchâtel se positionne devant l'Université de Fribourg (2.5%) parmi les institutions qui accueillent le plus d'étudiants africains. En Suisse alémanique, c'est l'Université de Bâle (2.6%) qui accueille le plus d'étudiants africains. Les étudiants africains sont nombreux à s'inscrire pour le master dans les filières de sciences exactes et naturelles (9.6%), les sciences humaines et sociales (9.2%), les sciences économiques (8.4%), les sciences techniques (5.3%), le droit (3.9%) et la médecine et la pharmacie (2.3%).

Le second graphique 6a de la figure 15.4 décrit l'évolution du nombre d'étudiants africains en master entre 2002 et 2014 ainsi que le nombre de diplômes de master décernés à des étudiants africains entre 2005 et 2014.

Les étudiants africains sont nombreux à poursuivre leurs études doctorales dans les universités de Genève (5.9%), de Bâle (3.1%) et de Lausanne (2.4%). Les autres institutions d'études supérieures accueillent des étudiants africains dans des proportions moindres, à l'image de l'ETH de Zurich (1.7%), de l'EPFL (1.3%), de l'Université Fribourg (1.5%) et de Neuchâtel (1.1%). Le graphique 15.5 montre l'évolution de l'effectif des doctorants africains en Suisse entre 1980 et 2014 ainsi que celle du nombre de diplômes de doctorats décernés à des étudiants africains durant cette période. Comme on peut le remarquer sur ce graphique, les années 1982, 1988, 2004 et 2011 sont les années durant lesquelles ont été enregistrés les plus grands effectifs d'étudiants africains au doctorat. Les années 2011 et 2004 étant les deux années ayant enregistré les effectifs les plus élevés avec des parts respectives de 109 et 103 doctorants. Les années 1989 et 1993 étant celles ayant enregistré le moins d'étudiants africains inscrits au doctorat avec des parts respectives de 30 et 43 doctorants.

Quant au nombre de titres de doctorats décernés à des étudiants africains en Suisse, les années 2006 et 2007 sont celles ayant enregistré les plus grands effectifs de diplômes décernés.

Nous complétons cette analyse descriptive par l'évolution de l'effectif des étudiants suisses dans les hautes écoles suisses. Le but de ce complément est de présenter l'évolution de l'effectif des étudiants suisses du bachelor au doctorat. Les données à disposition couvrent la période de 1980 à 2014 pour le doctorat et les périodes de 2000 à 2014 pour les niveaux de formation de bachelor et de master. Les données utilisées pour réaliser cette étude sont celles du LABB (2015). La base de données ne contenant que les étudiants suisses a été obtenue sur le critère de la nationalité en créant un filtre qui permet de ne sélectionner que les étudiants de nationalité suisse. L'analyse de la variable sexe montre qu'il y a majoritairement plus d'hommes que de femmes dans l'enseignement supérieur suisse avec des parts respectives de 54.5% d'hommes pour 45.5% de femmes.

5. Rapport entre les effectifs observés des Africains et les effectifs attendus sur la base de l'ensemble des étudiants. Si ce rapport est > 1, cela veut dire que le groupe est surreprésenté et vice-versa

6. Avant 2001, la réforme de Bologne n'était pas encore appliquée (le système bachelor, master n'existait pas)

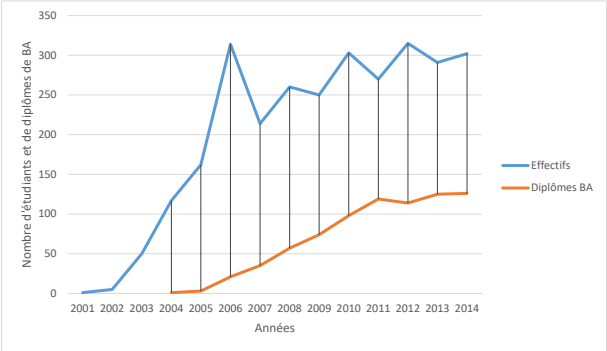


Figure 6 : Diplômes de bachelor

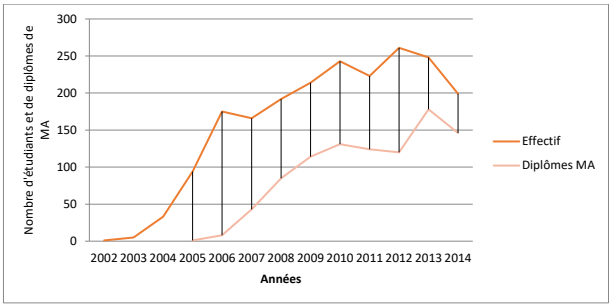


Figure 6a : Diplômes de master

FIGURE 15.4 – : Evolution du nombre d'étudiants africains en Suisse au niveau du bachelor ainsi que du nombre de diplômes de bachelor décernés à des étudiants africains entre 2001 et 2014.
Source : OFS (2015), Transitions et parcours dans le degré tertiaire : Edition 2015, Neuchâtel

15.2.1 Evolution de l'effectif des étudiants suisses au niveau du bachelor

Le nombre d'étudiants de nationalité suisse entrant au bachelor est passé de $1'498$ étudiants en 2001 à $17'978$ étudiants en 2005, avant de se situer à $27'199$ étudiants en 2010. En 2014, il y avait $28'850$ étudiants suisses au niveau du bachelor. Sur la période allant de 2001 à 2014, le nombre d'étudiants suisses au niveau du bachelor a été multiplié par plus de dix-neuf entre. Comme le montre le graphique 15.6, la plus forte augmentation du nombre d'étudiants suisses au bachelor a été enregistrée entre 2004 et 2005, l'effectif est passé de $6'290$ étudiants en $2'004$ à $17'948$ étudiants en 2005, soit une multiplication de l'effectif par presque trois sur cette période. Une légère baisse de l'effectif a été cependant enregistrée entre 2009 et 2010, en passant de $27'542$ étudiants en 2009 à $27'299$ étudiants en 2010 soit une diminution de l'effectif de 243 individus.

Le nombre de diplômes décernés à des étudiants suisses au niveau du bachelor a aussi considérablement augmenté depuis 2005. On est passé de $1'244$ diplômes en 2004 à $9'531$ diplômes trois ans plus tard, pour se situer à $29'089$ diplômes en 2013, puis à $28'850$ diplômes en 2014. La plus grande hausse du nombre de diplômes de bachelor décernés a été observée entre 2007 et 2009 où le nombre de diplômes décernés a été multiplié par plus de 3.5.

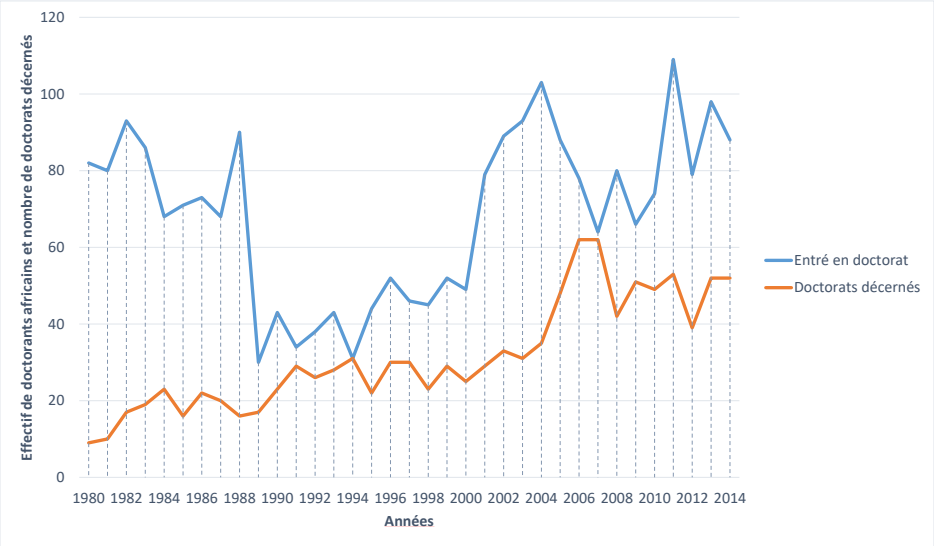


FIGURE 15.5 – : Evolution du nombre de doctorants africains en Suisse ainsi que du nombre de diplômes de doctorat décernés à des étudiants africains entre 1980 et 2014.
Source : OFS (2015), Transitions et parcours dans le degré tertiaire : Edition 2015, Neuchâtel

15.2.2 Evolution de l'effectif des étudiants suisses au niveau du master

Au niveau du master, c'est l'ETH de Zurich, l'Université de Berne et l'Université de Zurich qui accueillent le plus grand nombre d'étudiants suisses avec des parts respectives de $9'498$, $8'784$ et $8'690$ étudiants. L'ETH de Zurich et l'Université de Berne sont les deux hautes écoles suisses qui accueillent les plus grands effectifs d'étudiants suisses au niveau du master. Les Universités de Bâle et de Fribourg occupent la quatrième et la cinquième place avec des parts respectives de $6'508$ et $5'907$ étudiants. Depuis 2003, le nombre d'étudiants suisses au niveau du master n'a cessé d'augmenter en passant d'un effectif de $1'221$ étudiants en 2005, à un effectif de $6'666$ étudiants en 2008, puis à $11'323$ étudiants en 2010, avant de se situer à $13'248$ étudiants en 2014, comme l'illustre le graphique 15.7. Le nombre de diplômes de master décernés à des étudiants suisses a aussi considérablement augmenté en passant de 528 diplômes en 2006 à $3'246$ diplômes en 2009, puis à $10'773$ diplômes en 2014.

15.2.3 Evolution de l'effectif des étudiants suisses au niveau du doctorat

Contrairement à l'évolution de l'effectif des étudiants suisses au niveau du bachelor et au master où on observait une tendance nette à la hausse, pour le doctorat, on observe une fluctuation de la tendance : tantôt à la hausse et tantôt à la baisse. Les augmentations les plus marquées de l'effectif des étudiants suisses au niveau du doctorat sont entre les périodes de 1982 à 1983, puis entre 1995 et 1996 et pour finir entre 2007 et 2008. Entre 1982 et 1983, l'effectif des étudiants suisses en doctorat est passé de $1'909$ doctorants en 1982 à $2'621$ doctorants en 1983 ; entre 1995 et 1996, cet effectif est passé de $1'457$ doctorants à $2'553$ doctorants. La figure 15.8 montre cette évolution.

En ce qui concerne le nombre de diplômes de doctorat décernés, on observe une augmentation successive du nombre de diplômes pour les périodes de 1980 à 1986 et de 1988 à 1990 et à une baisse successive du nombre de diplômes pour les périodes de 2002 à 2004.

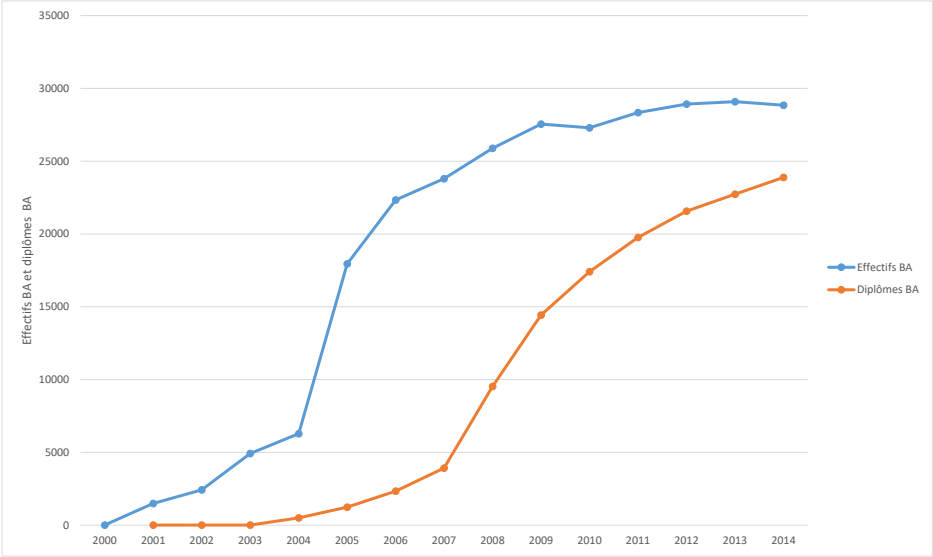


FIGURE 15.6 – Evolutions du nombre d'étudiants suisses et du nombre de diplômes de bachelor décernés à des étudiants suisses au niveau du bachelor entre 2000 et 2014. Source : LABB(2015)

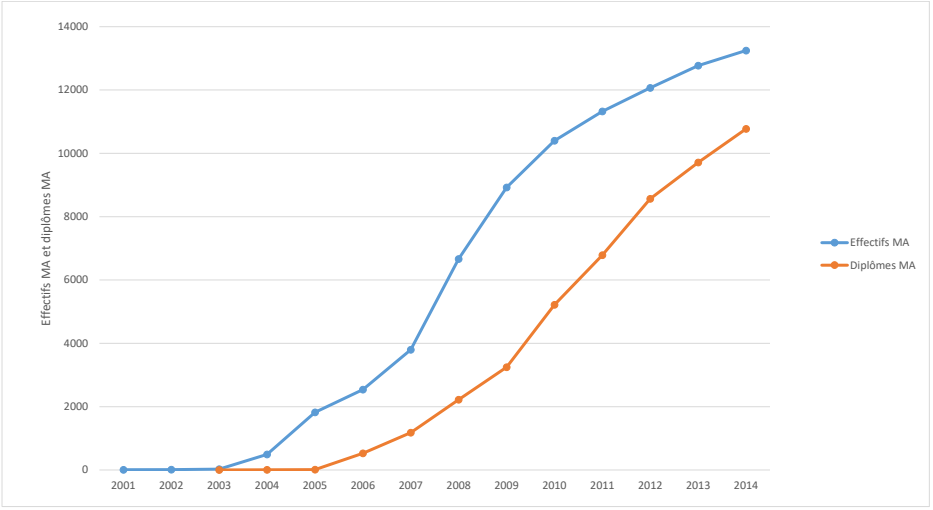


FIGURE 15.7 – Evolutions du nombre d'étudiants suisses et du nombre de diplômes de master décernés à des étudiants suisses au niveau du master entre 2001 et 2014. Source : LABB(2015)

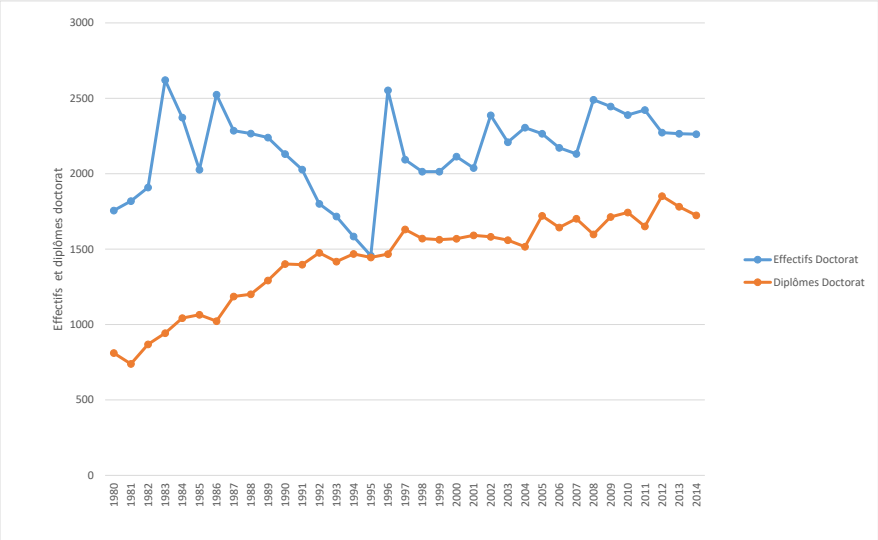


FIGURE 15.8 – : Evolutions du nombre d'étudiants suisses et du nombre de diplômes de doctorat décernés à des étudiants suisses au niveau du doctorat entre 1980 et 2014. Source : LABB(2015)

Ces chiffres sur l'évolution de l'effectif des étudiants suisses au niveau du bachelor, du master et du doctorat sont en parfaite harmonie avec une population suisse de plus en plus tournée vers le secteur tertiaire. En effet, selon une étude de 2006 du Secrétariat d'Etat à l'éducation et à la recherche (SER) et l'Office fédéral de la formation professionnelle et de la technologie (OFFT), en collaboration avec Présence Suisse et la Conférence universitaire suisse, il y a de plus en plus de Suisses qui font des études supérieures à l'image de ce qu'on observe dans les pays de l'OCDE (SER, 2006). L'étude mentionne que : *« Suivant la tendance internationale, le niveau d'éducation des Suisses augmente régulièrement. Alors que 42 % des personnes âgées aujourd'hui de plus de 65 ans ont terminé leur formation avec l'école obligatoire, elles ne sont que 10 % dans la catégorie d'âge de 20 ans. En parallèle, la proportion des personnes avec une formation de niveau tertiaire a augmenté. Elle est aujourd'hui de 26% parmi les personnes âgées de 25 à 34 ans ».*

Le taux de diplômés universitaires suisses reste cependant faible par rapport à la moyenne de l'OCDE. Cela s'explique par la performance du système de l'apprentissage qui se traduit par la formation de jeunes qui s'insèrent facilement sur le marché de l'emploi. Cette tendance s'observe également dans la statistique universitaire de l'Université de Genève où on a une part importante des étudiants des autres cantons suisses qui s'inscrivent dans une formation universitaire (Unige, 2016). Cette statistique mentionne le fait qu'en 2016, à l'Université de Genève, 43% de l'effectif des étudiants venaient du canton de Genève, 20% des autre cantons et 37% des étudiants venaient de l'étranger. La part des étudiants suisses dans cette université en 2016 se situerait donc à 53%. Parmi la répartition des étudiants issus des autres cantons (sans les étudiants genevois), le canton de Vaud est le plus représenté (près de 47%), suivi des cantons du Valais (environ 15%), du Tessin (environ 5.5%), de Neuchâtel (5.4%) et de Fribourg (environ 5.2%). Si les étudiants suisses sont majoritaires au bachelor et au mater, cette étude montre que les étudiants internationaux et étrangers sont plus nombreux aux formations MAS (Master of advanced studies), aux formations de spécialisation et au doctorat.

Les prévisions de l'Office fédéral de la statistique confirment également cette tendance à la hausse de l'effectif des étudiants dans les hautes écoles universitaires suisses. En effet, selon l'OFS (OFS, 2018), en 2027, il y aura 267'000 étudiants dans les hautes écoles suisses, soit une progression de 8% sur la période 2017-2027 (+6% dans les hautes écoles universitaires, + 10% dans les hautes écoles spécialisées, et +12% dans les hautes

écoles pédagogiques). Le nombre de diplômes délivrés n'est pas en marge de cette prévision, car la même étude mentionne qu'en ce qui concerne les nouveaux diplômés, les croissances les plus fortes sur les dix prochaines années seront dans le domaine des technologies de l'information et de la communication avec une hausse de +53% de masters délivrés entre 2017 et 2027 (OFS, 2018).

Nos résultats donnent une base pour des analyses plus approfondies sur les étudiants internationaux et africains en Suisse à travers l'étude de leurs parcours de vie à l'aide de méthodes statistiques. Nous commencerons d'abord par le parcours de vie des étudiants internationaux en Suisse.

15.3 Application aux étudiants internationaux

La première question de recherche est formulée de la manière suivante :

Le fait que les étudiants internationaux arrivés en Suisse entre 1997 et 2009 restent ou quittent la Suisse après leurs études dépend-il du mariage de ces étudiants en Suisse durant cette période, de l'âge de ces étudiants en arrivant en Suisse, du sexe, de la haute école fréquentée (Unifr, Unige, Unil, Unine, Uniba, UniZH, HES , EPFZ), du continent d'origine des étudiants (Afrique, Asie, Europe, Amérique) et de la cohorte d'entrées (entrées entre 1997 et 2000, entrées entre 2001 et 2005, entrées entre 2006 et 2009) ?

Le choix de la période allant de 1997 à 2009 s'explique d'une part par la législation suisse en matière d'accueil des étrangers en vigueur jusqu'en 2011, d'autre part, par la disponibilité d'une base de données des étudiants internationaux couvrant cette période. En effet, les étudiants internationaux entrés en Suisse durant cette période n'avaient que le mariage comme possibilité de rester en Suisse après leurs études ; le motif médical n'est pas considéré dans cette étude. La disponibilité d'une base de données couvrant le parcours migratoire des étudiants internationaux pour la période de 1997 à 2009 permet de fusionner directement cette base de données avec celle du LABB (base de données des étudiants) dans le but de mettre en commun le parcours migratoire et le parcours académique des étudiants. Cette base de données permet aussi une analyse du parcours de vie des étudiants internationaux sur une période de treize ans sans interruption et avec des données enregistrées dans le même format.

Les analyses descriptives avaient montré qu'en moyenne, les étudiants africains étaient plus âgés que les étudiants suisses au début des études parce que, généralement, les étudiants africains sont au bénéfice d'une première formation accomplie en Afrique avant le début des études en Suisse. Nous partons donc du principe que la variable âge est aussi importante pour expliquer le phénomène étudié. La variable continent d'origine nous permettra de comprendre le rôle de la nationalité (continent d'origine des étudiants) sur la probabilité de prolonger le séjour en Suisse après les études. La littérature sur les mobilités étudiantes s'est essentiellement centrée sur les facteurs qui influencent la décision de poursuivre des études supérieures à l'étranger ainsi que le choix du pays de destination et de l'institution d'éducation supérieure dans ce pays. Ces études sont basées surtout sur des données d'enquêtes transversales ne permettant pas d'examiner les trajectoires individuelles migratoires, scolaires, professionnelles et familiales de la population étudiante.

Les données à disposition étant des données administratives, nous présenterons dans la section suivante la méthodologie qui a été utilisée pour construire la variable dépendante qui n'est pas disponible dans les données à disposition.

15.3.1 Méthodes

La base de données ZAR à disposition couvre la période de 1997 à 2009 et ne contient que des étudiants internationaux et étrangers en Suisse pour la période mentionnée. Dans cette base de données, un étudiant peut apparaître sur plusieurs lignes selon la durée de son séjour en Suisse, cependant elle ne contient aucune variable académique. La base de données de LABB(2015) contient le parcours académique des étudiants internationaux et étrangers en Suisse pour la période de 1980 à 2014. Ces deux bases de données ont été fusionnées pour mettre en commun les parcours migratoires et académiques de ces étudiants internationaux et étrangers.

La définition des étudiants internationaux est celle de l'OCDE, de l'UNESCO et de l'Eurostat qui définit les étudiants internationaux comme des étudiants qui ont quitté leur pays pour se rendre dans un pays étranger dans le seul but d'y étudier. Ils ne doivent pas être ressortissants de ce pays d'accueil, ils ne doivent pas avoir résidé antérieurement dans ce pays et ils ne doivent pas avoir accompli une formation antérieure dans ce pays d'accueil. Il est dès lors essentiel de retirer de la base de données tous les étudiants étrangers en se basant sur le critère du certificat qui leur a permis d'accéder à l'enseignement supérieur suisse, ainsi que sur ceux du pays de naissance et du pays de résidence avant les études. Le certificat qui a permis à ces étudiants d'accéder à l'enseignement supérieur suisse doit être délivré par un pays étranger (formation antérieure non effectuée en Suisse), le pays de résidence avant le début des études doit être à l'étranger (résidence antérieure pas en Suisse) et le pays de naissance doit être à l'étranger. La disponibilité de ces variables dans les deux bases de données facilite la création d'un filtre permettant de ne sélectionner que les étudiants internationaux. La base de données ne contient donc que des étudiants qui ont quitté leurs pays pour venir étudier en Suisse, qui n'ont pas étudié antérieurement en Suisse et qui ne sont pas nés en Suisse. L'objectif est d'analyser les variables qui pourraient être déterminantes pour expliquer le fait que les étudiants internationaux prolongent leur séjour en Suisse après leurs études à l'aide d'un modèle statistique. Les variables explicatives sont toutes disponibles dans la base de données et seront présentées ultérieurement en détail avant de procéder aux analyses. La variable qui n'est pas disponible dans la base de données est la variable dépendante ou variable à expliquer ; cette variable est « prolonger le séjour en Suisse après les études ». La non disponibilité de cette variable rappelle l'une des limites des données administratives : ce ne sont pas des données qui en soit sont destinées à la recherche malgré l'immense potentiel qu'elles offrent en matière de données longitudinales à moindres coûts.

La non disponibilité de cette variable dépendante nous a conduits à explorer davantage la base de données dans le but de trouver une ou des variables qui permettraient de construire cette variable dépendante. Le choix s'est porté sur la variable « permis », car c'est le type de permis qu'on détient qui détermine le caractère temporaire ou non de notre séjour. Cette variable « permis » de notre base de données a dix modalités pour dix types différents de permis de séjour. Nous disposons également de la variable « nationalité du partenaire ». Parmi les modalités de la variable « permis », nous avons entre autres le permis étudiant (permis B court séjour et renouvelable chaque année), le permis d'établissement et d'autres permis de types courts séjours (permis frontalier, saisonnier,...). Nous formulons l'hypothèse qu'un étudiant qui finit les études pour la période antérieure à 2011 en étant détenteur d'un permis B pour études doit quitter la Suisse à cause du caractère temporaire de son permis. La variable nationalité du partenaire a trois modalités qui sont : célibataire, marié à un étranger (étrangère) et marié à un Suisse (Suissesse). L'idée principale était de dire que tous les étudiants qui ont un permis d'établissement ou qui sont mariés à un Suisse ou une Suissesse ou à un étranger établi en Suisse vont prolonger le séjour en Suisse après les études et tous ceux qui ont un permis court séjour ou qui sont célibataires vont partir après les études.

En réalisant un tableau croisé entre les variables permis et nationalité du partenaire, cette idée a été abandonnée car le tableau croisé a montré plus clairement les caractéristiques des étudiants qui ont des permis d'établissement.

En effet, le tableau croisé a montré que sur les 16'497 étudiants établis, 2'122 étudiants sont mariés à des Suisses ou Suissesses, 7'735 étudiants sont mariés à des étrangers ou étrangères et 6'640 étudiants célibataires sont établis. Ce tableau montre donc qu'il y a plus d'étudiants célibataires établis que d'étudiants mariés à des Suisses (Suissesses) d'une part, d'autre part, qu'il y a plus d'étudiants mariés à des étrangers que d'étudiants mariés à des Suisses ou Suissesses. Ceci nous a poussés à analyser la répartition de la variable nationalité du partenaire par continent d'origine des étudiants. Ce tableau croisé a montré que l'essentiel de ces étudiants viennent du continent européen. En effet, sur les $5'922$ étudiants mariés à des Suisses ou Suissesses, $4'079$ proviennent du continent européen, 515 du continent africain, 859 du continent américain, 446 du continent asiatique et 23 de l'Océanie.

Sur les $18'460$ étudiants mariés à des étrangers, $13'881$ proviennent du continent européen, $1'363$ du continent africain, $1'436$ du continent américain, $1'712$ du continent asiatique et 68 de l'Océanie. A partir de ce tableau croisé entre le type de permis et la nationalité du partenaire, nous avons fait l'hypothèse que tous les étudiants qui sont établis, qu'ils soient mariés à des Suisses, à des étrangers ou célibataires vont prolonger le séjour en Suisse après les études et tous ceux qui ont d'autres types de permis court séjour vont partir après les études. La

variable dépendante « prolonger le séjour en Suisse après les études » qui est dichotomique prendra la valeur 1 si l'étudiant prolonge le séjour en Suisse après les études et la valeur 0 sinon. En raison de la nature dichotomique de la variable dépendante et du fait que les données à disposition soient des données individuelles et annuelles, nous avons opté pour un modèle logit à temps discret. La particularité et la complexité de ce modèle réside dans l'organisation des données sous forme de personne-période. Une fois ces données organisées sous cette forme, on n'a plus qu'à réaliser un modèle de régression logistique classique sur ces données pour estimer les différents paramètres (Le Goff et al., 2013).

15.3.2 Données et description des variables

L'originalité de cette étude réside dans la nature longitudinale quantitative et administrative des données qui seront utilisées pour répondre aux questions de recherche d'une part, d'autre part, elle résulte du fait que très peu d'études ont été réalisées ailleurs et en Suisse pour étudier le parcours de vie des étudiants internationaux. La plupart des recherches quantitatives portant sur les migrations internationales d'Afrique subsaharienne sont basées sur les données produites par les principaux pays de destination sur les effectifs d'étudiants étrangers inscrits dans un établissement d'enseignement supérieur. Les variables qui seront utilisées pour répondre à la première question de recherche sont présentées dans le Tableau 15.1.

Tableau 15.1 – Description des variables de la première question de recherche

Variables	Description	Nature	Valeurs (Modalités)
séjour	prolonge ou pas	Quali. nominale	1= prolonge et 0 = sinon
age	Age de la personne en année	Quantitative	
mariage	Si marié ou pas	Quali. nominale	1= marié, 0 sinon
sexe	Sexe de la personne	Quali. nominale	1 =homme, 0 = femme
afrique	Etudiant africain	Quali. nominale	1=africain, 0 sinon
europe	Etudiant européen	Quali. nominale	1=européen, 0 sinon
asie	Etudiant asiatique	Quali. nominale	1=asiatique, 0 sinon
amerique	Etudiant sud américain	Quali. nominale	1=sud-américain, 0 sinon
coho1	Entrées entre 1997 et 2000	Quali. nominale	1=1997 à 2000, 0 sinon
coho2	Entrées entre 2001 et 2005	Quali. nominale	1= 2001 à 2005, 0 sinon
coho3	Entrées entre 2006 et 2009	Quali. nominale	1=2006 à 2009, 0 sinon
duree	Durée de séjour en année	Quantitative	
unifr	Etudes à l'uni de Fribourg	Quali. nominale	1 =unifr, 0 sinon
unige	Etudes à l'uni de Genève	Quali. nominale	1 =unige, 0 sinon
unil	Etudes à l'uni de Lausanne	Quali. nominale	1 =unil, 0 sinon
unine	Etudes à l'uni de Neuchâtel	Quali. nominale	1 =unine, 0 sinon
uniba	Etudes à l'uni de Bâle	Quali. nominale	1 =uniba, 0 sinon
HES	Etudes dans une HES	Quali. nominale	1 = HES, 0 sinon
EPFZ	Etudie à l'EPFZ	Quali. nominale	1 =EPFZ, 0 sinon
EPFL	Etudie à l'EPFL	Quali. nominale	1 =EPFL, 0 sinon
uniZH	Etudes à l'uni de Zurich	Quali. nominale	1 =uniZH, 0 sinon

Pour cette question de recherche, les catégories de référence sont les suivantes :

- Pour la variable mariage, la catégorie de référence qui prend la modalité 0 est non marié. On comparera ainsi le fait d'être marié par rapport à ne pas l'être.
- Pour la variable sexe, la catégorie de référence est « femme ».
- Pour la variable cohorte, la catégorie de référence est la cohorte 3, c'est-à-dire les étudiants arrivés en Suisse entre 2006 et 2009.
- La catégorie de référence pour les universités est l'Université de Berne.
- Pour les variables de continent, la catégorie de référence ce sont les étudiants ressortissants de l'UE. On comparera ainsi le fait de venir d'un continent donné par rapport au fait de venir d'un pays membre de l'UE.

15.3.3 Préparation des données en vue de l'analyse

La principale difficulté liée à l'application de cette méthode réside dans la préparation des données sous la forme de personne-période. Cette organisation consiste à faire ressortir le caractère longitudinal des données de manière à ce qu'une personne apparaisse sur autant de lignes qu'elle a été présente en Suisse, chaque ligne représentant une année. Le nombre de lignes pour une personne donnée correspond à la durée de séjour de cette personne en Suisse. Pour la première année de séjour en Suisse, la variable durée prendra la valeur 1, puis la valeur deux pour la deuxième année et la valeur k pour la k-ième année de séjour en Suisse. Organiser cette base de données sous cette forme n'a pas été compliqué grâce au travail réalisé en amont par l'OFS. En effet, la variable année de présence en Suisse a été créée par l'OFS dans les formats suivants : jj.mm.aaaa, aaaa et aa. Par exemple, pour quelqu'un qui est arrivé en Suisse le 15.08.2005, on a à disposition en plus de ce format, les formats 2005 et 05. Ces différents formats nous permettent de créer la variable durée pour cet étudiant qui prendra la valeur 1 pour l'année 2005, la valeur 2 pour l'année 2006, la valeur 3 pour l'année 2007, jusqu'à la dernière année de séjour. Chaque année sera sur une ligne avec les caractéristiques individuelles correspondantes ainsi que les différents états. Un extrait de la base de données organisée sous forme de personne période est présenté dans le tableau 15.2.

Tableau 15.2 – Un extrait de la base de données organisée sous la forme de personne-période

Id	presence	duree	age	sexe	coho1	coho2	coho3	mariage	...
2961703	1997	1	36	homme	1	0	0	non	...
2961703	1998	2	37	homme	1	0	0	non	...
2961703	1999	3	38	homme	1	0	0	non	...
2944775	1997	1	39	femme	1	0	0	non	...
2944775	1998	2	40	femme	1	0	0	non	...
2944775	1999	3	41	femme	1	0	0	non	...
2944775	2000	4	42	femme	1	0	0	non	...
2944775	2001	5	43	femme	1	0	0	non	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

15.3.4 Interprétation des résultats

Le modèle utilisé ici est le modèle logit à temps discret. L'effet d'une variable sur la chance d'occurrence de l'événement d'intérêt se mesure à travers l'odds ratio aussi appelé rapport de cote. Si l'odds ratio est supérieur à 1, l'effet de la variable correspondante sur la probabilité d'occurrence de l'événement d'intérêt augmente, s'il est inférieur à 1, l'effet sera négatif et s'il vaut 1, la variable sera sans effet sur les chances d'occurrence de l'événement étudié.

Avec des odds ratio tous inférieurs à 1, le tableau 15.3 des résultats montre que les universités fréquentées ont un impact négatif sur les chances de prolonger le séjour en Suisse après les études par rapport au fait de fréquenter l'Université de Berne. Les odds ratios associés aux continents d'origine des étudiants sont tous supérieurs à 1, tandis que le odds ratio associé à la variable durée de séjour est inférieur à 1. Ceci nous permet de dire que par rapport au fait de venir d'un pays de l'Union Européenne, le continent d'origine des étudiants et la durée de séjour ont respectivement des impacts positifs et négatif sur les chances de prolonger le séjour en Suisse après les études. Le fait d'être marié par rapport à ne pas l'être exerce un impact positif sur les chances de prolonger le séjour en Suisse après les études. Les cohortes entrées en Suisse entre 1997 et 2000, et entre 2001 et 2005 par rapport à celles entrées en Suisse entre 2006 et 2009, ainsi que le fait d'être un homme, ont un effet positif sur la probabilité de prolonger le séjour en Suisse après les études. L'effet significatif des variables au seuil de 5% sur la probabilité de prolonger le séjour en Suisse après les études est déterminé par la p-valeur associée à chacun des coefficients estimés du modèle.

En effet, une variable est significative, si la p-valeur du test associé à cette variable est inférieure au seuil de 5%. Nous effectuons ici des tests de Wald. Dans notre cas, les coefficients associés aux variables universités

Tableau 15.3 – Variables du modèle

Variables	Coeff.	Erreur standard	Test de Wald	P-valeur	Exp(Coeff.)
Université de Fribourg	-0.316	0.091	11.957	0.001	0.729
Université de Genève	-1.055	0.076	194.335	0.001	0.348
Université de Lausanne	-0.796	0.109	53.847	0.001	0.451
Université de Neuchâtel	-0.573	0.142	16.264	0.001	0.564
Université de Zurich	-0.736	0.109	45.266	0.001	0.479
Université de Bâle	-0.111	0.078	2.011	0.156	0.895
HES	-0.807	0.119	46.266	0.001	0.446
EPFZ	-1.723	0.130	176.619	0.001	0.178
EPFL	-0.875	0.108	66.077	0.001	0.417
age	0.054	0.001	1688.545	0.001	1.055
sexe	0.131	0.02	43.473	0.001	1.140
mariage	0.987	0.02	2376.557	0.001	2.684
Etudiants Européens, Non-UE	1.664	0.274	36.839	0.001	5.279
Etudiants Américains	1.467	0.276	28.206	0.001	4.335
Etudiants Africains	1.504	0.277	29.557	0.001	4.499
Etudiants Asiatiques	1.408	0.276	25.935	0.001	4.087
duree	-0.135	0.008	289.889	0.001	0.873
coho1	0.043	0.077	0.310	0.578	1.044
coho2	0.161	0.043	14.046	0.001	1.175
Constante	-3.908	0.294	176.420	0.001	0.020

fréquentées ont toutes des p-valeurs inférieures à 0.05 sauf pour l'Université de Bâle et des odds ratios inférieurs à 1. Cela veut dire que les hautes écoles fréquentées n'augmentent pas la probabilité de prolonger le séjour en Suisse après les études par rapport au fait de fréquenter l'Université de Berne et l'effet de ces variables est significatif.

Nous observons un effet positif et significatif du continent d'origine des étudiants par rapport au fait de venir d'un pays de l'Union Européenne. En effet, venir d'un pays non membre de l'Union Européenne par rapport au fait de venir d'un pays de l'Union Européenne, toutes choses étant égales par ailleurs, la chance de prolonger le séjour en Suisse après les études est environ 5.279 fois plus susceptible de se produire. Cette chance est environ quatre fois plus susceptible de se produire pour les autres continents par rapport au fait de venir d'un pays de l'Union Européenne. Les variables âge et sexe ont un effet positif sur la chance de prolonger le séjour en Suisse après les études, car lorsque l'âge augmente d'une unité, la chance de prolonger le séjour en Suisse après les études augmente de 5.5% et, être un homme par rapport à être une femme, augmente cette chance de prolonger le séjour après les études de 14%. Lorsque la durée de séjour augmente d'une unité, la chance de prolonger le séjour après les études diminue de 12.7%, toutes choses étant égales par ailleurs. On observe aussi un effet de cohorte car les étudiants entrés en Suisse entre 1997 et 2000 et ceux entrés entre 2001 et 2005 ont respectivement 1.044 fois et 1.175 fois plus de chance de prolonger le séjour en Suisse après les études, par rapport à ceux entrés entre 2006 et 2009. Comme on s'y attendait, la variable mariage a un impact sur la chance de prolonger le séjour en Suisse après les études avec un odds ratio (rapport de cote) significatif de 2.684. Cela implique que la variable mariage a un impact positif et significatif sur le fait de prolonger le séjour en Suisse après les études. Avec cet odds ratio d'environ 2.684, la prolongation du séjour après les études est 2.684 fois plus susceptible de se produire parmi les personnes mariées que parmi celles non mariées, toutes choses étant égales par ailleurs.

Le modèle a donné un pseudo R^2_N de Nagelkerke de 26.6% qui montre une qualité d'ajustement du modèle assez bonne pour des données en sciences sociales. Sur la base de cette analyse on conclut que notre hypothèse de recherche est corroborée.

15.4 Application aux étudiants africains

La deuxième question de recherche ne porte que sur les étudiants africains et est formulée de la manière suivante :

Le fait que les étudiants africains arrivés en Suisse entre 2003 et 2012 restent ou quittent la Suisse après leurs études dépend-il du mariage de ces étudiants en Suisse durant cette période, de l'âge de ces étudiants en arrivant en Suisse, du sexe de ces étudiants, des hautes écoles fréquentées (Unifr, Unige, Unil, Unine, Uniba, UniZH, HES , EPFZ), de la cohorte d'entrées en Suisse (entrées entre 2003 et 2005, entrées entre 2006 et 2008, entrées entre 2009 et 2012) et de la région d'origine de ces étudiants (AO, AE, AN, AC, AA) ?

AO = Afrique de l'Ouest, AE= Afrique de l'Est, AN = Afrique du Nord, AC= Afrique Centrale, AA=Afrique Australe.

Les données qui seront utilisées pour répondre à notre deuxième question de recherche sont présentées dans la partie qui va suivre.

15.5 Données et description des variables

Les variables qui seront utilisées pour répondre à cette question de recherche consacrée aux étudiants africains sont présentées dans le Tableau 15.4. La variable mariage est pour nous la variable la plus importante pour expliquer le fait que les étudiants africains restent ou quittent la Suisse à la fin de leurs études en tenant compte de l'âge et du sexe des étudiants. La variable nationalité a été recodée selon la situation géographique des pays d'origine des étudiants, l'objectif de ce découpage géographique est de voir si les étudiants provenant d'une région donnée ont plus de chance de prolonger le séjour en Suisse après les études que ceux provenant d'autres régions. L'effet de cohorte nous permettra de savoir si les étudiants africains entrés en Suisse durant une période donnée ont plus de chance de prolonger le séjour en Suisse après les études que ceux entrés durant d'autres périodes. Pour cette question de recherche aussi, la variable dépendante n'est pas disponible dans la base de données et sera à construire. La construction de la variable dépendante repose sur des hypothèses similaires à celles de la première question de recherche.

Tableau 15.4 – Description des variables de la deuxième question de recherche

Variables	Description	Nature	Valeurs(Modalités)
séjour	prolonge ou pas	Quali. nominale	1= prolonge, 0 sinon
mariage	Mariée ou pas	Quali. nominale	1= marié, 0 sinon
unifr	Etudes à l'uni de Fribourg	Quali. nominale	1 =unifr, 0 sinon
unige	Etudes à l'uni de Genève	Quali. nominale	1 =unige, 0 sinon
unil	Etudes à l'uni de Lausanne	Quali. nominale	1 =unil, 0 sinon
unine	Etudes à l'uni de Neuchâtel	Quali. nominale	1 =unine, 0 sinon
HES	Etudes dans une HES	Quali. nominale	1 = HES, 0 sinon
EPFZ	Etudes à l'EPFZ	Quali. nominale	1 =EPFZ, 0 sinon
EPFL	Etudes à l'EPFL	Quali. nominale	1 =EPFL, 0 sinon
uniZH	Etudes à l'uni de Zurich	Quali. nominale	1 =uniZH, 0 sinon
uniba	Etudes à l'uni de Bâle	Quali. nominale	1 =uniba, 0 sinon
age	Age de la personne en années	Quantitative	
sexe	Sexe de la personne	Quali. nominale	1 = homme, 0 = femme
coho1	Entrées en Suisse entre 2003 et 2005	Quali. nominale	1=2003 à 2006, 0 sinon
coho2	Entrées en Suisse entre 2006 et 2008	Quali. nominale	1=2007 à 2009, 0 sinon
coho3	Entrées en Suisse entre 2009 et 2012	Quali. nominale	1=2010 à 2012, 0 sinon
AO	Vient de l'Afrique de l'Ouest	Quali. nominale	1 si AO, 0 sinon
AE	Vient de l'Afrique de l'Est	Quali. nominale	1 si AE, 0 sinon
AN	Vient de l'Afrique du Nord	Quali. nominale	1 si AN, 0 sinon
AC	Vient de l'Afrique Centrale	Quali. nominale	1 si AC, 0 sinon
AA	Vient de l'Afrique Australe	Quali. nominale	1 si AA, 0 sinon
duree	Durée de séjour en année	Quantitative	

Les catégories de référence sont les suivantes :

- Pour la variable mariage, la catégorie de référence qui prend la modalité 0 est non marié. On comparera ainsi le fait d'être marié par rapport à ne pas l'être.
- Pour la variable sexe, la catégorie de référence est « femme ».

— Pour la variable cohorte, la catégorie de référence est la cohorte 3, c'est-à-dire les étudiants arrivés en Suisse entre 2009 et 2012.
— La catégorie de référence pour les universités est l'Université de Berne.
— Pour les variables de région d'origine des étudiants, la catégorie de référence est les étudiants ressortissants de l'Afrique Australe. On comparera ainsi le fait de venir d'une région d'Afrique donnée par rapport au fait de venir de la région de l'Afrique Australe.

15.6 Interprétation des résultats

Le modèle utilisé ici est également le modèle logit à temps discret. Le Tableau 15.5 contient les coefficients estimés ainsi que le test statistique associé à chacun des paramètres. Ce tableau montre un effet négatif sur la chance de prolonger le séjour en Suisse après les études pour les étudiants originaires des régions d'Afrique de l'Est et Centrale par rapport aux étudiants venant de l'Afrique Australe. En effet, par rapport à être originaire d'Afrique Australe, quand on est originaire de l'Afrique Centrale, la chance de prolonger le séjour en Suisse après les études diminue d'environ 38.5%, et lorsqu'on est originaire d'Afrique de l'Est, cette chance diminue d'environ 57.9%, toutes choses étant égales par ailleurs. On observe un effet non significatif pour les étudiants d'Afrique du Nord et d'Afrique de l'Ouest. Les Hautes Ecoles fréquentées, excepté l'Université de Zurich exercent aussi un effet négatif sur les chances de prolonger le séjour en Suisse après les études par rapport à fréquenter l'Université de Berne ; cet effet est significatif seulement pour les Universités de Genève, de Neuchâtel et l'EPFL avec des odds ratio respectifs de 0.635, 0.518 et 0.334. L'EPFZ avec un odds ratio de 0.001, n'exerce aucune influence sur les chances de prolonger le séjour en Suisse après les études par rapport à fréquenter l'Université de Berne. Lorsqu'on fréquente l'Université de Zurich par rapport à fréquenter l'Université de Berne, la chance de prolonger le séjour en Suisse après les études est environ 11 fois plus susceptible de se produire toutes choses étant égales par ailleurs. On observe aussi un effet positif du fait d'être une homme par rapport à être une femme sur la chance de prolonger le séjour en Suisse après les études alors que l'âge exerce un effet négatif sur cette chance. La prolongation du séjour après les études est environ 1.585 fois plus susceptible de se produire pour les hommes par rapport aux femmes, alors que lorsque l'âge augmente d'une unité, toutes choses étant égales par ailleurs, cette chance diminue d'envion 1.1%. On observe aussi un effet positif de la cohorte d'entrées sur la chance de prolonger le séjour en Suisse après les études. En effet, les cohortes d'étudiants africains entrées en Suisse entre 2003 et 2006 et celles entrées entre 2007 et 2009 ont respectivement 29% et 5.9% plus de chance de prolonger le séjour en Suisse après les études par rapport à la cohorte entrées en Suisse entre 2010 et 2012, toutes choses étant égales par ailleurs. Comme on s'y attendait, la variable explicative la plus importante de notre modèle est le mariage dans la mesure où cette dernière exerce un impact très élevé sur la chance de prolonger le séjour en Suisse après les études. La variable mariage a un effet positif sur le fait de prolonger le séjour en Suisse après les études pour les étudiants africains. Avec un odds ratio (rapport de cote) d'environ 176, la prolongation de séjour en Suisse après les études est 176 fois plus susceptible de se produire parmi les étudiants mariés que parmi ceux non mariés, toutes les autres caractéristiques individuelles restant les mêmes. Le tableau récapitulatif des modèles nous donne un R^2_N de Nagelkerke de 54.8%, qui montre une très bonne qualité d'ajustement du modèle. On a observé un impact très élevé du mariage sur la chance de prolonger le séjour en Suisse après les études, ce qui nous permet de dire que notre question de recherche est corroborée.

En observant de manière critique ces résultats, on remarque que le rapport de cote pour l'EPFZ est particulièrement bas. Ce qui nous a conduit à replonger dans la base de données à la recherche d'une explication de ce résultat. On a réalisé que la faiblesse du rapport de cote pour l'EPFZ qui est une institution d'enseignement supérieur de renommée internationale est liée au fait que l'effectif des étudiants africains dans cette institution d'enseignement supérieur est très faible. En effet, la base de données ayant servi aux analyses pour la deuxième question de recherche a été extraite en suivant la définition des étudiants internationaux selon l'OCDE. Cela a eu pour conséquence d'exclure tous les étudiants d'origine africaine qui : ont obtenu leur maturité en Suisse, sont nés en Suisse ou qui ont antérieurement résidé en Suisse (cet ensemble rentre dans la catégorie des étudiants étrangers alors que dans cette recherche on s'intéresse aux étudiants internationaux). En suivant la définition de l'OCDE, on diminue également l'effectif des étudiants africains en Suisse de manière générale, cela affecte

aussi ceux inscrits à l'EPFZ. C'est ce qui fait que sur 1190 étudiants internationaux africains inscrits dans les hautes écoles en Suisse, seulement 19 sont venus de l'Afrique pour faire des études à l'EPFZ, ce qui représente une part de 1.59% de l'effectif total des étudiants internationaux africains en Suisse.

Tableau 15.5 – Variables du modèle

Variables	Coeff.	Erreur standard.	Test de Wald	P-valeur	EXP(Coeff.)
AN	0.045	0.157	0.082	0.775	1.046
AE	-0.864	0.215	16.099	0.001	0.421
AO	0.238	0.156	2.307	0.129	1.268
AC	-0.486	0.216	5.064	0.024	0.615
Université de Bâle	-0.490	0.355	1.910	0.167	0.613
Université de Fribourg	-0.374	0.242	2.383	0.123	0.688
Université de Genève	-0.455	0.168	7.309	0.007	0.635
Université de Lausanne	-0.017	0.178	0.009	0.923	0.983
Université de Neuchâtel	-0.658	0.217	9.240	0.002	0.518
Université de Zurich	2.406	0.770	9.752	0.002	11.089
HES	-0.680	0.369	3.392	0.065	0.507
EPFZ	-19.339	3861.761	0.001	0.996	0.001
EPFL	-1.097	0.350	9.839	0.002	0.334
sexe	0.460	0.088	27.478	0.001	1.585
age	-0.021	0.08	6.752	0.009	0.979
duree	-0.020	0.011	3.575	0.059	0.980
mariage	5.172	0.190	741.321	0.001	176.278
coho1	0.254	0.109	5.448	0.020	1.290
coho2	0.722	0.10	51.947	0.001	2.059
Constante	-4.985	0.318	245.729	0.001	0.007

15.7 Discussion

Au cours de ces dernières années, nous avons assisté à une augmentation rapide du nombre d'étudiants en mobilité à travers le monde. Les étudiants africains, de par leur volonté de se former dans des universités renommées du monde entier et pour pallier les insuffisances des universités locales, ne sont pas en marge de cette mobilité grandissante et participent de manière active à cette dernière. L'effectif des étudiants africains en formation dans les hautes écoles suisses reste cependant modeste et est inégalement réparti selon l'origine géographique. L'Afrique du Nord et le Cameroun se partagent plus de la moitié de l'effectif des étudiants africains en Suisse. Les étudiants africains en Suisse sont majoritairement des hommes. Ils sont en moyenne plus âgés que les étudiants suisses et sont souvent mariés. Ils sont nombreux à passer leur bachelor dans les HES et le master dans les universités et les EPF ; ils s'orientent généralement vers les branches techniques, économiques, informatiques qui sont des branches dans lesquelles les universités locales sont défaillantes. Les Universités de Genève, de Lausanne et l'EPFL sont les hautes écoles qui accueillent le plus grand nombre d'étudiants africains en Suisse. Par comparaison aux étudiants africains, les étudiants suisses et européens sont nombreux à s'inscrire dans les sciences humaines, sociales et les sciences économiques. Le rythme d'augmentation de l'effectif des étudiants africains en Suisse reste cependant plus lent que pour les étudiants européens et les autres étudiants internationaux. Nous pouvons retenir que contrairement à la tendance qui se manifeste en Europe, l'effectif des étudiants africains en Suisse a considérablement augmenté ces dernières années et que cet effectif a été multiplié par presque quatre entre 1997 et 2014. Le nombre de diplômes de bachelor et de master décernés à des étudiants africains par des institutions d'enseignement supérieur suisses a aussi considérablement augmenté, les universités romandes étant celles qui accueillent le plus d'étudiants africains. Quant au doctorat, l'année 2011 est celle où il y a eu le plus d'étudiants africains inscrits au doctorat en Suisse, alors que les années 2006 et 2007, sont celles qui ont enregistré les plus grands effectifs de diplômés africains au niveau du doctorat. Nous avons complété cette analyse par une étude similaire sur les étudiants suisses au niveau du bachelor, du master et du doctorat. Comme on a pu le remarquer, nous avons assisté également à une augmentation très rapide de l'effectif des étudiants suisses à ces différents niveaux de formation.

L'augmentation rapide du nombre d'étudiants internationaux et africains nous a conduit à étudier le parcours de vie de ces étudiants en Suisse pour répondre à deux questions de recherche dont la première porte sur les étudiants internationaux de manière générale (y compris les étudiants africains) et la deuxième question de recherche ne concernant que les étudiants africains. Nous nous sommes intéressés aux rôles du mariage, de certaines caractéristiques démographiques, académiques et migratoires sur la probabilité de prolonger le séjour en Suisse après les études. Pour la première question de recherche, nous avons pu observer un impact fort et significatif du mariage sur la probabilité de prolonger le séjour en Suisse après les études. En effet, la variable mariage a le plus fort impact sur la probabilité de prolonger le séjour en Suisse après les études avec un odds ratio de 2.687. Nous avons pu observer que fréquenter une autre haute école exerce un effet négatif sur la probabilité de prolonger le séjour en Suisse après les études par rapport à fréquenter l'Université de Berne. Nous observons un effet positif et significatif du continent d'origine des étudiants par rapport au fait de venir d'un pays de l'Union Européenne. Venir d'un pays non membre de l'Union Européenne par rapport à venir d'un pays de l'Union Européenne, toutes choses étant égales par ailleurs, la chance de prolonger le séjour en Suisse après les études est environ 5.279 fois plus susceptible de se produire. Quand on vient d'un autre contient, la chance de prolonger le séjour en Suisse après les études est environ quatre fois plus susceptible de se produire par rapport au fait de venir d'un pays de l'Union Européenne. Ce résultat est quand même surprenant dans la mesure où les ressortissants d'un pays membre de l'UE ou de l'AELE ont plus de liberté et de facilité dans la mobilité en Suisse que les étudiants ressortissants des pays non membres de l'UE ou des Etats tiers. Nos résultats montrent aussi que le fait de fréquenter certaines hautes écoles suisses considérées comme parmi les meilleures du monde (Université de Genève, EPFL, EPFZ) n'offrent pas plus de chance que les hautes écoles les moins connues en ce qui concerne la prolongation du séjour en suisse après les études. Ce qui est encore surprenant est donc le fait que ces hautes écoles universitaires exercent un impact négatif sur les chances de prolonger le séjour en Suisse après les études. On se serait théoriquement attendu à ce que certaines hautes écoles de par leur renommée exercent un impact positif sur les chances de prolonger le séjour après les études par rapport à fréquenter l'Université de Berne.

Les variables sexe et âge ainsi que les cohortes d'entrées en Suisse entre 2001 et 2009 exercent aussi un impact positif sur la probabilité de prolonger le séjour en Suisse après les études. La chance de prolonger le séjour en Suisse après les études augmente de 14% quand on est un homme par rapport à quand on est une femme. La cohorte d'entrées en Suisse entre 1997 et 2000 et celle d'entrées en Suisse entre 2001 et 2005 augmentent la chance de prolonger le séjour en Suisse après les études pour les étudiants internationaux respectivement de 4.4% et 14.5% par rapport à la cohorte d'entrées en Suisse entre 2006 et 2009.

Dans la deuxième question de recherche consacrée aux étudiants africains, nous avons aussi remarqué un impact fort et significatif de la variable mariage sur la probabilité de prolonger le séjour en Suisse après les études. La chance de prolonger le séjour en Suisse après les études est environ 176 fois plus susceptible de se produire pour les étudiants mariés par rapport aux étudiants non mariés, toutes choses étant égales par ailleurs. Nous remarquons aussi un impact négatif des hautes écoles fréquentées sur la probabilité de prolonger le séjour en Suisse après les études par rapport à fréquenter l'Université de Berne sauf pour l'Université de Zurich où on remarque un impact fort et positif. La chance de prolonger le séjour en Suisse après les études est environ 11 fois plus susceptible de se produire si on fréquente l'Université de Zurich par rapport à fréquenter l'Université de Berne, toutes choses étant égales par ailleurs. Toutefois, on doit relativiser ce dernier résultat, car le nombre d'étudiants diplômés de l'Université de Zurich est très faible dans notre analyse. Par rapport à être originaire de l'Afrique Australe, on observe aussi un impact fort et négatif de la région d'origine des étudiants africains sauf pour l'Afrique du Nord (impact pas significatif) où on remarque un impact positif sur les chances de prolonger le séjour en Suisse après les études. Les variables âge et durée de séjour exercent un impact négatif sur la probabilité de prolonger le séjour en Suisse après les études. Les variables sexe et cohortes d'entrées en Suisse, elles, exercent un effet positif sur les chances de prolonger le séjour en Suisse après les études, toutes choses étant égales par ailleurs.

La réponse à nos questions de recherche nous ouvre maintenant la voie sur une discussion générale portant sur tous les éléments qui ont été présentés dans cette thèse. Cette discussion générale est faite dans la cinquième et dernière partie de ce travail.

Cinquième partie

Conclusion générale

Chapitre 16

Conclusion

L'objectif de cette thèse est d'étudier le parcours de vie des étudiants internationaux et africains en Suisse. Nous avons commencé par faire une revue de la littérature sur la mobilité des étudiants internationaux dans le monde et en Suisse dans le but d'effectuer un état des lieux des études qui existent dans cette thématique d'une part, d'autre part, de savoir comment ces études ont abordé cette mobilité étudiante. Les études qui ont été faites ont abordé la mobilité internationale des étudiants sous diverses formes : immersion dans une nouvelle culture, apport culturel pour le pays d'accueil, commerce lucratif pour les pays d'accueil, forme d'immigration déguisée ou fuite des cerveaux, facteurs de rétention dans le pays d'accueil, et apprentissage d'une nouvelle langue etc. Parmi ces études, certaines utilisent des données d'enquêtes (Gaillard, 2002; Ballatore et Blöss, 2008; Findlay et al., 2011), des données de panel (González Rodríguez et al., 2011), des données provenant d'institutions internationales (Flahaux et De Haas, 2016; Unesco, 2012) et d'autres études utilisent des données administratives ; ces études se limitent généralement à des analyses descriptives et n'abordent pas la mobilité des étudiants dans une perspective de parcours de vie.

Le manque d'études réalisées sur le parcours de vie des étudiants internationaux dans le monde de manière générale et en particulier en Suisse avec l'utilisation de données administratives motive notre recherche. Nous avons ensuite introduit nos deux questions de recherches dont la première porte sur les étudiants internationaux de manière générale (y compris les étudiants africains) et la deuxième ne concerne que les étudiants africains. Dans la première question de recherche, nous nous sommes intéressés au rôle du mariage, de certaines caractéristiques démographiques, académiques et migratoires sur la chance de prolonger le séjour en Suisse après les études pour les étudiants internationaux. Dans la deuxième question de recherche, nous nous sommes intéressés aux étudiants africains en étudiant également les facteurs explicatifs du fait que ces étudiants prolongent le séjour en Suisse après les études.

Pour répondre à ces questions de recherche, nous avons eu besoin de modèles statistiques et de données. Nous avons ainsi commencé par une introduction à l'analyse des parcours de vie en prenant des exemples de trajectoires d'étudiants puis, nous avons présenté les différentes fonctions à la base de ces méthodes, avant de développer des exemples littéraux de calcul de ces différentes fonctions. A l'aide de données de sources diverses, nous avons ensuite calculé ces différentes fonctions dans le but de montrer que lorsque nous avons de petites bases de données, il est possible de faire ces différents calculs à la main sans avoir besoin d'un ordinateur. L'objectif de ces calculs littéraux et de ces exemples développés avec des données est de procéder à une vulgarisation de ces différentes fonctions, car ce sont elles qui sont à la base des différents modèles qui ont été présentés dans cette thèse.

Nous nous sommes ensuite intéressés aux modèles statistiques les plus couramment utilisés dans l'analyse des parcours de vie. Ces modèles se subdivisent en trois groupes de modèles qui sont : les modèles non paramétriques, les modèles paramétriques et le modèle semi-paramétrique de Cox. Dans les modèles non paramétriques, les méthodes de Kaplan-Meier et actuarielle ont été présentées avec des exemples d'application. Nous avons également montré qu'avec ces méthodes, on peut calculer les différents estimateurs (l'estimateur de Kaplan-Meier et l'estimateur actuariel) à la main lorsque nous avons de petites bases de données et nous avons ensuite effectué des analyses avec un logiciel dans le cas de grandes bases de données. Pour finir, nous avons

comparé la méthode de Kaplan-Meier et la méthode actuarielle en dégageant les avantages et les inconvénients de l'une par rapport à l'autre. Il ressort de cette comparaison qu'il n'y a aucune considération technique qui pousserait à privilégier l'une des méthodes par rapport à l'autre.

Cependant, on privilégiera la méthode de Kaplan-Meier lorsque la taille de l'échantillon est petite et lorsque l'instant exact auquel les événements ont lieu est connu avec précision ainsi que le rythme d'occurrence de l'événement étudié.

On privilégiera la méthode actuarielle lorsque nous avons des échantillons de grandes tailles, lorsque la période d'observation est longue et lorsqu'on ne connaît pas l'instant exact auquel les événements ont lieu. Les méthodes de Kaplan-Meier et actuarielle étant des méthodes descriptives, elles ne permettent pas de mesurer l'impact d'une ou de plusieurs variables sur le risque d'occurrence d'un phénomène donné ; nous nous sommes alors penchés sur les modèles paramétriques qui sont des modèles de régression.

Dans les modèles de régression paramétriques, nous avons pu observer que les caractéristiques individuelles peuvent agir soit sur la fonction de risque soit sur la fonction de séjour. Lorsque les caractéristiques individuelles agissent sur la fonction de séjour, on parle de modèle à temps de sorties accélérées (AFT) et lorsqu'elles agissent sur la fonction de risque, on parle de modèles à risques proportionnels (HP). Nous avons aussi montré que pour les modèles à risques proportionnels, la « variable dépendante » est le risque, les caractéristiques individuelles agissent donc sur cette dernière. Ces modèles permettent ainsi de répondre à des questions de recherche qui s'intéressent à l'impact des variables explicatives sur le risque d'occurrence d'un événement quelconque au cours du temps. Dans les modèles à temps de sorties accélérées, la variable dépendante est la durée de séjour ; les variables explicatives agissent donc sur cette dernière. Les modèles à temps de sorties accélérées permettent donc de répondre à des questions de recherche qui s'intéressent à l'impact d'une ou de plusieurs variables sur la durée de séjour.

Nous avons ensuite introduit le modèle semi-paramétrique de Cox qui est aussi un modèle à risques proportionnels ; ce modèle combine l'approche paramétrique et l'approche non paramétrique. L'avantage de cette méthode par rapport aux autres réside dans le fait qu'aucune hypothèse n'est faite en ce qui concerne la distribution du risque au cours du temps. Dans ce modèle aussi, la « variable dépendante » est le risque. Ce modèle permet ainsi de répondre à des questions de recherche qui s'intéressent au risque d'occurrence d'un événement donné au cours du temps à l'aide d'une ou de plusieurs variables explicatives. A l'image des modèles paramétriques à hasard proportionnel, dans le modèle de Cox aussi, le risque n'est pas une caractéristique observée. Un exemple d'application du modèle de Cox a aussi été développé et on a pu remarquer, à travers les résultats, que le modèle de Cox est un modèle sans constante. Les modèles paramétriques et le modèle de Cox étant sensibles aux groupes d'égalité, nous avons introduit un autre groupe de modèles qui s'appelle les modèles de régressions logistiques à temps discret. Parmi ces modèles, le modèle logit à temps discret est le plus adapté lorsque nous avons des données annuelles récoltées à des périodes fixes de l'année, car il n'est pas sensible aux groupes d'égalités.

Une remarque importante que l'on peut faire est que la largeur de la durée d'observation des individus soumis à observation peut avoir un impact sur le modèle statistique à choisir. Les méthodes statistiques seront différentes selon l'étendue du temps durant lequel les événements sont mesurés (observations journalières, hebdomadaires, mensuelles, semestrielles ou annuelles) (Ray, 1988). Dans les modèles d'analyse de survie, lorsque la durée d'observation est courte et que les événements observés peuvent survenir à tout moment avec un risque faible que deux individus ou plus expérimentent simultanément le phénomène étudié, on parle de temps continu. Lorsque la période d'observation est longue et que le risque d'occurrence simultanée de l'événement étudié est élevé, on parlera de temps discret et les modèles en temps discret seront les plus adaptés pour analyser le phénomène étudié.

En plus des méthodes statistiques, nous avons aussi besoin de données pour mettre en application ces méthodes. Nous avons ainsi commencé par présenter les différentes sources de données longitudinales dans les sciences sociales en présentant les avantages et les inconvénients de chaque type de données. Nous avons aussi mis en évidence qu'à l'image des autres disciplines, dans les sciences sociales aussi, les données longitudinales peuvent être récoltées de toutes les manières possibles (rétrospective, prospective et administrative).

Les données administratives de l'OFS qui sont celles utilisées dans cette thèse ont ensuite été présentées ainsi

que les différents travaux de mise en commun de ces différentes bases de données. Les avantages et les inconvénients des données administratives ont aussi été présentés ainsi que les différents types de variables contenues dans les bases de données de l'OFS. Pour la mise en commun des différentes bases de données, nous avons mis en évidence la nécessité d'avoir un identifiant commun (clé unique) dans les différentes bases de données. Dans les données de l'OFS, cet identifiant est représenté par le numéro AVS anonymisé. Nous avons également ressorti l'importance de l'anonymisation pour la protection des données. Une fois la mise en commun des bases de données effectuée, nous avons expliqué en détail la procédure d'extraction de la base de données contenant les étudiants internationaux et les étudiants africains en utilisant les définitions de l'Unesco, de l'OCDE et de l'EUROSTAT.

Nous avons aussi pu mettre en évidence les conséquences de la protection des données pour la recherche. En effet, nous n'avons pas pu mener cette recherche à la hauteur de nos ambitions car certaines variables importantes n'ont pas été livrées par l'OFS au nom de la protection des données. L'idée de départ de cette recherche était d'étudier le parcours de vie des étudiants internationaux et africains en s'intéressant à la chronologie de certains événements dans leur parcours de vie. Notre objectif était de connaître la chronologie des événements suivants dans le parcours de vie des étudiants internationaux en Suisse : obtention du diplôme, mariage, naissance des enfants, premier emploi puis changement de permis. Nous étions intéressés à l'ordre dans lequel ces événements apparaissaient dans le parcours de vie des étudiants internationaux et africains en Suisse. Mais en raison de la protection des données, les variables « date de naissance des enfants » et la « date de mariage des étudiants » n'ont pas été livrées par l'OFS nous poussant ainsi à renoncer à des variables centrales pour nos questions de recherche de départ. En renonçant à ces variables, nous renonçons également à une méthode importante d'analyse des parcours de vie : l'analyse de séquences. En effet, cette méthode qui est à vocation descriptive offre l'avantage de visualiser graphiquement l'ordre dans lequel les différents événements apparaissent dans le parcours de vie des étudiants. Nous avons renoncé à cette méthode parce que nous n'avons finalement que très peu d'événements du parcours de vie de ces étudiants, la variable « emploi » n'étant en soi pas disponible dans la base de données, nous n'avons plus qu'un seul événement, l'année d'obtention du diplôme (BA, MA et doctorat).

Nous avons aussi pu tester les potentialités d'application de ces différentes méthodes à nos données administratives, ce qui nous a permis encore une fois de faire ressortir l'immense potentiel qu'offrent les données administratives comme source de données longitudinales. En appliquant les méthodes non paramétriques à nos données administratives, nous avons pu observer que la méthode actuarielle est à préférer à la méthode de Kaplan-Meier en tenant compte de la taille importante de notre base de données. En effet, la méthode actuarielle offre la possibilité de grouper les données par intervalle de temps, ce qui nous permet d'obtenir des graphiques plus lisibles et des tables de survie moins surchargées.

Nous avons aussi montré que les méthodes paramétriques peuvent s'appliquer à nos données, mais ces méthodes présentent le désavantage d'être sensibles aux groupes d'égalités. La sensibilité de ces méthodes aux groupes d'égalités a pour conséquence de conduire à l'estimation de paramètres biaisés, ce qui a conduit à dire que ces méthodes ne conviennent pas à des données annuelles récoltées à des périodes fixes de l'année. N'eût été la sensibilité de ces modèles aux groupes d'égalités, il n'y a aucun autre problème technique qui empêcherait l'application de ces méthodes à nos données. Le modèle semi paramétrique de Cox, qui est un modèle à hasard proportionnel, est une bonne alternative aux modèles paramétriques, mais il présente aussi l'inconvénient d'être sensible aux groupes d'égalités. Ceci nous a poussés à nous orienter vers le modèle logit à temps discret qui est adapté aux données présentant des groupes d'égalités et qui sont récoltées à des périodes fixes de l'année. Nos données ont un risque important d'avoir des groupes d'égalités d'une part, d'autre part, en tenant compte de leur caractère annuel, nous avons privilégié le modèle de régression logit à temps discret pour répondre à nos questions de recherche. Mais auparavant, nous avons repris la partie descriptive de notre article publié dans la Revue Géo-Regards (Barry, 2017) qui décrit de manière complète nos données. Cette analyse descriptive nous a permis de connaître l'évolution de l'effectif des étudiants internationaux et africains en Suisse. Nous avons complété cette analyse descriptive par une analyse similaire sur les étudiants suisses dans les hautes écoles suisses au niveau du bachelor, du master et du doctorat. Cette analyse complémentaire a permis de montrer l'évolution très rapide de l'effectif des étudiants suisses pour ces différents niveaux de formation. Nous avons aussi pu observer que malgré le faible effectif des étudiants africains en Suisse, cet effectif a continué d'augmenter ces dernières années. Cet article a également mis en évidence la distribution très inégalitaire en

Suisse de la répartition des étudiants africains par nationalité.

En effet, l'Afrique du Nord représente à elle seule près de la moitié de l'effectif des étudiants africains en Suisse. En additionnant les effectifs des étudiants d'Afrique du Nord et d'Afrique de l'Ouest, on obtient 80% de l'effectif total des étudiants africains en Suisse. Ces étudiants sont majoritairement concentrés dans les cantons romands ; en Suisse alémanique, c'est l'Université de Bâle qui accueille le plus grand effectif d'étudiants africains. Nous terminons ensuite par présenter nos deux questions de recherche, puis par répondre à ces dernières.

La première question de recherche porte sur le parcours de vie des étudiants internationaux (y compris les Africains) en Suisse. Nous arrivons à la conclusion que le mariage est une variable qui exerce un effet positif et significatif sur la probabilité de prolonger le séjour en Suisse après les études par rapport au fait de ne pas être marié. Pour cette question de recherche, nous arrivons à la conclusion que la chance de prolonger le séjour en Suisse est environ 2.687 fois plus élevée lorsqu'on est marié par rapport à lorsqu'on n'est pas marié, toutes choses étant égales par ailleurs. Nous avons pu observer que les hautes écoles fréquentées et la nationalité exercent un effet négatif sur la probabilité de prolonger le séjour en Suisse après les études par rapport à fréquenter l'Université de Berne ou d'être originaire d'un pays de l'UE. Les variables sexe et âge ainsi que les cohortes entrées en Suisse entre 2001 et 2009 exercent aussi un impact positif sur la probabilité de prolonger le séjour en Suisse après les études.

Dans la deuxième question de recherche, nous nous sommes intéressés aux parcours de vie des étudiants africains en Suisse. Pour cette question de recherche aussi, nous avons montré qu'il y a un impact très fort et significatif du mariage sur la chance de prolonger le séjour en Suisse après les études. En effet, avec un rapport de cote d'environ 176, la chance de prolonger le séjour en Suisse après les études est 176 fois plus susceptible de se produire parmis les étudiants mariés que parmis les étudiants non mariés, toutes choses étant égales par ailleurs. Nous remarquons aussi un impact négatif des hautes écoles fréquentées sur la probabilité de prolonger le séjour en Suisse après les études sauf pour l'Université de Zurich où l'on remarque un impact fort et élevé par rapport à fréquenter l'Université de Berne. La prolongation du séjour en Suisse après les études est environ 11 fois plus susceptible de se produire si l'on a fréquenté l'Université de Zurich par rapport à si l'on a fréquenté l'Université de Berne. On observe aussi un impact fort et négatif de la région d'origine des étudiants africains, sauf pour l'Afrique du Nord où l'on remarque un impact positif sur les chances de prolonger le séjour en Suisse après les études par rapport aux étudiants venant de l'Afrique Australe. Les variables âge et durée de séjour exercent un impact négatif sur la probabilité de prolonger le séjour en Suisse après les études. Les variables sexe et cohortes d'entrées exercent un effet positif sur les chances de prolonger le séjour en Suisse après les études. Les résultats des analyses statistiques nous ont conduits à mettre en évidence l'importance du mariage sur la probabilité de prolonger le séjour en Suisse après les études pour les étudiants internationaux et pour les étudiants africains par rapport au fait de ne pas être marié et par conséquent à corroborer nos deux hypothèses de recherche.

Le fait de suivre à la lettre la définition des étudiants internationaux selon l'Unesco, l'OCDE et l'EUROSTAT a eu pour conséquence de retirer de la base de données tous les étudiants africains qui : sont nés en Suisse, ont suivi une formation antérieure en Suisse ou qui ont résidé antérieurement en Suisse. Nous constatons ainsi que les étudiants d'origine africaine qui sont nés en Suisse, qui ont suivi toute leur scolarité en Suisse, et qui peuvent n'avoir jamais connu de migration ne sont comptabilisés ni dans la catégorie des étudiants internationaux, parce qu'ils ne sont pas venus en Suisse pour faire des études, ni dans la catégorie des étudiants suisses parce qu'il n'ont pas la nationalité suisse. Nous jugeons dès lors qu'il serait intéressant de comptabiliser ces étudiants dans la catégorie des étudiants suisses dans la mesure où d'un point de vue académique, ces étudiants ont exactement le même parcours que les étudiants suisses, mais un parcours académique totalement différent de celui des étudiants africains, par exemple, qui ne sont venus que pour faire des études universitaires en Suisse. Exclure ces étudiants sous le prétexte qu'ils sont nés en Suisse constituerait un biais de sélection. Remédier à ce problème impliquerait de revoir le critère du pays de naissance dans la définition des étudiants étrangers.

Il serait dès lors intéressant de voir l'impact sur les paramètres estimés du fait de ne pas exclure tous ces étudiants étrangers de la base de données. En effet, nous avons pu montrer dans cette thèse que malgré le fait que l'effectif des étudiants africains en Suisse soit faible, il a beaucoup augmenté ces dernières années. Le fait d'extraire de la base de données les étudiants étrangers a eu pour conséquence de réduire la taille de

l'échantillon des étudiants africains et comme on a pu le voir a eu un impact sur le rapport de cote obtenu pour les étudiants africains de l'EPFZ.

Les méthodes statistiques présentées dans cette thèse proviennent du domaine de la recherche médicale et sont largement diffusées dans ce secteur. La transposition de ces méthodes dans le domaine des sciences sociales dans le but d'analyser le parcours de vie des étudiants internationaux n'est pas une chose aisée à faire. En effet, quand on parle de risque de décès suite à une implantation cardiaque, cela fait naturellement du sens. Mais quand on parle de « risque » d'obtenir un diplôme de bachelor après un certain temps depuis le début des études de bachelor, cela crée immédiatement des confusions. Pourtant, d'un point de vue statistique, ces deux événements se mesurent de la même manière. Dans le cas de l'implantation cardiaque, on va juste observer le nombre de patients décédés (le premier transplanté cardiaque a survécu 73 jours) après l'implantation cardiaque entre la date de l'opération et une date quelconque dans le futur. La durée d'observation est généralement exprimée en jours, semaines ou mois.

Dans le cas de l'obtention du diplôme de bachelor, on va juste observer le nombre de diplômes de bachelor obtenus entre le début des études de bachelor et une durée supérieure ou égale à trois ans. En comparant les deux durées d'observation, on réalise que la durée d'observation est plus courte dans le cas de l'implantation cardiaque que dans le cas des études de bachelor qui vont durer au moins trois ans. Cela entraîne déjà un premier problème dans l'utilisation et dans la transposition de ces méthodes dans le domaine de la mobilité des étudiants. Le fait que la durée d'observation pour les études de bachelor soit large pose donc problème et le fait que plusieurs étudiants obtiendront le diplôme de bachelor la même année entraîne aussi un autre problème : celui des groupes d'égalité qui ne conviennent pas aux modèles paramétriques et au modèles de Cox. En effet, les méthodes non paramétriques n'ont pas les mêmes objectifs que les méthodes paramétriques et ne fonctionnent pas de la même manière. A l'intérieur même des modèles paramétriques, nous avons deux groupes de modèles qui sont différents à travers le lien entre les variables explicatives et la variable dépendante. Dans les modèles à hasard proportionnel, la « variable dépendante » est le risque, mais cette fonction de risque est différente d'un modèle à un autre, même si nous avons toujours la possibilité de retrouver un ou plusieurs modèles à partir des autres modèles. Dans les modèles à hasard proportionnel, le risque n'est pas une caractéristique observée malgré le fait que ce dernier puisse s'écrire comme une combinaison linéaire des variables explicatives.

Dans les modèles à temps de sorties accélérées, la variable dépendante est la durée de séjour ; cette durée de séjour aussi s'écrit différemment d'un modèle à temps de sorties accélérées à un autre, ce qui rend difficile la comparaison de ces modèles. Nous avons également pu observer que la manière dont un modèle est spécifié peut avoir des effets inverses sur la variable dépendante et ainsi impacter les résultats obtenus par certains modèles. En effet, pour le modèle exponentiel, on passe d'un modèle à hasard proportionnel à un modèle à temps de sorties accélérées, en changeant simplement le signe des coefficients obtenus dans les modèles à hasard proportionnel et vice versa. Cela veut dire qu'une variable qui a un impact positif sur la « variable dépendante » dans un modèle à hasard proportionnel aura un impact négatif sur celle-ci dans un modèle à temps de sorties accélérées. Pour la régression de Weibull, on passe d'un modèle à hasard proportionnel à un modèle à temps de sorties accélérées en multipliant les coefficients obtenus dans le modèle à hasard proportionnel par $-1/p$, p représentant le paramètre de forme de la loi de Weibull. On peut passer d'un modèle à AFT à un modèle à HP en multipliant les coefficients estimés dans le modèle à AFT par $-p$, p étant toujours le paramètre de forme de la distribution de Weibull.

La complexité de ces méthodes réside dans les expressions mathématiques des fonctions à la base de certains modèles : cette complexité ressort souvent dans la littérature pour les distributions gamma et log logistique par exemple. L'approche graphique s'est avérée être très utile dans la comparaison, car selon l'allure de la courbe de la fonction de risque, on peut savoir quel est le modèle le plus approprié pour analyser les données à disposition. Une autre critique qu'on peut faire à l'égard des modèles paramétriques est l'hypothèse très forte qui est faite en ce qui concerne la forme de la distribution du risque d'occurrence de l'événement étudié au cours du temps. Si l'on fait par exemple l'hypothèse que le risque est distribué de manière monotone au cours du temps, alors le meilleur modèle pour analyser les données est le modèle de Weibull. Si cela est vrai, alors aucun autre modèle ne peut mieux estimer les paramètres que le modèle de Weibull ; si ce ne pas le cas, on estimera des paramètres biaisés.

Nous avons également abordé la question du choix d'un modèle statistique parmi les nombreux modèles pré-

sentés ; ce choix peut se faire en tenant compte du fait que les modèles soient imbriqués ou pas. Si les modèles sont imbriqués, le choix du meilleur modèle est simple et se fera par simples tests statistiques sur les paramètres estimés. Si les modèles ne sont pas imbriqués, nous utiliserons le critère AIC pour choisir le meilleur modèle : dans ce cas, le meilleur modèle sera celui qui aura la plus petite valeur du critère AIC. L'avantage de ce critère est qu'il permet de comparer toute une série de modèles différents. On peut ainsi comparer à la fois tous les modèles paramétriques (modèles à risque proportionnel et modèles à temps de sorties accélérées) et le modèle de Cox.

Cependant, la recherche du meilleur modèle dans le but d'analyser nos données a engendré la comparaison de nombreux modèles avant de trouver le modèle statistique adapté à nos données. On peut ainsi se poser la question suivante : pourquoi ne pas prendre directement le modèle adapté et répondre à nos questions de recherche ?

La réponse à cette question réside dans les objectifs du projet de thèse qui était de comparer tous les modèles habituellement utilisés dans l'analyse de parcours de vie. Ces modèles devaient être présentés et discutés même s'ils ne s'appliquent pas à nos données en mentionnant pourquoi ils ne s'appliquent pas avant de passer aux modèles suivants. Cela explique aussi le fait que la partie consacrée aux modèles soit plus longue que toutes les autres parties de la thèse.

Il faut aussi rappeler que la non livraison de certaines variables très importantes nous a conduit à modifier le projet de départ qui était celui de procéder d'abord à une analyse descriptive des données à travers la visualisation des trajectoires des étudiants avec l'analyse de séquence. Cette méhode statistique nous aurait permis de connaître la chronologie de certains événements du parcours de vie de ces étudiants avant d'aborder les modèles qui ont été présentés et vulgarisés dans cette thèse.

Les résultats obtenus après les différents travaux de vulgarisation nous offrent une entière satisfaction dans la mesure où tous ces modèles ont été présentés avec des exemples d'application, ce qui est indispensable dans la compréhension et la mise en pratique de ces méthodes.

La complexité de ce travail ne réside pas seulement au niveau des méthodes, elle réside aussi au niveau des données qui ont été mobilisées pour réaliser cette thèse. Ces données sont administratives et à la base, elles ont pour objectif d'aider les autorités à savoir qui est en Suisse et qui fait quoi en Suisse. Ces données administratives ont été mobilisées dans le but d'analyser les parcours de vie des étudiants internationaux et africains en Suisse. Ces données n'étant donc à la base pas destinées à la recherche, on pouvait s'attendre à beaucoup de surprises dans le sens où elles peuvent tout contenir, sauf « l'essentiel » ou contenir des informations surprenantes. Nous avons pu remarquer dans le chapitre 13, que la variable permis de séjour contient des types de permis que l'on n'a pas l'habitude de voir des étudiants détenir. Les données ont montré qu'il y a des étudiants qui détiennent des permis saisonniers (statut supprimé en 2002) ou des permis frontaliers (permis réservés aux travailleurs frontaliers). Ceci a pour conséquence qu'une analyse explorartoire des données est absolument essentielle avant de passer aux analyses pour être sûr que les données soient cohérentes avec les objectifs de la recherche.

La complexité de certains parcours de vie fait qu'un simple registre ne permet pas d'apporter des explications à certaines interrogations que l'on peut avoir sur des parcours de vie complexes. Nous parlerons ici des étudiants qui étudient en Suisse sans titre de séjour valable ; les étudiants qui, à un moment donné ont obtenu un titre de séjour et qui l'ont perdu pour diverses raisons (échec définitif, changement du plan d'études initial, changement de canton de résidence) et qui continuent les études en Suisse ; des étudiants qui arrivent en Suisse et qui accomplissent ou pas une première formation et qui quittent la Suisse pour y revenir après un certain temps de séjour à l'étranger ; ou ces étudiants qui basculent simplement dans la clandestinité sans être inscrits dans une institution d'enseignement supérieur en Suisse. Pour cette dernière catégorie d'étudiants, la dernière information que nous aurons sur ces étudiants en Suisse est l'année du dernier titre de séjour valable. Ces étudiants disparaîtront complètement des registres ZEMIS et LABB.

Les étudiants qui suivent une formation sans être détenteurs de permis de séjour valable et ceux qui avaient un permis de séjour et qui l'ont perdu, mais qui continuent de suivre une formation, ne sont pas enregistrés dans ZEMIS, mais ils sont enregistrés dans LABB, en raison du fait qu'ils sont inscrits dans une haute école universitaire suisse. Il ne sera pas possible de savoir pourquoi ces étudiants ont perdu leur titre de séjour, ou

pourquoi ils n'en n'ont pas, sans procéder à une enquête rétrospective, ce qui est impossible dans le cadre de cette recherche en raison de la protection des données. Cependant, le parcours académique de ces étudiants sera toujours disponible dans LABB. A partir de ce parcours académique : il sera possible de retrouver une partie du parcours migratoire de ces étudiants, notamment l'année d'arrivée en Suisse et le pays de résidence avant le début des études en Suisse. La seule information très importante qu'on ne pourra pas obtenir à partir de LABB est l'état civil de ces étudiants, parce que la variable état civil n'est pas disponible dans cette base de données. En comparant l'année d'entrée en Suisse des étudiants dans ZEMIS, nous avons pu remarquer que cette année coïncide toujours avec l'année de début des études dans LABB. Ce qui signifie que les étudiants commencent toujours les études à l'année d'arrivée en Suisse. On peut ainsi connaître, à partir de LABB, l'année d'arrivée d'un étudiant en Suisse, indépendamment du registre ZEMIS.

Les étudiants qui font une première formation en Suisse et qui quittent la Suisse après cette formation, pour y revenir après un certain temps de séjour à l'étranger pour suivre une autre formation, ont un parcours plus complexe par rapport à ceux qui n'ont pas interrompu le séjour en Suisse. En effet, pour cette catégorie d'étudiants, nous aurons deux dates d'entrées en Suisse qui sont : la date d'entrée pour la première formation et la date d'entrée pour la seconde formation. Pour cette catégorie d'étudiants aussi, les données de registre ne permettent pas de savoir pourquoi ces étudiants ont interrompu leur séjour pour revenir en Suisse après un certain temps passé à l'étranger. Pour ces étudiants, la date d'entrée à prendre en compte est la dernière date parce que celle-ci englobe la première, d'un point de vu du parcours académique antérieur en Suisse, ainsi que la date de la première entrée, qui correspondra à la date de début des premières études en Suisse. Toutes les informations sur le parcours académique antérieur seront disponibles dans LABB, ce qui fait que le parcours académique antérieur se fusionnera au nouveau parcours académique entamé pour offrir des trajectoires complètes en ce qui concerne le séjour en Suisse.

Nous remarquons donc que la base de données LABB est un excellent complément au parcours migratoire enregistré dans ZEMIS, ce qui lui confère une importance centrale pour comprendre le parcours de vie des étudiants internationaux en Suisse. Un étudiant qui se fait naturaliser suisse disparaît de la base de données ZEMIS mais reste toujours enregistré dans LABB tant qu'il suit une formation dans une haute école en Suisse. Ceci nous permet de dire que la base de données LABB offre dans le long terme des trajectoires plus complètes que ZEMIS, parce que les étudiants restent enregistrés dans LABB, tant qu'ils sont inscrits dans une institution d'enseignement supérieur suisse.

A partir des données de l'OFS, nous avons pu observer que les entrées multiples étaient très rares chez les étudiants africains et plus fréquentes chez les étudiants européens. Cela nous laisse penser que pour des raisons à déterminer, les étudiants africains, une fois en Suisse, sont moins mobiles que les étudiants européens.

Cette situation évoque aussi une autre catégorie d'étudiants : ceux qui étudient sans être détenteurs de permis de séjour valable. Il serait à ce sujet intéressant de faire une étude sur ces étudiants internationaux extra européens qui étudient dans les hautes écoles suisses sans être au bénéfice d'un permis de séjour valable. Dans notre cas, cette catégorie d'étudiants pourrait conduire à sous-estimer l'effectif des étudiants internationaux si l'on travaillait seulement avec le registre ZEMIS. L'utilisation de la base de données de LABB permet de corriger ce type de problèmes, parce que même si les étudiants ne sont pas enregistrés dans ZEMIS, s'ils fréquentent une haute école suisse, ils seront d'office enregistrés dans la base de données de LABB.

Ce qui ne compromet pas la qualité de ces données et n'est pas une critique mais simplement la mise en évidence d'une difficulté rencontrée lors de la préparation des données.

En effet, nous avions à notre disposition toutes les variables explicatives, sauf les variables dépendantes pour nos deux questions de recherche. Ces variables ont dû être construites pour pouvoir continuer cette recherche, mais on aurait aussi pu aboutir à une impasse s'il n'y avait eu aucune possibilité de construire ces variables dépendantes. Il serait dès lors recommandé d'être prudent lors de l'utilisation potentielle de données administratives en se renseignant clairement auprès des autorités sur la disponibilité ou non des variables qui nous intéressent. Les données livrées par l'OFS sont très volumineuses et les différents travaux de mise en commun de ces bases de données nous ont offert une solide expérience en matière de gestion de base de données et de data management. Nous avons appris à tirer les informations utiles dans des bases de données très volumineuses, à documenter des bases de données, tout en archivant les différentes étapes du data management.

Ces différents travaux ont conduit à l'élaboration d'un codebook documentant l'ensemble des données ZEMIS et LABB en vue d'une utilisation future par d'autres chercheurs intéressés par ces données. Ce codebook, qui constitue un des apports de cette thèse, contient 78 pages qui documentent toutes les variables contenues dans les bases de données livrées par l'OFS (ZAR, ZEMIS et LABB). Nous ne nous sommes pas contentés de documenter uniquement les variables qui nous intéressent pour répondre à nos questions de recherche, nous sommes allés au-delà des objectifs de cette thèse en mettant à la disposition des futurs chercheurs intéressés par ces données, un codebook prêt à l'emploi qui constituera pour eux un immense gain de temps.

Les travaux de préparation des données en vue de l'analyse, le contenu des bases de données, ainsi que les difficultés rencontrées, nous permettent de faire les recommandations suivantes en vue de compléter les données et permettre également un gain de temps pour d'éventuelles futures recherches.

L'introduction de la variable état civil dans la base de données des étudiants permettrait de réaliser toute cette recherche sans avoir forcément besoin de fusionner la base de données des étudiants à celle des étrangers. Il serait aussi intéressant d'introduire la variable « prolonger le séjour » dans le registre des étrangers en demandant la collaboration des étudiants internationaux lorsqu'ils quittent la Suisse d'informer l'autorité cantonale compétente de leur départ.

Il serait très intéressant, dans le but de pouvoir suivre les étudiants au-delà de leurs formations universitaires, de relier la base de données du LABB à celle de l'enquête [1] auprès des personnes diplômées des hautes écoles. En effet, la base de données LABB livrée par l'OFS couvre le parcours académique des étudiants internationaux et suisses depuis 1980, alors que la première enquête auprès des personnes diplômées des hautes écoles date de 1977. En reliant les deux bases de données à travers le numéro AVS, nous obtiendrions une base de données longitudinale très riche en informations sur les étudiants d'une part, d'autre part, cette base de données couvrirait une période très longue sur le parcours de vie de ces étudiants, au-delà de la formation universitaire.

A l'image de certains pays scandinaves, il serait intéressant de créer un « registre miroir » qui permettrait la mise à disposition rapide des données aux chercheurs dans le sens où ces données seraient déjà anonymisées et prêtes à être exploitées selon l'intérêt des chercheurs. En effet, en prenant l'exemple de ZEMIS, un registre miroir consisterait à créer en plus du vraie registre ZEMIS détenu par les autorités un autre registre ZEMIS toujours géré par les autorités, et totalement anonymisé et mis à jour dans le but de répondre rapidement aux besoins des chercheurs. La mise en place d'un registre miroir aurait permis un immense gain de temps dans la livraison des données par l'OFS. Je rappelle aussi que nous avons attendu près de deux ans avant de recevoir toutes les données en vue de mener cette recherche, sans compter les conséquences au niveau du temps, de la livraison par vague des données. Cette livraison par vague a eu pour conséquences de refaire plusieurs fois les analyses dans le but d'intégrer à chaque fois aux analyses les nouvelles données reçues.

Cette thèse a permis de montrer que, contrairement aux idées reçues, l'effectif des étudiants africains en Suisse bien qu'étant faible, a augmenté ces dernières années et continue d'augmenter. La répartition de l'effectif des étudiants africains en Suisse est très inégalitaire, car l'Afrique du Nord représente près de la moitié de l'effectif des étudiants africains en Suisse.

A travers cette thèse, nous avons pu mettre en évidence le fait que tous les étudiants internationaux ne bénéficient pas des mêmes conditions d'accueil et de séjour en Suisse. Venir étudier en Suisse est ainsi beaucoup plus complexe pour un étudiant africain que pour un étudiant venant d'un pays membre de l'UE ou d'un pays membre de l'AELE. On peut dès lors faire l'hypothèse que les étudiants africains une fois arrivés en Suisse pour faire des études, ne s'intéressent pas à faire un échange européen durant leur cursus académique en Suisse.

Il serait dès lors intéressant de savoir dans le cadre d'une recherche future s'il y a des étudiants africains ou des étudiants issus de pays extras européens qui s'intéressent à la mobilité européenne durant leurs cursus académique en Suisse.

1. *Cette enquête s'intéresse principalement à la situation professionnelle des personnes diplômées des hautes écoles, un an et cinq ans après l'obtention de leur diplôme, et de la formation qu'elles ont suivie depuis. Elle recherche plus précisément des réponses aux questions suivantes : comment évolue le taux des personnes diplômées actives ? Quels sont les facteurs déterminants pour réussir son entrée dans le monde du travail ? Par les nouveaux enseignements qu'elle fournit, cette enquête représente un instrument d'information de choix pour les hautes écoles et pour les instances de la politique de la formation et de l'emploi (https ://www.bfs.admin.ch/bfs/fr/home/statistiques/education-science/enquetes/ashs.assetdetail.7757.html).*

Nous sommes convaincus que la mise en dialogue de méthodes statistiques d'analyse des parcours de vie avec les problématiques de mobilité constitue une démarche prometteuse pour y répondre.

Liste des tableaux

Liste des figures

Bibliographie

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology 13*, pp.61–98.

Allison, P. D. (1984). *Event history analysis : Regression for longitudinal event data*, Volume 46. SAGE publications.

Allison, P. D. (2010). *Survival Analysis Using SAS : A Practical Guide. Second Edition* (Second Edition ed.). SAS Institut Inc., Cary, North Carolina 27513, USA.

Andrade, M. (2006). International students in english-speaking universities : Adjustment factors. *Journal of Research in International Education 2*(5), 134–154.

Andress, H.-J., K. Golsch, et A. W. Schmidt (2013). *Applied panel Data Analysis for Economic and Social Surveys*.

Aronowitz, R. A. (2011). The framingham heart study and the emergence of the risk factor approach to coronary heart disease, 1947-1970. *Revue d'histoire des sciences 64*(2), 263–295.

Aslanbeigui, N.and Montecinos, V. (1998). Foreign students in us doctoral programs. *The Journal of Economic Perspect 3*(12), 171–182.

Auriat, N. (1991). Who forgets ? an analysis of memory effects in a retrospective survey on migration history. *European Journal of Population / Revue Européenne de Démographie 7*(4), 311–342.

Ballatore, M. et T. Blöss (2008, Juillet-septembre). L'autre réalité du programme erasmus : affinité sélective entre établissement et reproduction sociale des étudiants. *Formation emploi Fuite ou mobilité des cerveaux ?* (103).

Barry, M. P. (2017). Les étudiants africains dans l'enseignement supérieur suisse : Pays d'origine, filières d'études et nouvelles tendances. pp. 93–107. Société Neuchâteloise de Géographie, Institut de Géographie de l'Université de Neuchâtel
Editions Alphil, Presses Universitaires Suisses.

Bijwaard, G. E. et Q. Wang (2016, January). Return migration of foreign students. *Eur J Population* (32), 31–54.

Black, C. et L. L. Roos (2005). Linking and combining data to develop statistics for understanding the populations health. *Health Statistics : Shaping Policy and Practice to Improve the Population1s Health*, 214–240.

Blossfeld, H.-P., K. Golsch, et G. Rohwer (2009). *Event History Analysis with Stata*. Psychology Press, Taylor and Francis Group.

Blossfeld, H.-P. et G. Rohwer (2002). Techniques of event history modeling : New approaches to causal analysis.

Bolzman, C. et I. Guissé (2015). *Etudiants du sud et internationalisation des hautes écoles : entre illusion et espoirs*. Collection du Centre de recherche sociale. ISBN : 9782882241429.

Brad, T. (1999, June). Probabilistic record linkage software : A statistics canada evaluation of grls and automatch. *Proceedings of the Survey Methods Section*.

Campus France, C. F. (2016, Octobre). « la mobilité internationale des étudiants africains », les notes de campus france hors-série. (16).

Carolyn J., A. (2005). *Journal of the American Statistical Association 100*(469), 352–353.

Cattan, N. (2004). Genre et mobilité étudiante en europe. *Espace populations sociétés*, 15–24.

CIMO (2016). In finland, at work, elsewhere ? status of international higher education students in finland 5 years after their graduation, facts express. Technical Report 1B, Helsinki : Centre for International Mobility.

Cleves, M., G. Roberto G., G. William, et M. Yulia V. (2010). *An Introduction to Survival Analysis Using Stata*. Stata Press.

Collett, D. (2015). *Modelling survival data in medical research* (3rd Edition ed.). Chapman and Hall/CRC Press.

Coulibaly, S. Z. (2011, Juillet). De la nécessite de dynamiser le traitement des données des sources administratives : Elements de reflexion. *Observatoire économique et statistique d'Afrique Subsaharienne*.

Courgeau, D. (1991, Janvier-Février). Analyse de données biographiques erronées. *Institut d'études Démographiques 46*(1), pp.89–104.

Courgeau, D. et E. Lelièvre (1989). *Analyse démographique des biographies*. Institut national d'études démographiques.

Courgeau, D. et E. Lelièvre (1993). Nouvelles perspectives de l'analyse biographique. *Cahiers québécois de démographie 22*(1), 23–43.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Desrosières, A. (2011, septembre). Enquêtes versus registres administratifs : réflexions sur la dualité des sources statistiques. *Courrier des statistiques* (111).

Diekmann, A. (1990). Diffusion and survival models for the process of entry into marriage. *Event History Analysis in Life Course Research*, 170–183.

Diggle, P. et M. G. Kenward (1994). Informative drop - out in longitudinal data analysis. *Journal of the Royal Statistical Society* (43), pp. 49–93.

Dodge, Y. et V. Rousson (2004). *Analyse de régression appliquée*. Dunod.

Doll, R. et A. B. Hill (2004). The mortality of doctors in relation to their smoking habits : A preliminary report. *BMJ : British Medical Journal 328*(7455), 1529–1533.

Doll, R., R. Peto, J. Boreham, et I. Sutherland (2004). Mortality in relation to smoking : 50 years' observations on male british doctors. *BMJ : British Medical Journal 328*(7455), 1519–1528.

Dormont, B. (1989). Petite apologie des données de panel. *Économie and prévision* (87), pp. 19–32.

Droesbeke, J.-J., B. Fichet, et P. Tassi (1989). *Analyse statistique des durées de vie : Modélisation des données censurées*. Association pour la statistique et ses utilisations.

Dupont, L. (2006, June). *Contribution à l'étude de la durée de vie des assemblages de puissance dans des environnements haute température et avec des cycles thermiques de grande amplitude*. Theses, École normale supérieure de Cachan - ENS Cachan.

Efionayi, D. et E. Piguet (2014). Les étudiants d'afrique de l'ouest face à la globalisation du savoir. *Revue internationale de politique de développement 5 International Development Policy 116*(2).

Elder, G. H. (1998). The life course as developmental theory. *Child Development 69*(1), 1–12.

Elder, G. H. et S. Y. C. Kenneth (1987). Life course dynamics : Trajectories and transitions, 1968-1980. by Glen H. Elder, jr. *Social Forces 66*(2), 587–588.

Elder, G. H. et S. K. Steinmetz (1976). Children of the great depression-social change in life experience. *Contemporary Sociology 5*(3), 287–288.

Endrizzi, L. (2010, Avril). La mobilité étudiante, entre mythe et réalité. *Institut National De Recherche Pédagogique. Revue de littérature qui appartient à la collection des Dossiers d'actualité de la Veille scientifique et technologique* (51). https ://halshs.archives-ouvertes.fr/halshs-00473752.

Erlich, V. (2012). Les mobilités étudiantes. *Revue Française de pédagogie*, 129–132.

Estève, J., E. Benhamou, et L. Raymond (1994). Statistical methods in descriptive epidemiology. *International Agency for Research on Cancer 4*(128).

European Commission (1998). Swiss-Swedish joint study on cohort-based asylum statistics. *Eurostat* (2).

Evans, M. (2017, October). A review of statistical failure time models with application of a discrete hazard based model to 1cr1mo-0.25v steel for turbine rotors and shafts. *Materials, Open Access Journal* (1190). doi :10.3390/ma1010119.

Ewers, M. C. et J. M. Lewis (2008, January). Risk and the securitisation of student migration to the united states. *Jounal of Economic and Social Geography 99*.

Findlay, A. M. (2011, September). An assessment of supply and demand-side theorizations of international student mobility. *International Migration 49*. https ://doi.org/10.1111/j.1468-2435.2010.00643.x Cited by : 72.

Findlay, A. M., R. King, F. M. Smith, A. Geddes, et R. Skeldon (2011, January). Worls class ? an investigation of globalisation, difference and international student mobility. *Royal Geographical Society*.

Finn, M. G. et L. A. Pennington (2003, March). Stay rates of foreign national doctorale students in u.s. economic programs. Technical report, Science Education Programs Oak Ridge Institute for Science and Education.

Flahaux, M.-L. et H. De Haas (2016). African migration :trends, patterns, drivers. *Comparative Migration Studies 4*.

Gaillard, A.-M. et J. Gaillard (2002, Février). Fuite des cerveaux, circulation des compétences et développement : un enjeu politique. *Mots pluriels* (20).

Gaillard, J. (2002). Entre science et subsistance quel avenir pour les chercheurs africains ? *Fondation internationale pour la science (IFS) et Institut de Recherche pour le Développement* (Novembre-décembre).

Garneau, S. (2007). Les expériences migratoires différenciées d'étudiants français. de l'institutionalisation des mobilités étudiantes à la circulation des élites professionnelles ? *Revue Européenne des Migrations Internationales 23*(1), pp. 139–161.

Giroux, E. (2011). Origines de l'étude prospective de cohorte : Épidémiologie cardio-vasculaire américaine et étude de framingham. *Revue d'histoire des sciences 64*(2), 297–318.

González Rodríguez, C., B. R. Mesanza, et P. Mariel (2011, December). The determinants of international student mobility flows : an empirical study on the erasmus programme. *High Educ* (62), 413–430.

Gravlee, C. C., D. P. Kennedy, R. Godoy, et W. R. Leonard (2009). Methods for collecting panel data : What can cultural anthropology learn from other disciplines ? *Journal of Anthropological Research 65*(3), 453–483.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *The Public Opinion Quarterly 70*(5), 646–675.

H., E. G. (1994). Time, human agency, and social change : Perspectives on the life course. *Social Psychology Quarterly 57*(1), 4–15.

Habibi, D., M. Rafiei, A. Chehrei, et T. S. Shayan, Zahra (2018, January). Comparison of survival models for analyzing prognostic factors in gastric cancer patients. *Asian Pacific Journal of Cancer Prevention 3*(19), 749–753. doi : 10.22034/APJCP.2018.19.3.749.

Hill, C., C. Com-Nougueé, A. Kramar, T. Moreau, J. O'Quigley, R. Senoussi, et C. Chastang (1991). *Analyse statistique des données de survie* (1 ed.). 4, rue Casimir-Delavigne 750006 Paris : Médecine Sciences Flammarion.

Hillygus, D. S. et S. Snell (2015, December). *Longitudinal Surveys : Issues and Opportunities*. Political Science, Political Methodology, Political Behavior.

Hofsten, E. et H. Lundstrom (1976). Swedish population history. main trends from 1750-1970. *Population* (2), 471.

Hoseini, M., A. Bahrampour, et M. Moghaddameh (2017, February). Comparison of weibull and lognormal cure models with cox in the survival analysis of breast cancer patients in rafsanjan. *Journal of Research in Health Sciences 17*(1).

Hosmer, D. W. et S. Lemeshow (1999). *Applied Survival Analysis : Regression Modeling of Time to Event Data*. Wiley series in probability and Statistics.

Iglesias, K., P. Gazareth, et C. Suter (2017). *Explaining the Decline in Subjective Well-Being Over Time in Panel Data*. University of Neuchâtel : Springer.

Jabine, T. B. et F. Scheuren (1985). Goals for statistical uses of administrative records : The next 10 years. *Journal of Business and Economic Statistics 3*(4), 380–391.

Kabbanji, L., A. levatino, et F. Ametepe (2013). Migrations internationales étudiantes ghanéennes et sénégalaises : caractéristiques et déterminants. *42*(2), 303–333.

Kankoé, S. (2010, août). Méthode actuarielle d'estimation des courbes de survie : principe, différences avec la méthode de Kaplan-Meier.

Kaplan, E. L. et P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), pp. 457–481.

Kargarian-Marvasti, S., S. Rimaz, J. Abolghasemi, et I. Heydari (2017, October). Comparing of cox model and parametric models in analysis of effective factors on event time of neuropathy in patients with type 2 diabetes. *Journal of Research in Medical Sciences 22*(115). J Res Med Sci. 2017 ; 22 : 115.

Keller-Gerber, A. (2017). Poursuivre sa carrière à l'étranger. histoires de parcours d'étudiants immigrants en Suisse : du récit de mobilité au récit d'établissement. *Journal of international Mobility* (5), 93–114.

Khan, S. A. (2017, March). Exponentiated weibull regression for time-to-event data. *Lifetime Data Anal* (24), 328–354. https ://doi.org/10.1007/s10985-017-9394-3.

Kim, D., C. A. S. Bankart, et L. Isdell (2011). International doctorates : trends analysis on their decision to stay in us. *Higher Education 62*(2), 141–161.

Klein, J. P. et M. L. Moeschberger (2003). *Survival analysis : techniques for censored and truncated data* (second ed.). Springer.

Kleinbaum, D. G. (1996). *Survival Analysis. A Self-Learning Text*. Statistics in the Health Sciences.

Latreche, A. (2001, Septembre-Octobre). Les migrations étudiantes de par le monde. *Homme et Migrations* (1233), 13–27.

Le Goff, J.-M. (2003). Modélisation des événements du parcours de vie : Une introduction. In *Centre lématique des parcours et modes de vie(PaVie) et laboratoire de démographie et détudes familiales*.

Le Goff, J.-M. (2012). Time dependency in diffusion models : Gamma-diffusion models as an alternative to the Hernes model. *LIVES Working Papers 2012*, 1–24.

Le Goff, J.-M., Y. Forney, J.-P. Antonietti, et A. Berchtold (2013, Février). Méthodes non-paramétriques de l'analyse des événements du parcours de vie (event history analysis) estimations avec SPSS méthode de Kaplan-Meier et méthode actuarielle. *Cahiers Recherche et Méthodes* (2).

Le Goff, J.-M., F. Yannic, A. Jean-Philippe, et B. André (2013, Février). Analyse des événements de l'histoire de vie : estimation de modèles logistiques à temps discret avec SPSS. Technical Report 3, Université de Lausanne, Faculté des SSP CH-1015 LAUSANNE.

Levy, R., J.-A. Gauthier, et E. Widmer (2006). Entre contraintes institutionnelle et domestique : les parcours de vie masculin et féminin en Suisse. *Canadian journal of sociology 31*(4), 461–489.

Lindsey, J. (1999). *Models for Repeated Measurements* (2 ed.). Oxford University Press.

Lollivier, S. (2000). Récurrence du chômage dans l'insertion des jeunes : des trajectoires hétérogènes. *In : Economie et statistique* (334), 49–63. http ://www.persee.fr/doc/estat-0336-1454-2000-num-334-1-7530.

Lynn, P. (2009). *Methodology of Longitudinal Surveys*. Departement of Survey Methodology, Institute for Social and Economic Research, University of Essex, UK : Willey.

Meier, P., T. Karrison, R. Chappell, et H. Xie (2004). The price of kaplan-meier. *Journal of the American Statistical Association 99*(467), pp. 890–896.

Menard, S. (2002). *Longitudinal Research*. Thousand Oaks, CA : SAGE Publications.

Mendy, A. F. (2014). La carrière du médecin africain en europe :être médecin avec un diplôme africain au royaume-uni, en france et en suisse. *Swiss Journal of Sociology 40*(1), 55–77.

Missine, L.-E. (1968). Problèmes concentrant l'éducation supérieure en Afrique. *Revue Internationale de l'éducation 14*(1), 62–74.

Montaseri, M., J. Y. Charati, et E. Fateme (2016, November). Application of parametric models to a survival analysis of hemodialysis patients. *Nephrology and Urology Research Center 6*(8). doi : 10.5812/numonthly.28738.

Mosneaga, A. et L. Winther (2013). Emerging talents ? international students before and after their career start in denmark. *Population Space and Place*.

Mueller, N. S. (2011). *Inégalités sociales et effets cumulés au cours de la vie : concept et méthodes*. Ph. D. thesis, Université de Genève.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica 46*(1), 69–85.

Nations-Unies (2003). Principes et recommandations pour un système de statistiques de l'état civil. Technical Report 2, ONU.

NUFFIC (2016). « analysis of stay rate of international graduates : 2008-2009 », the hague : Nuffic. Technical report, Netherlands Universities Foundation for International Cooperation.

OCDE (2014). Regards sur l'éducation 2014 : Les indicateurs de l'OCDE, Éditions OCDE. Technical report, OCDE.

OFS (2015). Les étudiants internationaux dans les Hautes Ecoles Suisses, rapport thématique de l'enquête 2013 sur la situation sociale et économique des étudiant-e-s. Technical report, OFS.

OFS (2018). Scénarios 2018-2027 pour les hautes écoles-étudiants diplômés. Technical report, OFS.

Olivas, M.and Li, C. S. (2002). Understanding stressors of international students in higher education : What college counselors and personnel need to know. *Journal of Instructional Psychology*, 217–222.

Olivieri, I. et R. Vitalis (2001, Janvier). La biologie des extinctions. *médecine/sciences* (1).

OSCE (2009). Guidelines on population registration. Technical report, OSCE-ODIHR.

Piguet, E. et J.-H. Ravel (2002). Les demandeurs d'asile sur le marché du travail suisse. Rapport de recherche 19, Université de Neuchâtel Forum suisse pour l'étude des migrations et de la population.

Poulain, M. et A. Hern (2013). Le registre de population centralisé, source de statistiques démographiques en europe. *Institut national d'études démographiques (INED) « Population » 68*, 215–247.

Poulain, M., B. Riandey, et J.-M. Firdion (1991). Enquête biographique et registre belge de population : une confrontation des données. *Population* (1), pp. 65–87.

Rakotonarivo, A. (2013). Mobilité internationale étudiante et insertion professionnelle : parcours différenciés de migrants congolais en Belgique. *Cahiers québécois de démographie 42*(2), 273–302.

Ray, J.-C. (1988). Données censurées et modèles de durées. *Recherche et Applications en Marketing 3*(2), 77–88.

Ritschard, G. (2004). Estimer un modèle de Cox en temps continu avec SPSS. In *Modélisation des événements et transitions du parcours de vie(Event History Analysis)*.

Robette, N. (2011). *Explorer et décrire les parcours de vie : les typologies de trajectoires*. CEPED.

Rowebottom, L. (1978). L'utilisation des dossiers administratifs à des fins statistiques. *Techniques d'enquête 4*(1). Statistique Canada, no 12-001-X au catalogue.

Rudolph, J. E., S. R. Cole, et J. K. Edwards (2018, November). Parametric assumptions equate to hidden observations : comparing the efficiency of nonparametric and parametric models for estimating time to aids or death in a cohort of hiv-positive women. *BMC Medical Research Methodology 1*(18). doi : 10.1186/s12874-018-0605-8.

SER (2006). Panorama de l'enseignement supérieur en suisse. Technical report, Secrétariat d'Etat à l'éducation et à la recherche SER et Office fédéral de la formation professionnelle et de la technologie OFFT en collaboration avec Présence Suisse et la Conférence universitaire suisse, Hallwylstrasse 4, CH-3003 Bern.

Stoop, I. A. L. (2005). The hunt for the last repondent : Non response in sample surveys. the hague :social and cultural planning office of Netherlands. Technical report, SCP.

Sykes, B. et E. Ni Chaoimh (2013). Mobile talent ? : the staying intentions of international students in five eu countries, sachverstandigenrat deutscher stiftungen fur integration und migration (svr) gmbh, berlin.

Tansel, A. et N. D. Güngör (2003). Brain drain from turkey : Survey evidence of student non-return. *Carrer Development International 2*(8), 52–69.

Terrier, E. (2009a). Les migrations scientifiques internationales pour études : facteurs de mobilité et inégalités Nord-Sud. *L'information géographique 73*, 69–75.

Terrier, E. (2009b). Les mobilités spatiales des étudiants internationaux. déterminants sociaux et articulation des échelles de mobilité. *Armand Colin* (670), pp. 609–636.

Terrier, E. (2010, décembre). Mobilités spatiales / inégalités sociales : le cas de la mobilité internationale pour études. *ESO-Rennes* (30).

Tourangeau, R. (2003). Recurring surveys : Issues and opportunities. Technical report, National Science Foundation Based on a Workshop.

Unesco (2012). New patterns in student mobility in the Southern Africa development community. *UIS information bulletin*.

Unige (2016). Statistique universitaire etudiantes et étudiants, diplômes et personnel. Technical report, Université de Genève.

Verger, J. (1991). La mobilité étudiante au moyen Âge. *Education Médiévales : L'Enfance, l'Ecole, l'Eglise en Occident : Ve- XVe siècle* (50).

Vermunt, J. K. (1997). *Log-linear Models for Event Histories*. Newbury-Park :Sage publications.

Villar, E., L. Frimat, R. Ecochard, et M. Labeeuw (2008). Spécificités méthodologiques de l'analyse de survie des patients dialysés. *Néphrologie & Thérapeutique 4*(7), 553–561.

Wanner, P. et Y. Forney (2007). Une reconstitution de trajectoire de vie des personnes décédées en suisse, 1990-2004. Colloque de Bordeaux.

Yamagushi, K. (1991). Event history analysis. *Applied Social Research Methods Series 28*.

Yang, R. (2003). Globalisation and higher education development : A critical analysis. *International Review of Education 49*(3/4), 269–291.

Zare, A., M. Mahmoodi, K. Mohammad, H. Zeraati, H. Mostafa, et K. H. Naieni (2013). Comparison between parametric and semi-parametric cox models in modeling transition rates of a multi-state model : Application in patients with gastric cancer undergoing surgery at the iran cancer institute. *Asian Pacific Journal of Cancer Prevention*.

Zhang, Z. (2016, June). Parametric regression model for survival data : Weibull regression model as an example. *Annals of Translational Medicine 4*(24). doi : 10.21037/atm.2016.08.45.