UNIVERSITÉ DU QUÉBEC À MONTRÉAL

DONNÉES DE PANEL D'ENQUÊTE DANS LE CONTEXTE D'UN PROCESSUS DE RENOUVELLEMENT ALTERNÉ TRONQUÉ ET CENSURÉ PAR INTERVALLE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

PATRICK-HERVÉ TIAN

FÉVRIER 2023

REMERCIEMENTS

Comment ne pas commencer cette série de reconnaissance en rendant gloire au Seigneur des seigneurs JÉSUS qui lorsque mes forces s'amenuisaient, les renouvelait continuellement. Durant toutes les épreuves que j'ai traversées, IL a été et est, un soutien indéfectible. Sa présence me rassure !

Cette merveilleuse aventure a pris forme par celle que j'appelle affectueusement « la mère », Sorana Froda, qui m'a encouragé à entreprendre ces travaux et, par ses conseils m'a permis de postuler pour les différents concours de bourses d'excellence. Un très grand merci Sorana. Je vous souhaite une retraite remplie de belles aventures et pleine de quiétude.

Les honneurs vont à ma directrice de recherche, Juli Atherton, qui m'a toujours soutenu et accompagné dans cette quête de la connaissance statistique avec ses idées inspirantes et ses analyses pointues. Ses conseils ont été précieux autant dans ma formation académique qu'en dehors. En plus d'être une excellente professeure, elle s'est avérée être une très bonne mentore. Merci madame.

Les données ! Faire de la statistique sans celles-ci ? Impensable. Grand merci à Lenin A. Castillo qui m'a initié aux données ayant pour point focal ce mémoire. Il est celui qui m'a guidé dans l'obtention des accréditations nécessaires pour avoir accès à elles au centre de données et de recherche (CDR) et le « mentor SAS ».

Merci à Franck Larouche, analyste au CDR de l'UdeM, pour ses conseils et son aide.

Je n'oublie pas toutes ces personnes et organismes qui m'ont apporté leurs financements et leurs soutiens dans la réalisation de mes études. Je pense notamment à :

— Juli Atherton qui, avant d'être la directrice de ce mémoire a été, pendant toutes ces années passées à l'UQÀM au Baccalauréat et à la Maîtrise, celle qui a financé et m'a soutenu à obtenir des financements à travers le conseil de recherches en sciences naturelles et en génie du Canada (CRSNG). Merci aussi d'avoir été ma référante pour mon actuel poste en tant que méthodologiste dans le plus prestigieux organisme national de statistique qu'est Statistique Canada ;

— Alain Desgagné. Merci pour toutes ces opportunités offertes de faire mes preuves en tant qu'assistant-enseignant au sein de l'université et tous les sages conseils prodigués. Merci aussi d'avoir été mon référant pour l'obtention d'un poste à Statistique Canada ;

— René Ferland, pour la bourse Fisher-Tukey octroyée dans mes années de Baccalauréat : un soutien considérable ;

— le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) pour les bourses aux stages d'été dans mes années de Baccalauréat. Ces stages m'ont permis de développer un goût pour la recherche et d'avoir un regard minutieux sur l'échantillonnage et les données d'enquête. Cela n'a pas seulement été une source de motivation pour mon mémoire ;

— le Fonds de recherche du Québec – Nature et technologies (FRQNT) pour le financement de ma maîtrise ;

— le centre de recherche facultaire STATQAM du département de mathématiques de l'UQÀM pour les bourses de soutien à la recherche ;

— le Département de mathématiques de l'UQÀM pour son accueil et son assistance. En particulier, à Isabella Couture qui a toujours fourni des informations de qualité et de l'aide dans la prolongation lorsque j'étais pris avec des obligations familiales et professionnelles.

À ma famille et mes amis :

— mon épouse Brigitte. Parfois psychologue, parfois amie ; parfois mère, parfois gen-

darme. Merci Bri, d'être un soutien inestimable. Je suis chanceux de t'avoir à mes côtés. Je t'aime chérie ;

— Yanelle, ma fille, 12 ans. Très bienveillante et compréhensive quand papa manquait quelques fois de temps pour nos activités communes. Tu es un trésor mademoiselle Tian. Que les grâces, les bénédictions et la miséricorde du Seigneur Jésus t'accompagnent tous les jours.

— Onénihi, soit le bienvenu. Que l'Éternel te bénisse et te positionne où coulent le lait et le miel ;

— mes frères et sœur : Anicet, Diane et Brice. Merci pour les prières et les encouragements ;

— la famille Quenum : Flora Tapé, Marie Holy, Harrison et Léopold Quenum. Vous avez été, et continuez d'être présents dans les moments-clés de ma vie. Que le Seigneur se souvienne toujours de vous et vous bénisse au-delà de vos espérances ;

— mes oncles Apollinaire et Mathias pour les encouragements ;

— René Kouarfaté pour son soutien.

Soyez, chacun et tous, immensément bénis et comblés de grâces au delà de vos espérances. Que le Seigneur veille sur vous.

À Élisabeth Zihon et Benoît Dagri, mes parents.
À Oscar, mon cadet.

Je puis tout par celui qui me fortifie.
Philippiens 4 : 13 LSG

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Modèles statistiques décrivant différents états d'un processus ainsi que les différentes transitions possibles entre ces états, les modèles multi-états sont utiles, par exemple, pour les mesures répétées dans le temps et les données longitudinales. Un cas particulier est le processus de renouvellement alterné dans lequel l'individu pivote entre deux états (0 et 1, Sain et Malade, ou Chômage et Emploi comme dans notre cas de figure). Le sujet de la présente recherche est motivée par des données d'enquête par panel telles que l'Enquête sur la dynamique du travail et du revenu (EDTR), une enquête de Statistique Canada, qui suit les individus du début du panel à la fin de celui-ci. Enquête longitudinale à participation volontaire, l'EDTR est composée de deux panels dont les individus transitent entre chômage et emploi et sont suivis sur une durée de six ans consécutifs avec pour objectif de comprendre le bien-être économique des Canadiens et les changements ayant des incidences dans le temps sur leurs conditions de vie. Les données, parfois collectées rétroactivement avant le début du panel, emmènent des défis quant aux dates de début et de fin des périodes de temps.
Dans un processus de renouvellement alterné censuré par intervalle PRACI *(window censored alternating renewal process WCARP)*, les bornes supérieure et inférieure interceptent la première durée de séjour observée ainsi que la dernière durée de séjour observée respectivement. Motivé par différentes approches vues dans la littérature, dans ce mémoire nous développons trois fonctions de vraisemblance dans les situations suivantes :

1. lorsqu'aucune donnée de temps de début de la période n'est collectée avant l'entrée dans le panel : la première durée de séjour est incomplète. Dans cette approche, elle est supprimée. La dernière durée de séjour observée est traitée en utilisant la censure à droite ;
2. lorsque les dates de début des périodes de temps sont collectées avant introduction dans le panel : de ce fait, nous connaissons la valeur complète de la première durée de séjour observée dans le panel. Une approche conditionnelle est utilisée pour ces premiers temps de séjour ; et à nouveau une censure à droite est utilisée pour les dernières périodes incomplètes de l'intervalle ;
3. lorsqu'aucune donnée n'est collectée avant le début du panel, au lieu de « jeter » les premières périodes (comme dans la première approche), en supposant la stationnarité, nous incluons ces périodes incomplètes dans notre approche ; puis traitons à nouveau la dernière période incomplète en utilisant encore une censure à droite.

L'EDTR produit un ensemble de données très complexe qui comprend une pondération pour différents échantillons. Une description en détail et une présentation de quelques données tabulaires non pondérées de l'EDTR (tout en mettant l'accent sur les temps de transition manquants) ainsi que la description de bon nombre de difficultés provenant de l'ensemble

de données lors de l'analyse des périodes de chômage et d'emploi est faite.
Dans une étude de simulation, nous générons un jeu de données semblables à celles de l'EDTR puis nous estimons, dans chacune des trois méthodes décrites plus haut, les paramètres des lois ayant permis de générer cet échantillon. Dans cette étude, nous faisons varier différents paramètres tels que la largeur de la fenêtre d'observation et la taille de l'échantillon. D'autres paramètres comme le taux de chômage et les paramètres des fonctions permettant la génération des durées ont été fixés. Une réplication de 100 fois chaque situation a permis d'obtenir les résultats qui sont présentés dans le Chapitre 4.
Dans ce mémoire est apporté un traitement à l'une parmi les nombreuses complexités présentées par les données de panel d'enquête en général et par l'EDTR en particulier qui a motivé ce travail.

MOTS-CLÉS : modèles multi-états, processus de renouvellement alterné, enquêtes longitudinales, panel, EDTR, vraisemblances, EMV, chaîne continue de Markov homogène, stationnarité, optimisation, simulation Monte Carlo.

INTRODUCTION

Dans ce mémoire, nous considérons une analyse longitudinale de données de panel d'enquête. Nous nous concentrons, en particulier, sur une analyse des périodes d'emploi et de chômage. Les termes périodes d'emploi et périodes de chômage font référence à la période de temps passé en emploi et hors emploi. Tout au long de ce document, nous utilisons les termes *période, période de temps* et *temps de séjour* de manière interchangeable. Le terme *période* est couramment utilisé dans la littérature économique tandis que le terme *temps de séjour* apparaît dans la littérature sur la modélisation multi-états et tous les deux font référence au temps passé dans un état. Dans l'application motivant notre travail, les états sont salarié (ou employé) et chômeur.

Notre travail découle d'une collaboration en cours, Arango-Castillo *et al.* (2019), basée sur l'enquête sur la dynamique du travail et du revenu (EDTR). L'objectif principal d'Arango-Castillo *et al.* (2019) est d'analyser les effets de la récession de 2008 sur les périodes de chômage. Au cours de cette enquête, un certain nombre de questions se sont posées concernant la meilleure façon d'analyser les périodes de chômage (ou d'autres types de données sur les périodes) à partir des données de panel d'enquête.

En utilisant l'EDTR à titre d'illustration, nous commençons, au Chapitre 1, par présenter les données de panel d'enquête et les difficultés qui surviennent lorsqu'on utilise ces données pour faire une analyse longitudinale des périodes de temps. Ici, notre application concerne les périodes d'emploi et de chômage. Cependant on pourrait envisager d'autres applications comme une analyse des temps de séjour dans les états scolaire et extra-scolaire en sciences sociales pour étudier le décrochage scolaire ou en biostatistique pour modéliser le temps de séjour passé dans un état d'une certaine maladie ou en dehors (sain).

Dans ce mémoire, nous considérons deux cas : 1) un modèle simple à deux états appelé processus de renouvellement alterné faisant partie de la grande famille des modèles multi-états et 2) une chaîne de Markov homogène continue. L'analyse consiste à estimer les intensités de transition entre les états.

En utilisant des données de panel d'enquête, on souhaite estimer les paramètres de ces modèles. À chaque fois, il y aura des temps de séjour ou des périodes de temps qui sont coupés au début et à la fin du panel. Notre objectif principal dans ce mémoire est de proposer différentes méthodes pour faire des inférences sur ces périodes sectionnées par le début et la fin du panel.

Nous commençons à la Section 1.1 du Chapitre 1 en décrivant les principales caractéristiques de l'EDTR. Avant d'aborder la complexité des données d'enquête en général à la Section 1.3, nous proposons à la Section 1.2 une analyse descriptive des durées d'emploi et de chômage dans l'EDTR en accordant une attention particulière aux temps de transition manquants entre les états *Emploi* et *Chômage* Nous concluons le Chapitre 1 en résumant les difficultés rencontrées lors de l'utilisation de données de panel d'enquête pour analyser de telles périodes. Les difficultés sont causées par le plan d'échantillonnage complexe (dans le cas de l'EDTR), les données manquantes ou incomplètes (en particulier les dates de transition manquantes entre les périodes d'emploi et celles de chômage) et les complexités de modélisation. Les complexités de modélisation comprennent des covariables longitudinales variant dans le temps et la présence de différentes échelles de temps (par exemple, le temps calendaire, la durée de la période et l'âge des individus).

De toutes les complications énumérées dans le Chapitre 1, nous nous concentrons dans les Chapitres 2, 3 et 4 sur la façon de traiter les périodes sectionnées par le début et la fin du panel. Certains analystes, tels que Kovačević et Roberts (2007) suppriment simplement les périodes qui sont coupées au début du panel et traitent les périodes incomplètes à la fin en

utilisant la censure à droite. Au lieu de simplement supprimer les périodes incomplètes au début du panel, les auteurs d'Arango-Castillo *et al.* (2019) se sont demandés si les périodes incomplètes pouvaient être conservées lors de l'inférence. Il a été noté que bien que les dates de début du chômage soient connues avant le début du panel pour certaines observations de l'EDTR, ce n'était pas le cas pour toutes les observations de cette enquête. Cette question de savoir comment inclure les périodes incomplètes dues à la section par le début du panel et aussi le constat que toutes les dates de début (les temps calendaires initiaux) ne sont pas connues pour les périodes de chômage ou d'emploi commençant avant l'entrée dans le panel constituent la base des travaux dans ce mémoire.

Par conséquent, lorsque l'on considère comment faire des inférences à partir des périodes de temps sectionnées par le début du panel, nous accordons une attention particulière au type de données requises, en particulier si les temps calendaires initiaux sont nécessaires. Les hypothèses utilisées pour chaque approche, et si ces hypothèses sont réalistes ou non, sont également importantes. Encore une fois, nous nous référons à l'enquête de l'EDTR à titre d'illustration pour déterminer si les différentes hypothèses sont réalistes ou pas. À noter qu'à la fin du panel les sujets quittent l'EDTR, nous ne considérons donc pas les situations où les temps calendaires finaux (les dates de fin des périodes de chômage ou d'emploi) sont connus après la fin du panel.

En bref, les trois approches que nous considérons sont les suivantes :

1. l'approche consistant à simplement laisser tomber toutes les périodes incomplètes au début du panel. Elle est abordée dans Kovačević et Roberts (2007). Étant donné que les périodes incomplètes sont exclues de l'analyse, les dates de début avant introduction dans le panel ne sont pas requises. Aucune hypothèse de modélisation supplémentaire n'est formulée ;

2. l'approche classique de l'utilisation d'une vraisemblance conditionnelle : voir le Chapitre 7 de Cook et Lawless (2018). L'utilisation de cette approche nécessite de connaître les temps calendaires initiaux (les dates de début des périodes), mais aucune hypothèse de modélisation supplémentaire n'est requise ;

3. étant donné que les dates de début des périodes avant introduction dans le panel (les temps calendaires initiaux) sont souvent inconnues pour de nombreux individus, nous avons considéré une approche qui suppose que le processus de renouvellement alterné d'être en emploi et sans emploi a atteint la stationnarité. En faisant cette hypothèse, nous avons plus d'informations sur les périodes qui sont interrompues par le début du panel et nous pouvons utiliser ces périodes incomplètement observées pour faire une inférence même si leurs dates de début avant introduction dans le panel sont manquantes. Ce type d'analyse apparaît souvent sous le nom de processus de renouvellement alterné censuré par intervalle - (PRACI) ou *window censored renewal processes - (WCRP)* en anglais. Cette approche est le principal sujet présenté dans Alvarez (2006).

Tout au long de ce mémoire, nous suivons largement la notation et les résultats de l'article théorique Alvarez (2006) qui n'envisage aucune application. Contrairement à Alvarez (2006), nous indiquons quel type de données est nécessaire pour chaque approche lorsque nous traitons les périodes incomplètes au début et à la fin du panel de l'enquête dans le Chapitre 2. En utilisant l'EDTR comme exemple, nous considérons quels types de données seraient disponibles de manière réaliste. Pour faciliter la lisibilité du mémoire, nous simplifions la présentation d'Alvarez (2006). Les détails techniques sont laissés au lecteur pour référence dans l'annexe et dans l'article original.

En résumé, après avoir présenté les données en notre possession dans le Chapitre 1, nous commençons au Chapitre 2 par présenter la théorie ; à savoir le modèle de renouvellement alterné avec les états C de chômage et E d'emploi ainsi que les hypothèses spécifiques à

chaque approche évoquée ci-dessus pour traiter les périodes incomplètes dues à la collecte des données de panel. Dans le Chapitre 2 nous présentons les hypothèses de modélisation en général et nous ne précisons pas encore les distributions f_E et f_C pour les temps de séjour dans les états E et C respectivement.

Au Chapitre 3, nous maximisons chaque vraisemblance pour estimer les paramètres du processus de renouvellement à deux états en utilisant chacune de nos trois approches. Nous simplifions le problème en utilisant une chaîne continue de Markov homogène à deux états, un processus de renouvellement alterné très simple, impliquant que les périodes d'emploi et de chômage proviennent de distributions exponentielles f_E et f_C.

Le Chapitre 4 présente quelques résultats de simulation. Ces résultats sont obtenus en simulant cent fois chaque situation présentée plus haut pour lesquelles nous avons généré une base de données semblable à l'EDTR à partir d'un processus de renouvellement alterné retardé en fixant certains paramètres comme ceux des fonctions ayant généré les durées d'emploi et de chômage ainsi que le taux de chômage. D'autres paramètres comme la taille de l'échantillon et la largeur de la fenêtre d'observation (le panel) sont variés.

La conclusion, quant à elle, résume brièvement le mémoire et fournit des idées pour quelques extensions à ce que nous avons étudié.

CHAPITRE 1
ENQUÊTE SUR LA DYNAMIQUE DU TRAVAIL ET DU REVENU (EDTR)

Les données ! Faire de la statistique sans celles-ci ? Impensable.

En plus de produire des statistiques sur plusieurs caractéristiques de la population, Statistique Canada, à travers ses centres de données et de recherche (CDR) ainsi que leurs antennes, met à la disposition des communautés universitaire et professionnelle des bases de données de très grande qualité.

La confidentialité et la sécurité occupant une place de choix dans cette organisation de renommée mondiale, nous avons dû nous soumettre au protocole de sécurité pour obtenir les accréditations nécessaires d'accès aux données. Cette procédure qui va de plusieurs semaines à quelques mois était l'un des obstacles, dans notre cas, à surmonter pour accéder aux données de l'enquête sur la dynamique du travail et du revenu (EDTR). Dans cette enquête, nous nous sommes intéressés au panel 4 qui est un suivi de personnes pendant six ans, de 2002 à 2007.

Le travail que nous présentons dans ce mémoire découle d'une collaboration en cours (Arango-Castillo *et al.*, 2019) basée sur l'EDTR. L'objectif principal de ces chercheurs est d'analyser les effets de la récession de 2008 sur les périodes de chômage. Au cours de cette enquête, un certain nombre de questions se sont posées concernant la meilleure façon d'analyser les périodes de chômage (ou d'autres types de données sur les périodes) à partir des données de panel d'enquête.

En utilisant l'EDTR à titre d'illustration, nous commençons dans ce chapitre en présentant

les données de panel d'enquête et les difficultés qui surviennent lorsqu'on utilise ces données pour faire une analyse longitudinale des périodes de temps. Ici, notre application concerne les périodes d'emploi et de chômage.

Dans ce chapitre, nous présentons la base de données de l'EDTR dans la Section 1.1. Étant donné que l'échantillon de l'EDTR est sélectionné à partir de l'enquête sur la population active (EPA), nous présentons également des détails pertinents concernant cette dernière à la Section 1.1.1. La Section 1.3 servira à présenter les difficultés liées à la base de données de l'EDTR utilisée pour notre analyse tandis que dans la Section 1.2, une analyse descriptive des données de cette enquête sera faite. À cette fin, l'accent sera mis sur les données du panel 4 de l'EDTR qui ont été collectées entre 2002 et 2007 (Statistique Canada, 2019) [1].

L'EDTR est une enquête longitudinale, à participation volontaire, par panels dont les données proviennent de l'EPA. Ces enquêtes sont menées auprès des ménages par Statistique Canada. Une particularité de l'EDTR se trouve dans le fait que les échantillons de celle-ci sont issus de l'EPA (Statistique Canada, 2018). L'EPA est une enquête toujours en cours tandis que l'EDTR a été remplacée par l'Enquête sur la Consommation et le Revenu (ECR) en 2012 (Statistique Canada, 2018).
Comme mentionné dans l'introduction, bien que l'EDTR soit le point focal des données de ce mémoire, les résultats de ce document sont généralisables à d'autres types de données de panel d'enquête. Nous citons quelques exemples d'enquêtes de données de panel :

— Penn World Table : ensemble de jeu de données pour mesurer le produit intérieur brut (PIB) de différents pays, Feenstra *et al.* (2015) ;

— British Household Panel Survey : enquête auprès des ménages au Royaume-Uni, Tay-

1. Ces analyses descriptives ont été réalisées au laboratoire du centre interuniversitaire québécois de statistiques sociales (CIQSS) de l'Université de Montréal (UdeM) affilié aux Centres de Données et de Recherche (CDR) de Statistique Canada.

lor *et al.* (1993) ;

— Panel Study of Income Dynamics : panel auprès des ménages aux États-Unis d'Amérique, Institute for Social Research (2021).

Un panel statistique (ou panel) est un type d'enquête longitudinale, type particulier d'enquête répétée à échantillon constant. On y collecte des données sur des unités constantes de l'échantillon à plusieurs occasions sur une « fenêtre » ou un intervalle de temps avec une date de début et une date de fin bien définies.

1.1 Les enquêtes

Bien que notre cadre de travail soit focalisé sur l'EDTR, nous décrirons l'EPA, dans la Section 1.1.1, dont la base de sondage sert dans l'élaboration des échantillons de la première citée. Les informations dans cette section sont essentiellement inspirées des excellents recueils Statistique Canada (2017) et Statistique Canada (2018). Dans notre présentation, nous utilisons une terminologie de base issue des méthodes d'enquête pour lesquelles Lohr (2009) et Ardilly (2006) sont de bonnes références.

1.1.1 Enquête sur la population active (EPA)

Conçue au sortir de la Seconde Guerre mondiale, l'EPA est une enquête menée par Statistique Canada auprès des ménages. D'abord trimestrielle, elle devint mensuelle au début des années cinquante. Elle avait pour objectifs initiaux de répartir la population en âge de travailler en trois catégories formant une partition de celle-ci (les personnes occupées, les chômeurs et les inactifs) et de fournir des données fiables et actuelles sur le marché du travail notamment sur le taux d'activité, le taux d'emploi et le taux de chômage. C'est une enquête à participation obligatoire couvrant 98% de la population dans les provinces (des personnes âgées de 15 ans et plus).

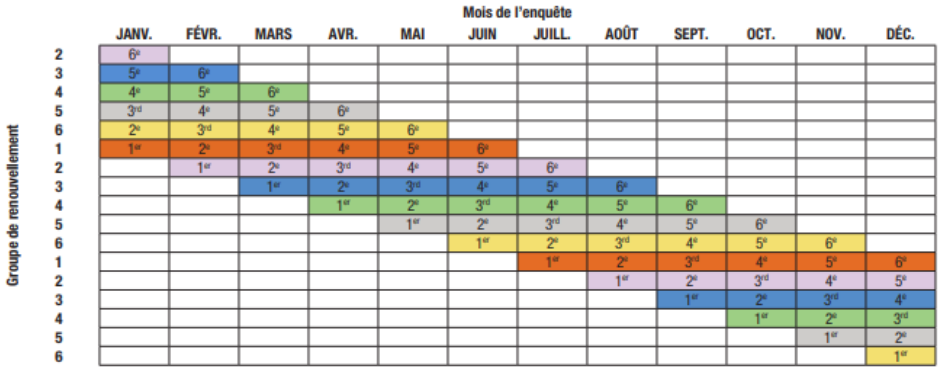
L'EPA utilise un plan de sondage stratifié par grappes à deux degrés sauf pour l'Île-du-Prince-Édouard du fait de sa caractéristique particulière. Pour cette dernière, le plan utilisé est l'échantillonnage à un degré. Les autres provinces sont réparties en régions qui constituent les grandes strates du premier degré d'échantillonnage : ce sont les unités primaires d'échantillonnage (UPE). Une fois l'échantillon d'UPE constitué, une liste des logements est établie dans chacune de ces UPE et un nouvel échantillon y est choisi. Les logements constituent ainsi les unités secondaires d'échantillonnage (USE).

Toutes les personnes des logements (ménages) sélectionnés dans la population-cible constituent l'échantillon de l'EPA et demeurent dans celui-ci pour une période de six mois. Après six mois, les ménages sortants sont remplacés par des logements de la même UPE ou d'une UPE semblable si l'UPE précédente a été éliminée et remplacée. Il en résulte un chevauchement des cinq sixièmes de l'échantillon d'un mois à l'autre : on dit que l'EPA utilise des groupes de renouvellement [2]. Cette caractéristique fait la particularité de l'EPA. La rotation de l'échantillon permet l'amélioration de la qualité des estimations et la diminution du fardeau du répondant : l'interview de l'EPA est réalisée une fois le mois pour un temps d'entrevue de moins de dix minutes.

Un autre avantage important du renouvellement se trouve dans le fait que l'échantillon obtenu de cette technique reflète la population actuelle, éliminant ainsi le potentiel biais de sélection qui serait apparu s'il n'y avait pas eu de renouvellement.

2. On parle aussi de rotation de l'échantillon.

Figure 1.1: Groupe de renouvellement de l'EPA.



La Figure 1.1 provient de la Figure 2.1 du rapport Statistique Canada (2017, p. 26). Sur celle-ci, un code de couleur est utilisé pour indiquer de quel groupe de renouvellement les logements sont issus. La couleur orange par exemple, correspond au premier groupe de renouvellement : groupe sélectionné en janvier pour la première fois et prenant fin en juin. Les logements de ce groupe seront remplacés par des logements de même UPE en juillet et demeureront dans l'échantillon pour une période de six mois, jusqu'en décembre avant d'être à nouveau remplacés par des logements en janvier. Les chiffres dans chaque cellule correspondent au nombre de mois pendant lesquels le groupe de renouvellement fait partie de l'échantillon de l'EPA depuis sa sélection. Cette procédure permet le renouvellement d'un sixième de l'échantillon chaque mois.

La collecte des données se tient tous les mois, habituellement pendant la semaine de référence : semaine qui correspond à celle contenant le 15^{e} jour du mois. L'Annexe A.1 contient plus de détails sur la façon dont les données de l'EPA ont été recueillies.

Dans la Section suivante 1.1.2, nous présentons les données de l'Enquête sur la dynamique du travail et du revenu (EDTR).

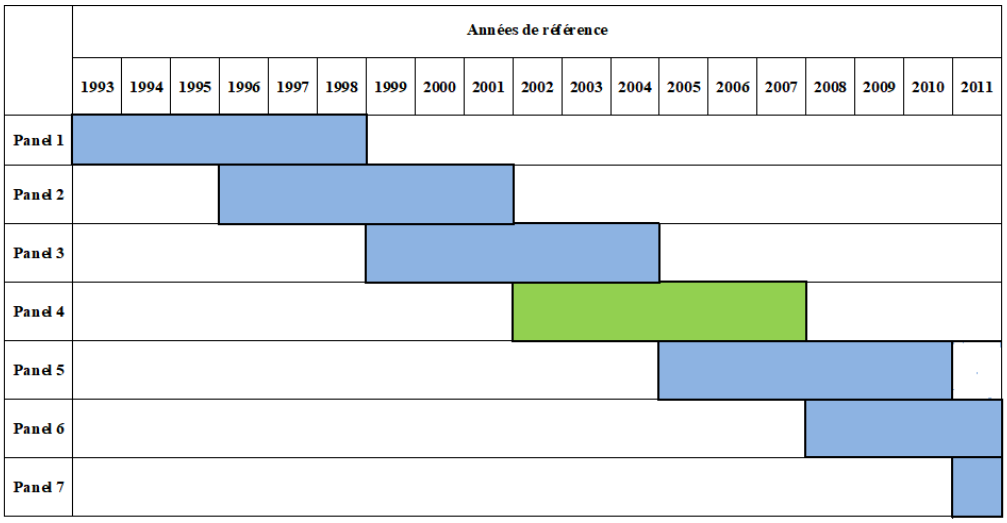
1.1.2 Enquête sur la dynamique du travail et du revenu (EDTR)

Introduite pour la première fois en 1993, l'EDTR est une source appréciable de données sur le revenu pour les familles, les ménages et les individus au Canada. Il s'agit d'une enquête longitudinale, à participation volontaire, qui donne une dimension supplémentaire aux enquêtes traditionnelles sur l'activité et le marché du travail : les changements vécus par les individus et les familles au fil du temps. L'un des objectifs principaux de l'enquête est de comprendre le bien-être économique des Canadiens et les changements ayant des incidences dans le temps sur les conditions de vie des personnes. Ainsi, en plus de fournir un portrait ponctuel, « une image » de la population canadienne, l'EDTR fournit un portrait évolutif, « un film » de cette population.

Composé de deux panels suivis sur une durée de six ans consécutifs, l'échantillon de l'EDTR est obtenu des répondants issus de la base de sondage de l'EPA. De ce fait, l'EDTR partage plusieurs caractéristiques avec cette dernière ; notamment le même plan de sondage. Elle couvre, comme l'EPA, environ 98% de la population du Canada soit toutes les personnes à l'exception des personnes vivant dans les réserves indiennes, dans les Territoires du Nord-Ouest, du Nunavut et du Yukon, et des personnes vivant dans des établissements pénitentiaires ou sanitaires et dans des casernes militaires pour plus de six mois.

Les ménages sortants de l'EPA aux deux premiers mois de la période de référence de l'EDTR (janvier et février) sont choisis pour faire partie de la base de sondage de cette dernière. Parmi ceux-ci, les ménages ayant répondu pendant leur dernier mois respectif de participation à l'EPA sont sélectionnés. L'échantillon de l'EDTR se compose, depuis le 1er janvier 1996, de deux panels dont chacun compte près de 17 000 ménages (soit environ 40 000 personnes).

Figure 1.2: Chevauchement des panels de l'EDTR.



La Figure 1.2, qui est inspirée de LaRoche (2007, p. 8), donne un aperçu de la sélection des panels et du chevauchement entre ceux-ci. Comme on peut le voir, le panel 4 part de 2002 à 2007 et chevauche le panel 3 (de 2002 à 2004) ainsi que le panel 5 (de 2005 à 2007). Il est bon de remarquer que le panel 4 débute lorsque le panel 2 s'achève de sorte qu'il n'y a que deux panels qui se chevauchent sur une année donnée.

Chaque panel représente la population cible au 31 décembre de l'année précédant son introduction. Par exemple, le panel 4 représente la population cible au 31 décembre 2001 et est introduit en janvier 2002 tandis que le panel 5 représente celle au 31 décembre 2004 alors qu'introduit en 2005. Deux panels consécutifs se chevauchent sur trois ans. Dans le cas des panels 4 et 5, on peut apercevoir que le chevauchement se tient sur les années 2005 à 2007.

Jusqu'en 2005, les personnes sélectionnées dans le cadre de l'EDTR étaient interviewées

en janvier, pour la collecte d'informations relatives à l'activité du marché du travail, à l'éducation, au logement, puis en mai, relativement à l'information sur le revenu [3]. Depuis 2005, ces deux entrevues ont été fusionnées en une seule dans le but de réduire les coûts de collecte de données et le fardeau des répondants. L'unique entrevue, qui se tient chaque année entre janvier et mars, est administrée au téléphone directement auprès de toute personne âgée de 16 ans ou plus du ménage sélectionné par un intervieweur assisté par ordinateur. Les informations sur le revenu et sur les activités liées au marché du travail relatives à l'année précédente sont recueillies.

Contrairement à l'EPA qui ne comptait que les personnes de 15 ans et plus dans sa population cible, l'échantillon de l'EDTR prend en compte tous les membres du ménage sélectionné, indépendamment de leurs âges. Ces personnes demeurent dans l'échantillon pour la durée du panel, soit de six années consécutives, quel que soit l'évènement qui pourrait leur arriver (déménagement, admission dans une institution, décès, etc.). L'échantillon longitudinal est donc défini au moment de la constitution du panel c'est-à-dire au 1er janvier de l'année de référence et aucune autre personne ne peut devenir membre de l'échantillon longitudinal durant l'existence de ce panel : c'est un échantillon constant (Statistique Canada, 2003).

C'est pour cet échantillon longitudinal de l'EDTR que nous souhaitons analyser les périodes d'emploi et de chômage. Un ensemble de poids longitudinaux [4] et de poids bootstrap longitudinaux [5] sont fournis par Statistique Canada avec l'échantillon. En fait, avec l'EDTR, trois types de poids et leur correspondant bootstrap sont fournis (LaRoche, 2007). Les

3. Le mois de mai a été choisi afin de faciliter la collecte sur le revenu du fait de la complétion des déclarations d'impôt.

4. Les poids longitudinaux sont utilisés pour l'estimation des paramètres de population au moment de la sélection de l'échantillon.

5. Les poids bootstrap longitudinaux sont utilisés pour l'estimation de la variance des estimateurs.

poids de bootstrap permettent le calcul de la variance des différents estimateurs.

Chaque type de poids décrit un échantillon différent. Brièvement :

- il y a l'échantillon longitudinal (avec les poids longitudinaux et les poids de bootstrap longitudinaux) mentionné ci-dessus. Il s'agit de l'échantillon initial sélectionné à partir de l'EPA. Les poids longitudinaux ainsi que les poids bootstrap longitudinaux changent d'année en année afin de refléter la population à l'année d'introduction de l'échantillon qui est sujette au phénomène d'attrition (LaRoche, 2007). Les poids longitudinaux sont présentés à l'Annexe A.2.1 ;
- vient ensuite l'échantillon transversal avec les poids correspondants fournis aussi par Statistique Canada. Cet échantillon comprend les personnes longitudinales [6] ainsi que des cohabitants [7] (Naud, 2004). Les poids transversaux sont discutés à l'Annexe A.2.2 ;
- comme décrit à la Figure 1.2, il y a une période de chevauchement entre les panels. Statistique Canada en a profité pour fournir des poids longitudinaux avec panels combinés afin de prendre avantage des échantillons des panels se chevauchant sur les trois années communes, en doublant la taille de l'échantillon résultant des échantillons des deux panels (Naud, 2004). Par exemple, des poids longitudinaux avec panels combinés sont fournis pour les panels 4 et 5 sur les années 2005 à 2007. Les poids longitudinaux avec panels combinés sont présentés dans l'Annexe A.2.3.

À l'Annexe A.2, nous fournissons une description détaillée des poids longitudinaux, transversaux et longitudinaux avec panels combinés (nous omettons la discussion sur les poids bootstrap correspondants). Bien que notre objectif dans ce mémoire porte sur la façon de traiter les périodes de temps sectionnées par le début et la fin de l'intervalle dans un pro-

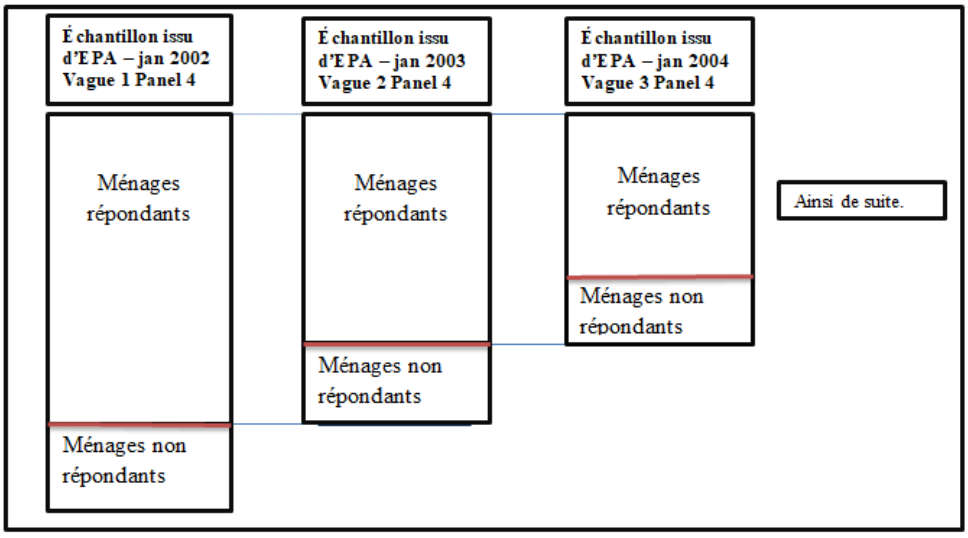
6. Les personnes longitudinales sont celles domiciliées dans les ménages sélectionnés en début de panel (au sortir de l'EPA).

7. Les cohabitants sont les personnes qui se sont jointes, au cours des années, aux ménages des personnes longitudinales.

cessus de renouvellement alterné causé par la collecte des données de panel, il est important de réaliser que les poids doivent être utilisés lors de l'analyse des données de l'EDTR. À partir des travaux antérieurs d'Arango-Castillo *et al.* (2019), nous nous rendons compte que l'utilisation des poids est particulièrement importante lors de l'estimation de la variance des estimateurs par exemple.

L'utilisation d'une enquête longitudinale offre plusieurs avantages mais encore plus d'inconvénients. L'un des problèmes survenant pour les données de type panel est l'attrition. L'attrition est un phénomène qui se caractérise par la diminution de la taille de l'échantillon au fil du temps. À Statistique Canada, le traitement utilisé pour compenser cette perte d'effectif dans l'échantillon consiste à faire de la repondération chaque année. Cette méthode permet de remédier à ce problème de sorte que l'échantillon demeure représentatif de la population canadienne à l'année de référence où le panel a été constitué.

Figure 1.3: Attrition dans un échantillon longitudinal.



La Figure 1.3, inspirée de LaRoche (2007, p. 9), donne une illustration du phénomène d'at-

trition sur un échantillon. Elle montre les ménages de l'EPA sélectionnés pour faire partir de l'échantillon longitudinal du panel 4 entre 2002 et 2007. En janvier de chaque nouvelle année, les ménages non répondants sont exclus du panel et une nouvelle pondération est établie afin que l'échantillon « réduit » reflète la population-cible lors de l'introduction du panel.

1.2 Analyse descriptive des données

Dans cette section, nous donnons quelques statistiques sur l'ensemble de données. En mettant notamment l'accent sur les caractéristiques liées aux périodes de chômage.

La plupart des éléments présentés dans cette section ont été préparés et partagés avec Arango-Castillo *et al.* (2019). Les codes ont été réalisés au Centre de données et de recherches (CDR) de l'Université de Montréal (UdeM) à Montréal à l'aide du logiciel SAS, dont l'accès fut autorisé après s'être soumis aux procédures de sécurité et d'éthique auprès de Statistique Canada.

Comme indiqué au début de ce chapitre, notre objectif est d'utiliser les données du panel 4 pour illustrer les caractéristiques typiques sur les périodes d'emploi et de chômage dans l'EDTR. Bon nombre des caractéristiques sur lesquelles nous insistons sont liées aux données manquantes, ce qui complique l'analyse des données de l'EDTR.

Tableau 1.1: Nombre d'individus du panel 4 de l'EDTR selon qu'ils ont ou non au moins une période de chômage.

Nombre de périodes de chômage	Panel 4	
	Effectif	Fréquence (%)
Aucune	25 909	61.35
Au moins une	16 323	38.65
Total	42 232	100

Dans le Tableau 1.1, nous présentons une vue de l'ensemble des données du panel 4 soumises à notre étude relativement au nombre de périodes de chômage. De cela, il ressort que 25 909 individus n'ont manifesté aucune période de chômage dans le panel 4. Cela représente 61.35% de l'ensemble des individus de ce panel (ces personnes n'ont donc vécu que des périodes de travail selon le modèle que nous étudions). *A contrario*, 38.65% des effectifs dudit panel ont vécu au moins une période de chômage.

Plus de détails sur ces périodes de chômage seront fournis dans les lignes suivantes. Déjà à partir du Tableau 1.1, nous voyons qu'en ce qui concerne l'emploi, il semble y avoir au moins deux sous-groupes avec des comportements différents. En effet, dans Arango-Castillo *et al.* (2019), il a été décidé de se concentrer sur la sous-population ayant au moins une période de chômage.

Tableau 1.2: Nombre d'individus du panel 4 de l'EDTR selon le nombre de période de chômage.

Nombre de périodes de chômage	Panel 4		
	Effectif	Fréquence (%)	Fréquence cumulée
1	9 934	60.86	60.86
2	3 516	21.54	82.40
3	1 514	9.28	91.68
4	759	4.65	96.33
5 et plus	600	3.68	100
Total	16 323	100	

Le Tableau 1.2 est un regard minutieux porté sur le nombre de périodes de chômage pour chaque individu. Dans ce tableau, nous nous intéressons donc aux individus ayant vécu au moins une perte d'emploi. On voit dans cette catégorie que, plus de la majorité des personnes ont vécu exactement une période de chômage, soit 60.86%.

Sur cent personnes présentes dans l'échantillon du panel 4 et ayant vécu minimalement un épisode de chômage, un peu plus de 17% ont expérimenté 3 périodes de chômage ou plus.

Les Tableaux 1.1 et 1.2 ont décrit l'échantillon longitudinal du panel 4 dont les individus ont, ou non, vécu au moins une période de chômage (Tableau 1.1) ; et parmi ceux ayant expérimenté au moins une période de chômage, le Tableau 1.2 indique les effectifs des individus ayant une période de chômage ou plus.

Tableau 1.3: Nombre de périodes de chômage selon que la date de début et la date de fin étaient connues ou non.

Panel 4	Dates de début			Dates de fin		
	Connues	Inconnues	Total	Observées	Non observées	Total
Nombre de périodes	19 763	8 195	27 958	14 742	13 216	27 958
Fréquence (%)	70.63	29.31	100	52.73	47.27	100

Dans le Tableau 1.3, nous illustrons toujours à l'aide de l'échantillon longitudinal du panel 4, les nombres de périodes (de chômage ou d'emploi) pour lesquelles certaines transitions *Chômage -> Emploi* ou *Emploi -> Chômage* sont manquantes. En clair, le Tableau 1.3 présente les périodes de chômage dont les dates de début et/ou de fin sont connues ou non.

On peut lire du Tableau 1.3 que, sur 27 958 périodes de chômage dans le panel 4, 29.31% ont des dates de début de période de chômage inconnu. Concernant la date de fin, cette proportion est de 47.27%. Les personnes interviewées ont une plus grande propension à se rappeler les dates d'entrée en chômage que les dates de fin de chômage.

Évidemment, le fait d'avoir autant de dates de transition manquantes peut rendre difficile l'estimation des paramètres d'un processus de renouvellement alterné. L'Annexe A.3 présente les données par périodes de chômage dans le panel 4 de l'EDTR selon la date de début et la date de fin.

Dans ce mémoire, nous considérons un modèle de renouvellement alterné pour les durées d'emploi et de chômage dans un échantillon longitudinal d'un panel. Notre travail est

motivé par l'EDTR. Bien que l'accent ne soit pas mis sur l'analyse populationnelle à travers l'utilisation des poids longitudinaux, nous présentons des informations agrégées de ces poids pour le panel 4 de l'EDTR dans le Tableau 1.4 afin de montrer la nature complexe de telles données. Les modalités 07 et 08 de la variable *resp*99, `statut de réponse`, entraînent manifestement des données de périodes incomplètes mais les autres modalités peuvent aussi induire des périodes de temps incomplètes.

Tableau 1.4: Nombre d'individus longitudinaux de l'échantillon de l'EDTR par statut de réponse, poids longitudinal et année dans le panel 4.

Panel 4								
Années	Poids (w)	Statut de réponse				Total partiel	Total	Taux de réponse
		(01)	(02-06)	(07)	(08)			
2002	w = 0	7 936				7 936	42 232	0.81
	w > 0	33 792	504			34 296		
2003	w = 0	6 721		361	~ 15 [8]	$\sim 7\,097$	42 232	0.83
	w > 0	34 115	1 021			35 136		
2004	w = 0	3 300		5 835	28	9 163	42 232	0.78
	w > 0	31 662	1 407			33 069		
2005	w = 0	2 568		7 953	35	10 556	42 232	0.73
	w > 0	29 912	1 764			31 676		
2006	w = 0	2 866		9 078	37	11 981	42 232	0.72
	w > 0	28 178	2 073			30 251		
2007	w = 0	3 248		9 859	37	13 144	42 232	0.69
	w > 0	26 708	2 380			29 088		

Le Tableau 1.4 présente une répartition des effectifs totaux de l'échantillon de l'EDTR dans

le panel 4 selon le statut de réponse. La variable correspondante est déclinée par *resp*99 ayant pour modalités :

- (01), si l'individu répondant se trouve toujours dans l'enquête et vit dans l'une des dix provinces canadiennes ;
- (02-06), si l'individu est une personne longitudinale [9]
  - (02) résidant au Yukon, Territoires du Nord-Ouest ou Nunavut,
  - (03) résidant en dehors du Canada et de la zone continentale des États-Unis,
  - (04) résidant dans la zone continentale des États-Unis,
  - (05) institutionnalisée pour plus de six mois,
  - (06) décédée ;
- (07), si l'individu est perdu de vue ou a abandonné l'enquête, éliminé de l'échantillon (refus ferme, incapable de dépister, etc.) ;
- (08), si l'individu est identifiée plus tard comme n'étant pas une personne réelle.

Les valeurs précédées du symbole $\sim$ dans le Tableau 1.4 indiquent une approximation pour préserver la confidentialité des données.

Les poids (w) correspondent aux poids longitudinaux. Rappelons que ces poids ont été conçus afin que l'échantillon de l'EDTR représente la population-cible lors de l'introduction du panel dans l'enquête : les poids longitudinaux du panel 4 (de 2002 à 2007), permettent à l'échantillon de représenter la population-cible de 2002. Voir la section relative à la pondération longitudinale (Annexe A.2.1).

Les poids longitudinaux nuls caractérisent la non-réponse, c'est-à-dire que dans le ménage aucun individu n'a répondu à l'interview. À l'inverse, ce poids est non nul si au moins un individu est répondant. Ainsi un individu peut avoir un poids nul lors d'une année de présence dans le panel, puis un poids non nul lors d'une autre année de présence selon que

9. Pour rappel, une personne est longitudinale si elle est domiciliée dans le ménage sélectionné en début de panel (au sortir de l'EPA).

le ménage auquel cet individu appartient soit non répondant ou répondant respectivement à l'enquête. De ce fait, un individu longitudinal décédé peut avoir un poids non nul si le ménage auquel il appartenait est répondant.

Dans un ménage répondant, il existe deux classes de poids : celle des enfants (15 ans et moins) et celle des adultes (16 ans et plus). Au sein de ce ménage, les poids longitudinaux sont les mêmes à l'intérieur de chacune des classes pour chacun des individus. Plus de détails sont fournis dans l'Annexe A.2.1.

Le taux de réponse est le quotient du nombre d'individus ayant répondu par le nombre total d'individus dans l'échantillon : division du nombre de répondants (dont le poids $w > 0$) par le nombre d'individus longitudinaux (42 232 pour le panel 4).

À travers l'effectif des individus de poids nul qui croît au fil des années de participation à l'enquête (à l'exception de l'année 2003) et du taux de réponse qui décroît, on observe le phénomène d'attrition évoqué précédemment. Pour corriger cela, une repondération est effectuée chaque année.

1.3 Complexités liées à la base de données

Comme mentionné dans les sections précédentes, les échantillons de l'EDTR sont issus de l'EPA avec laquelle ils partagent le plan d'échantillonnage. C'est un plan d'échantillonnage complexe : stratifié par grappes et à plusieurs degrés. Ces plans d'échantillonnage ont pour principaux objectifs de réduire les coûts de collecte des données tout en améliorant la qualité des estimations par réduction de la variance des estimateurs. Le plan d'échantillonnage et les poids correspondants ne sont pas les seules complexités que l'on rencontre dans cette base de données.

Même en restreignant notre attention à une analyse de l'échantillon longitudinal sur un modèle de renouvellement alterné entre les états *Emploi* et *Chômage*, nous observons de nombreuses complexités :

— les informations sur les individus sont collectées de manière rétrospective, c'est-à-dire, pour une année donnée t, les intervieweurs de Statistique Canada collectent les informations relatives à l'année $t-1$. De ce fait, certaines personnes interviewées ont quelques fois de la difficulté à se souvenir de certaines informations demandées ;

— inclure les poids longitudinaux et les poids bootstrap longitudinaux dans le modèle : le fait que les poids changent chaque année et qu'ils décrivent des ménages et non des individus compliquent davantage leur utilisation. De plus, il faut utiliser les poids bootstrap pour l'estimation de la variance puisque certaines informations sont masquées pour des raisons de confidentialité ;

— les périodes d'emploi et de chômage sectionnées par le début du panel et la fin du panel : les périodes (d'emploi ou de chômage) les plus longues ont plus de chance d'être interceptées par le début et la fin du panel. Ce phénomène plus connu sous le « paradoxe de l'inspection » (Grimmett et Stirzaker, 2001) complique la situation par le fait que plusieurs dates de début des périodes de temps avant l'introduction dans le panel(les temps calendaires initiaux) sont manquantes. La complexité décrite dans ce point sera notre objectif dans la suite de ce mémoire ;

— données incomplètes : les temps de transition d'un état à l'autre à l'intérieur du panel sont souvent des données manquantes ;

— plusieurs covariables sont fournies et nombreuses parmi elles sont dépendantes du temps. Par exemple, des covariables comme l'éducation et la région peuvent changer au cours du panel de 6 ans. On peut aussi avoir des données manquantes pour ces variables ;

— comme mentionné dans la Section 1.1.1, il peut y avoir des individus en emploi, d'autres en recherche active d'emploi et d'autres encore non en recherche active

d'emploi (actifs, chômeurs et inactifs) alors que le modèle utilisé dans ce mémoire est dichotomique (en emploi et non en emploi). Pour obtenir une description réaliste de la population, on peut utiliser une sorte de modélisation mixte pour tenir compte des différents sous-groupes de la population et de leurs différents comportements face au travail. C'est la raison pour laquelle Arango-Castillo *et al.* (2019) ont limité leur analyse aux personnes ayant vécu au moins une période sans emploi ;

— il existe différentes échelles de temps dans l'analyse : dans le cas de l'EDTR, les transitions entre l'emploi et le chômage se font en fonction du temps calendaire (par exemple pendant la récession de 2008) et qui diffère de l'échelle de temps sur lequel sont observés l'âge et la durée des périodes d'emploi et de chômage. Cependant, la durée des périodes d'emploi et de chômage est très probablement fonction du temps calendaire.

Traiter toutes ces complications à la fois dépasse le cadre de ce mémoire de maîtrise. Bien que bon nombre des complications énumérées ci-dessus rendraient très difficile l'application d'un modèle de renouvellement alterné, dans les Chapitres 2, 3 et 4, nous nous concentrons uniquement sur les manières de traiter les périodes qui sont sectionnées par le début et la fin du panel. Notre modèle d'état alterné ainsi que les données de panel entrent dans ce que l'on appelle communément le processus de renouvellement alterné censuré par intervalle (PRACI) ou *window censored alternating renewal process (WCARP)* en anglais.

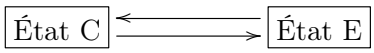
Les données étant recueillies dans une fenêtre de six ans (entre le 1er janvier 2002 et le 31 décembre 2007 pour le panel 4), les périodes de temps les plus longues commençant avant le début du panel sont observées : on parle de troncature à gauche.
En fin de panel, il arrive que la plupart des données soient incomplètes. Soit parce que l'individu faisait encore partie de l'étude mais n'a pas subi l'évènement, soit parce qu'il a été perdu de vue : on parle alors de censure à droite. La troncature à gauche et la censure à droite sont les formes les plus répandues de troncature et de censure respectivement.

CHAPITRE 2
MODÈLE DE RENOUVELLEMENT ALTERNÉ

Dans ce chapitre, nous présentons le modèle de renouvellement alterné et les différentes hypothèses faites pour les périodes tronquées par le début et celles sectionnées par la fin du panel. Comme mentionné dans l'introduction de ce mémoire, nous présentons trois approches pour traiter ces périodes tronquées en portant une attention particulière au type de données dans chaque approche.

1. La première approche que nous désignons par *Suppression* est celle pour laquelle les premières périodes de temps tronquées par le début du panel sont supprimées (non prises en compte dans l'élaboration de la vraisemblance) et les dernières périodes de temps sectionnées par la fin du panel sont traitées en utilisant la censure à droite non informative. Cette approche est abordée dans Kovačević et Roberts (2007).
2. Dans la deuxième approche désignée par *Conditionnelle*, les dates de début (c'est-à-dire les temps calendaires initiaux) des périodes d'emploi et de chômage tronquées par le début du panel sont connues (et de ce fait la première période est entièrement connue). Nous utilisons alors une approche par vraisemblance conditionnelle pour le traitement de ce type de période. Il s'agit d'une méthodologie très courante et abordée au Chapitre 7 de (Cook et Lawless, 2018).
3. Pour la troisième approche, désignée dans ce mémoire par *Stationnarité*, les dates de début des périodes tronquées par le début du panel sont inconnues. L'hypothèse de stationnarité est faite afin de développer une vraisemblance dans le but d'apporter un traitement à de telles périodes incomplètes. Cette approche est le principal sujet présenté dans Alvarez (2006).

Figure 2.1: Espace d'états pour un modèle de renouvellement alterné et les différentes transitions possibles entre ces états indiquées par les flèches.



La Figure 2.1 illustre un processus de renouvellement alterné typique où les individus pivotent entre deux états. Nous utilisons les termes périodes, périodes de temps et temps de séjour de manière interchangeable pour indiquer les temps passés dans chaque état et nous supposerons des distributions pour ces variables aléatoires de temps passés dans chaque état. Pour illustrer l'EDTR, dans la Figure 2.1, l'état C représente *Chômage* tandis que l'état E représente *Emploi*.

Nous commençons, dans la Section 2.1, par présenter le processus de *renouvellement alterné pur*. Bien que ce modèle s'inscrive dans le cadre plus large des modèles multi-états (Cook et Lawless, 2018), nous simplifions notre présentation en utilisant uniquement les notations typiquement utilisées pour le modèle de renouvellement alterné (voir par exemple, Alvarez (2006) et Ross (1996)). Dans la Section 2.2, nous considérons un modèle de *renouvellement alterné retardé*. Ceci est nécessaire car la date du début du panel d'enquête est arbitraire au regard du processus de renouvellement alterné de l'emploi et du chômage dans la population. Ainsi, nous souhaitons modéliser un processus de renouvellement alterné retardé avec les données du panel de l'enquête. La Section 2.3 sert à présenter des résultats importants et bien connus pour les processus de renouvellement stationnaire. Nous nous appuyons sur ces résultats en utilisant la stationnarité pour traiter les temps tronqués par le début du panel (approche *Stationnarité*).

Nous utilisons des densités exponentielles pour modéliser les distributions des durées d'emploi et de chômage. Cette hypothèse signifie que dans les chapitres suivants, notre modèle

de renouvellement alterné est une chaîne de Markov continue homogène à deux états. Pour le moment cependant, nous commençons par la Section 2.1 sans imposer de densités particulières pour les distributions des temps d'emploi et de chômage, ce qui a pour avantage d'être flexible et d'envisager plusieurs types de lois. Puisque le temps est continu, nous supposons d'abord qu'il s'agit de densités. La loi exponentielle étant sans mémoire, cette présentation plus générale permet de mettre en évidence les hypothèses de chacune des trois approches avant de simplifier les vraisemblances en utilisant des lois exponentielles

2.1 Modèle de renouvellement alterné pur

Dans cette section et dans la suite, nous considérons deux distributions marginales distinctes f_C et f_E pour les temps de séjour dans les états C et E respectivement. Nous faisons l'hypothèse que tous les temps de séjour sont *iid* c'est-à-dire indépendants et identiquement distribués. Typiquement, le temps passé dans l'état C est représenté par la variable Z et le temps passé dans l'état E est noté Y :

$$Z \sim f_C(z), \qquad Y \sim f_E(y). \tag{1}$$

Les processus de renouvellement alternés (voir Alvarez (2006) et Ross (1996) par exemple) sont généralement présentés avec des temps de renouvellement $X_j = Z_j + Y_j$. Bien que les temps de renouvellement X_j soient supposés *iid*, les Z_j et Y_j pourraient être dépendants. Habituellement, la distribution conjointe de Z_j et Y_j est notée Q. Lorsque l'indice de l'individu i est inclus dans la notation, nous écrivons $X_{ij} = Z_{ij} + Y_{ij}$ pour le j^{e} temps de renouvellement du i^{e} individu.

Dans notre présentation, nous supposons que les Z_{ij} et Y_{ij} sont indépendants pour un individu i. C'est plus simple comme hypothèse *iid* pour les paires (Z_j, Y_j). Dans les modèles

1. Souvent, pour les processus de renouvellement alternés, la fonction de répartition de la variable aléatoire Z est notée H avec fonction de densité h tandis que celles de la variable aléatoire Y sont notées respectivement G et g. Ici, nous avons choisi d'utiliser f_C et f_E au lieu de h et g respectivement.

plus sophistiqués, on pourrait autoriser une dépendance entre les temps Z_{ij} et Y_{ij} car il est très plausible que les temps, d'emploi et de chômage par exemple, pour le même individu i soient dépendants. On pourrait, à un autre degré, supposer la dépendance entre les renouvellements X_{ij}, $j = 1, 2, \ldots$ pour le même individu i.

Pour alléger la notation, dans la suite, nous abandonnons l'indice i et présentons d'autres constructions couramment utilisées pour décrire les processus de renouvellement alternés. Désignons par S_k le cumul des k temps de renouvellement X_j pour un individu donné :

$$S_k = \sum_{j=1}^{k} X_j,$$

avec la convention que $S_0 = 0$.

Désignons par N le processus de comptage (2.1) indiquant le nombre de temps de renouvellement X_j se trouvant dans l'intervalle de temps $[0, t]$ pour un individu donné :

$$N(t) = \sum_{k=1}^{\infty} \mathbb{1}_{[0,t]}(S_k). \tag{2.1}$$

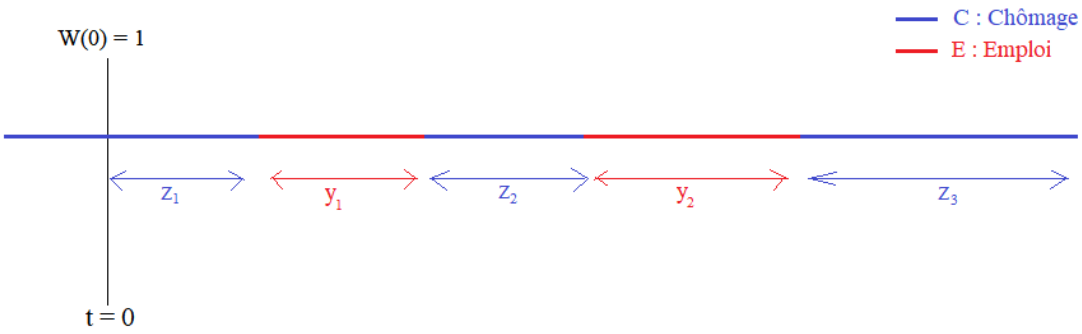
La fonction (2.2) indiquant l'état (C ou E) dans lequel se trouve le processus à l'instant t, pour un individu donné, est désignée par W :

$$W(t) = \mathbb{1}(S_{N(t)} + Z_{N(t)+1} > t). \tag{2.2}$$

Dans Alvarez (2006), S_k est appelé *processus de renouvellement* associé à $W(t)$ et $W(t)$ est désigné comme le *processus de renouvellement alterné* associé à S_k.

Lors d'une inspection rapide de l'équation (2.2), nous voyons qu'à l'instant t, $W(t) = 1$ si le processus de renouvellement alterné est dans l'état C et $W(t) = 0$ si le processus de renouvellement alterné est dans l'état E. Dans le processus de renouvellement alterné pur formalisé dans l'équation (2.2), tous les individus commenceraient au temps $t = 0$ dans l'état C.

Figure 2.2: Représentation d'un chemin dans un processus de renouvellement alterné pur.



La Figure 2.2 illustre un exemple de chemin dans un processus de renouvellement alterné pur. Les individus commencent dans l'état C avec $W(0) = 1$ au temps $t = 0$.

2.2 Modèle de renouvellement alterné retardé

Pour notre travail, nous allons considérer un processus de renouvellement alterné retardé, c'est-à-dire un processus de renouvellement alterné pur ayant débuté dans le passé au temps initial 0, mais pour lequel les observations sont prélevées plus tard au temps $t > 0$.

En pensant aux données de panel telles que l'EDTR, nous avons une population que nous échantillonnons au début du panel et suivons pendant toute la durée de celui-ci. Au début du panel, tous les individus seront dans l'état C (chômage, sans emploi) puis plus tard en un temps t, certains se retrouveront dans l'état E alors que d'autres seront dans C. Cela se prête bien à l'échantillon en notre possession pour analyse.

Le processus de renouvellement alterné retardé débute à un instant positif t (ce temps

représente le début de notre panel). Une séquence de renouvellements alternés $\{(Z_j^t, Y_j^t), j > 1\}$ est alors construite pour le processus retardé. En termes simples, pour le premier temps de séjour dans le panel, nous sectionnons et ne prenons pas en compte tout ce qui se trouve avant le début du panel, à gauche de t. Le processus de renouvellement alterné retardé est [2] :

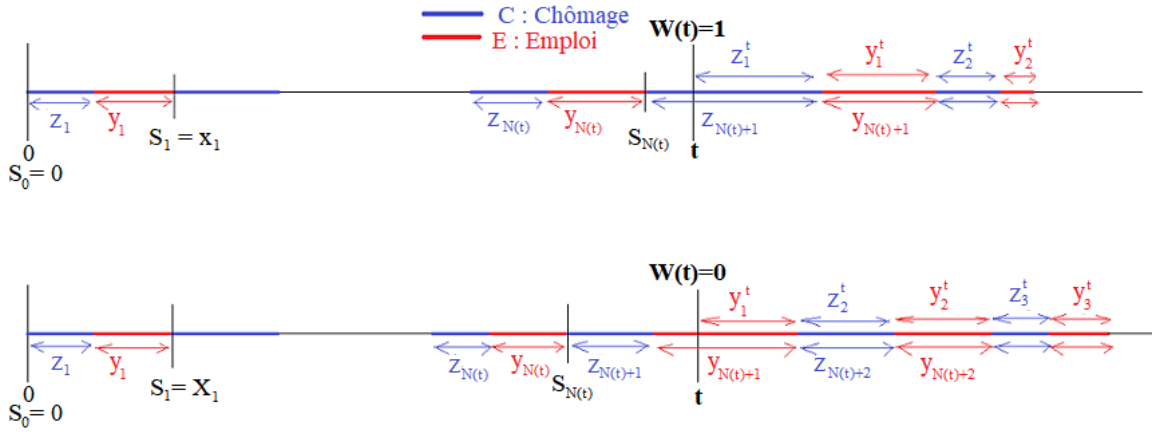
$$Z_1^t = (S_{N(t)} + Z_{N(t)+1} - t)^+$$
$$Z_j^t = Z_{N(t)+j},\ j = 2, 3, \ldots$$
$$Y_1^t = Y_{N(t)+1} - (t - S_{N(t)} - Z_{N(t)+1})^+$$
$$Y_j^t = Y_{N(t)+j},\ j = 2, 3, \ldots$$

où l'opérateur $(\eta)^+$ désigne le maximum entre 0 et η, avec $\eta \in \mathbb{R}$ par exemple, $(-2)^+ = 0$ et $(7)^+ = 7$.

C'est le processus de renouvellements alternés retardé $(Z_1^t, Z_j^t, Y_1^t, Y_j^t)$, $j = 2, 3, \ldots$ qui sera observé dans le panel. Dans la suite pour simplifier les notations, on laissera tomber le t en exposant et on écrira (Z_1, Z_j, Y_1, Y_j), $j = 2, 3, \ldots$. Pour cette nouvelle séquence de renouvellements alternés, nous allons donc redéfinir la notation en supprimant le t en exposant et redimensionner le repère de manière à ce que le processus débute à 0 et non à t.

2. Pour s'assurer que $W(t)$ est bien défini pour le processus de renouvellement alterné retardé, nous nous écartons légèrement de Alvarez (2006) et remplaçons Z_0 par Z_1 et Y_0 par Y_1.

Figure 2.3: Représentation de deux exemples de chemins d'un processus de renouvellement alterné retardé à l'instant t. L'un commence dans l'état C ($W(t) = 1$) et l'autre dans l'état E ($W(t) = 0$).



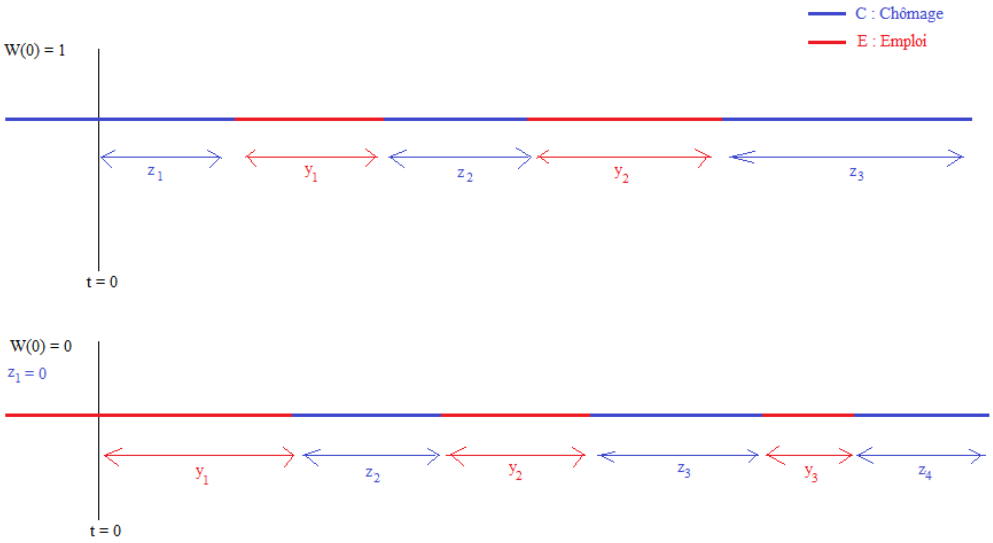
Dans le processus de renouvellement alterné retardé :

$$S_0 = 0, \tag{2.3}$$

$$S_k = \sum_{j=1}^{k} X_j, \tag{2.4}$$

où le premier temps de renouvellement retardé est $X_1 = Z_1 + Y_1$. Notons que Z_1 peut être nul, ce qui signifie que la première apparition dans le panel se fait dans l'état E et que la période correspondante a été sectionnée par le début du panel. Voir la Figure 2.4 avec $W(0) = 0$.

Figure 2.4: Représentation de deux exemples de chemins d'un processus de renouvellement alterné retardé dans le panel.



Sur la Figure 2.4, nous ne présentons que la partie après t de l'illustration du processus de renouvellement alterné retardé de la Figure 2.3 pour laquelle nous avons laissé tomber le t en exposant et avons redimensionné le repère afin de faire débuter ce « nouveau » processus en 0. En d'autres mots, nous avons laissé tomber toutes les périodes de temps à gauche de t qui ne sont pas interceptées par t et avons changé l'origine de notre repère.

Une fois de plus,

$$N(t) = \sum_{n=1}^{\infty} \mathbb{1}_{[0,t]}(S_k) \tag{2.5}$$

et

$$W(t) = \mathbb{1}(S_{N(t)} + Z_{N(t)+1} > t) \; . \tag{2.6}$$

2.2.1 Processus de renouvellement alterné retardé dans le panel

Pour chaque individu i, il y a quatre cas de figure qui peuvent se produire dans le panel $[0, T]$, où 0 correspond au début du panel tandis que T en est la fin. Ces cas de figure sont déterminés par le fait qu'un chemin d'échantillonnage commence soit dans l'état C, soit dans l'état E et se termine dans l'état C ou dans l'état E. Puisqu'il y a deux choix d'états au début du panel et encore deux choix d'états à la fin du panel, il y a donc $2^2 = 4$ cas à considérer.

Tableau 2.1: Synthèse des différents aboutissements en début $t = 0$ et fin $t = T$ de panel pour les différents parcours d'un processus de renouvellement alterné retardé.

Cas	$W(0)$	$W(T)$	Commentaires
1	0	1	Débute dans l'état E et prend fin dans l'état C
2	1	0	Débute dans l'état C et prend fin dans l'état E
3	1	1	Débute dans l'état C et prend fin dans l'état C
4	0	0	Débute dans l'état E et prend fin dans l'état E

Le Tableau 2.1 présente les divers cas pouvant se produire dans le panel $[0, T]$.

Pour le processus de renouvellement alterné dans le panel, nous introduisons τ, une variable qui compte le nombre de temps de renouvellement X_j dans le panel. Par exemple, si $\tau = 1$ alors la fin du panel s'est produite pendant le temps de séjour X_1 (avec deux cas de figure possibles : soit en Z_1, soit en Y_1). Si $\tau = 2$, la fin du panel s'est produite en X_2 (ici aussi, avec deux cas de figures pour la fin : soit en Z_2, soit en Y_2), etc. En général, si $\tau = j$ alors la fin du panel s'est produite pendant le temps de séjour X_j (avec une fin en Z_j ou en Y_j). Autrement dit, τ est le nombre minimal de renouvellement X_j pour lequel $S_j > T$. Nous noterons $\tau = \inf\{j : S_j > T\}$. Avec la notation qui vient d'être introduite pour le processus de renouvellement alterné retardé, dans la Section 2.4, nous

présentons une vraisemblance pour le processus de renouvellement alterné retardé dans le panel. Nous présentons une vraisemblance (2.11) sous une forme générale, puis nous la modifierons selon chacun des trois cas. Certaines des notations utilisées dans l'écriture de la vraisemblance (2.11) auront des significations différentes selon l'approche utilisée. Notre objectif est de rendre la présentation concise et intuitive. Pour plus de détails techniques, nous renvoyons le lecteur soit à l'annexe, soit à la littérature.

2.3 Modèle de renouvellement alterné stationnaire

Un processus de renouvellement alterné devient *stationnaire* lorsque t, illustré à la Figure 2.3 qui est le début du processus de renouvellement retardé, est suffisamment grand ($t \to \infty$) et que le processus est en « équilibre ». En d'autres termes, le processus a commencé indéfiniment dans le passé.
Une définition plus formelle de processus de renouvellement alterné stationnaire est proposée par Alvarez (2006). Dans la suite, nous verrons d'autres manières de vérifier la stationnarité.

Dans l'approche *Stationnarité*, pour la vraisemblance, nous supposons la stationnarité. Dans cette approche, les dates de transition avant le début du panel (les temps calendaires initiaux) sont inconnues mais l'hypothèse de stationnarité nous permet d'obtenir une vraisemblance en utilisant les segments de période au début et à la fin du panel.

Ci-dessous, nous présentons quelques résultats bien connus et importants pour les processus de renouvellement stationnaires. Les références Ross (1996), Grimmett et Stirzaker (2001) et Cox et Isham (2000) sont utiles pour comprendre la stationnarité. Parmi ces trois références, Ross (1996) contient le plus d'informations concernant le processus de renouvellement alterné stationnaire. Pour les processus de renouvellement alternés qui n'ont pas encore atteint la stationnarité, Grimmett et Stirzaker (2001) est une bonne référence.

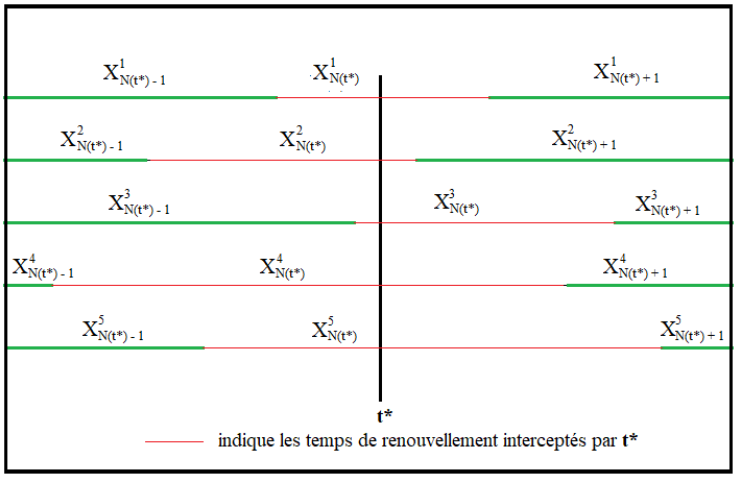
Le paradoxe de l'inspection

Il est bien connu que si nous avons des exemples de chemins d'un processus avec temps de renouvellement $X_j, j = 1, 2, 3, \ldots$ et sélectionnons un temps fixe arbitraire t^*, que les temps de renouvellement observés contenant t^* seront stochastiquement plus grands que les temps X_j dans le processus d'origine. Voir Grimmett et Stirzaker (2001).

Lorsque le processus est stationnaire, nous pouvons déterminer la densité de l'ensemble des temps de renouvellement sélectionnés interceptés par t^*. Elle est donnée par (voir Ross (1996) et Cox et Isham (2000)) [3] :

$$f(S_{N(t^*)} - S_{N(t^*)-1}) = \frac{xf(x)}{\mu} \quad \text{où} \quad X \sim f(x) \quad \text{et} \quad \mu = E[X]. \tag{2.7}$$

Figure 2.5: Représentation de cinq exemples de chemin d'un processus de renouvellement interceptés par un temps fixe t^*.



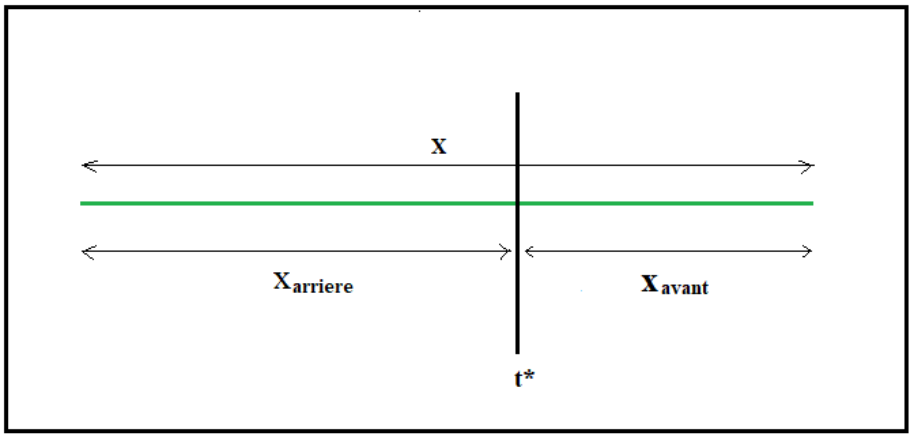
3. Pour comprendre pourquoi l'expression (2.7) diffère de celle de la densité d'un temps de renouvellement X_j, rappelons que $N(t^*)$ est une variable aléatoire.

La Figure 2.5 illustre un ensemble de temps de renouvellement $X_j^i, i = 1, 2, ..., 5$ pour cinq individus d'un processus de renouvellement alterné interceptés par un temps fixe t^*. Les temps en rouge $X^i_{N(t^*)}$ sont stochastiquement plus longs que les temps de renouvellement habituels X_j^i.

Temps de récurrence avant et temps de récurrence arrière

Pour les temps de renouvellement interceptés par un temps t^*, nous définissons le *temps de récurrence avant* (*forward recurrence time*) comme étant le temps restant dans la période de temps de renouvellement après t^* (aussi connu comme *durée de vie excédentaire* ou *durée de vie résiduelle*). De même, nous définissons le *temps de récurrence arrière* (*backward recurrence time*) comme étant le temps entre le début de la période de temps de renouvellement et t^* (aussi connu comme *l'âge au temps* t^*). La Figure 2.6 en donne des illustrations.

Figure 2.6: Illustration des temps de récurrence avant et arrière.



Une fois que le processus a atteint la stationnarité, les temps de renouvellement avant et

arrière ont tous deux la même distribution. Si S est la fonction de survie pour les temps de renouvellement X_j dans le processus de renouvellement, la fonction de répartition pour les temps de renouvellement avant et arrière est (voir Lawless (2003)) :

$$F_{\text{avant}}(x) = F_{\text{arriere}}(x) = \frac{\int_0^x S(u)du}{\mu} \quad \text{où} \quad \mu = \mathbb{E}[X]. \tag{2.8}$$

Disponibilité asymptotique

La disponibilité asymptotique est définie dans Alvarez (2006) comme étant la probabilité que le système soit dans l'état pour lequel $W(t) = 1$ à tout instant t. D'après Ross (1996), on voit facilement que la probabilité qu'un processus de renouvellement alterné stationnaire soit dans l'état C à tout instant t est :

$$P(\text{état } C) = \frac{\mathbb{E}[Z_j]}{\mathbb{E}[Z_j] + \mathbb{E}[Y_j]}. \tag{2.9}$$

De même, le processus de renouvellement alterné est non disponible lorsqu'il se trouve dans l'état pour lequel $W(t) = 0$ à tout instant t. La probabilité correspondante est :

$$P(\text{état } E) = \frac{\mathbb{E}[Y_j]}{\mathbb{E}[Z_j] + \mathbb{E}[Y_j]}. \tag{2.10}$$

Pour simplifier la Section 2.5.3 et éviter beaucoup de détails théoriques comme dans Alvarez (2006), nous utiliserons les résultats de cette section pour rendre notre présentation plus intuitive et moins technique.

2.4 Vraisemblance dans le cadre général

Avant de considérer les trois approches présentées au début de ce chapitre pour traiter les périodes sectionnées par le début et la fin du panel, nous développons la vraisemblance

dans un cadre général. Pour simplifier la présentation, notons que nous considérons pour le moment que les situations de vraisemblance où $\tau > 2$, c'est-à-dire les situations où il y a au moins 3 renouvellements dans le panel ($X_1, X_2, ..., X_\tau$, avec $\tau \geq 3$). De la présentation, le lecteur peut facilement déduire comment traiter les cas $\tau = 1$ et $\tau = 2$. Ces cas sont traités en Annexe A.4.

Notre objectif est de construire une expression de vraisemblance suffisamment générale et flexible pour s'adapter à la fois aux quatre situations présentées dans le Tableau 2.1 et aux différentes approches décrites dans les énumérations [1], [2] et [3]. En s'autorisant une certaine ambiguïté dans la notation, nous développons l'expression de vraisemblance générale (2.11). Nous décrivons dans les lignes suivantes l'ambiguïté de la notation.

Les densités $f_{C,\text{début}}$ et $f_{C,\text{fin}}$ sont des formes modifiées de la densité f_C au début et à la fin du panel respectivement pour l'état de chômage C. Tandis que $f_{E,\text{début}}$ et $f_{E,\text{fin}}$ sont les pendants respectifs de la densité f_E pour l'état d'emploi E. Leur modification dépend de la manière dont la troncature et la censure sur l'intervalle constituant le panel sont traitées dans chacune des trois approches. Rappelons que l'approche utilisée dépend du type de données disponibles. De plus, selon le cas présenté (rappel du Tableau 2.1) et les données disponibles, z_1, y_1, z_τ et y_τ peuvent représenter un temps de séjour complet ou un segment de temps de séjour (une période incomplète de temps de séjour). Un résumé des différentes interprétations pour z_1, y_1, z_τ et y_τ est donné dans le Tableau 2.2.

Tableau 2.2: Résumé des différentes interprétations pour les notations z_1, y_1, z_τ et y_τ en début ($t = 0$) et en fin ($t = T$) de panel.

W	État correspondant	Commentaires
$W(0) = 0$	Emploi E	$z_1 = 0$, $y_1 =$ segment
$W(0) = 1$	Chômage C	$z_1 =$ segment, $y_1 =$ complet
$W(T) = 0$	Emploi E	$z_\tau =$ complet, $y_\tau =$ segment
$W(T) = 1$	Chômage C	$z_\tau =$ segment, $y_\tau =$ non observé

Ces ambiguïtés de notation permettent de présenter une seule vraisemblance qui peut être modifiée pour chaque cas que l'on souhaite traiter. Dans ce chapitre, nous nous limitons aux vraisemblances pour un seul individu (l'indice i a été supprimé). Nous généralisons cela pour n individus *iid* au Chapitre 3.

Les contributions à la vraisemblance, pour un individu donné, sont :

premier renouvellement retardé X_1 : $[f_{C,\text{début}}(z_1) f_E(y_1)]^{\mathbb{1}(w(0)=1)} \times [f_{E,\text{début}}(y_1)]^{\mathbb{1}(w(0)=0)}$;

dernier renouvellement X_τ : $[f_{C,\text{fin}}(z_\tau)]^{\mathbb{1}(w(T)=1)} \times [f_C(z_\tau) f_{E,\text{fin}}(y_\tau)]^{\mathbb{1}(w(T)=0)}$;

les autres renouvellements (si $\tau > 2$) : $\prod_{j=2}^{\tau-1} [f_C(z_j) f_E(y_j)]$.
En combinant le tout dans une seule vraisemblance on obtient :

$$\begin{aligned} l = {} & [f_{C,\text{début}}(z_1) f_E(y_1)]^{\mathbb{1}(w(0)=1)} \times [f_{E,\text{début}}(y_1)]^{\mathbb{1}(w(0)=0)} \\ & \times \underbrace{\prod_{j=2}^{\tau-1} [f_C(z_j) f_E(y_j)]}_{\text{si } \tau > 2} \\ & \times [f_{C,\text{fin}}(z_\tau)]^{\mathbb{1}(w(T)=1)} \times [f_C(z_\tau) f_{E,\text{fin}}(y_\tau)]^{\mathbb{1}(w(T)=0)} . \end{aligned} \tag{2.11}$$

2.5 Trois approches de traitement des périodes incomplètes de temps de séjour

En référence à la vraisemblance (2.11), nous abordons les trois approches décrites dans l'introduction. Dans chaque sous-section, nous concluons en résumant les avantages et les inconvénients de l'approche considérée. Dans l'ensemble, il est utile de se référer aux Tableaux 2.1 et 2.2.

2.5.1 Non prise en compte du premier temps de séjour

La non prise en compte du premier temps de séjour est l'approche adoptée dans Kovačević et Roberts (2007). La période incomplète au début du panel est simplement abandonnée et la période incomplète à la fin est traitée avec une censure à droite (voir Kalbfleisch et Prentice (2011) pour une description de la censure à droite). En modifiant ainsi la vraisemblance (2.11), nous avons

$$\begin{aligned} l =& [f_E(y_1)]^{\mathbb{1}(w(0)=1)} \times \underbrace{\prod_{j=2}^{\tau-1} [f_C(z_j) f_E(y_j)]}_{\text{si } \tau > 2} \\ & \times [S_C(z_\tau)]^{\mathbb{1}(w(T)=1)} \times [f_C(z_\tau) S_E(y_\tau)]^{\mathbb{1}(w(T)=0)} , \end{aligned} \tag{2.12}$$

où dans le cas de $W(T) = 1$, $f_{C,\text{fin}}(z_\tau) = 1 - F_C(z_\tau) = S_C(z_\tau)$ et dans le cas de $W(T) = 0$, $f_{E,\text{ fin}}(y_\tau) = 1 - F_E(y_\tau) = S_E(y_\tau)$.

Puisque dans chaque cas nous savons que les temps de séjour réels étaient respectivement supérieurs à z_τ et y_τ (les segments résultant des temps de séjour étant sectionnés à la fin du panel), nous insérons une expression paramétrique pour la probabilité que les temps de séjour réels soient plus longs que les segments observés. Il s'agit de la contribution à la probabilité d'une censure à droite indépendante et non informative.

Dans le Chapitre 3, nous revisiterons la vraisemblance (2.12) dans le contexte d'une chaîne de Markov continue homogène à deux états alternés.

Nous présentons les avantages et les inconvénients de la méthode de non-prise en compte de la première période de temps dans l'encadré suivant.

Avantage : En laissant tomber la première période de temps (celle interceptée par le début du panel), nous n'avons pas besoin des dates de début : les temps calendaires initiaux (de cette période de temps) avant le début du panel. De plus, aucune hypothèse forte telle que la stationnarité n'est requise.
Inconvénient : Nous jetons des données, c'est-à-dire la période de temps observée au début du panel. Par exemple, pour un individu i ayant 3 renouvellements X_{ij}, ce serait une perte de 16% à 25% [a] des données.

a. Ces pourcentages de sont déterminés théoriquement pour un individu ayant 3 renouvellements et selon qu'il commence et termine le panel en C ou en E.

2.5.2 Vraisemblance conditionnelle

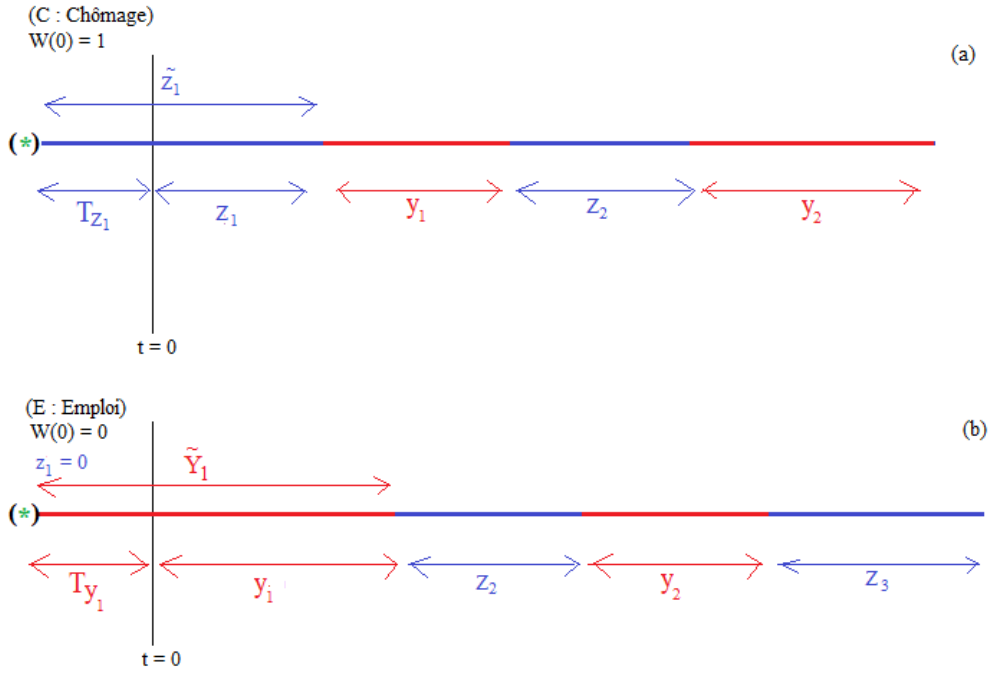
Pour l'approche de vraisemblance conditionnelle, l'expression (2.11) s'applique avec $f_{C,\text{début}}(z_1)$ et $f_{E,\text{début}}(y_1)$ correctement spécifiées. Dans cette approche, nous remplaçons $f_{C,\text{début}}(z_1)$ et $f_{E,\text{début}}(y_1)$ par des distributions conditionnelles. Une notation supplémentaire est nécessaire :

— lorsque $W(0) = 1$, la première période interceptée par le début du panel est la période de chômage Z_1. Cette période sera donc un segment de période pour un temps de récurrence avant (*forward recurrence time*). Nous introduisons la notation $\tilde{Z}_1$ pour toute la période de temps et T_{Z_1} pour le segment de période de temps avant le début du panel c.-à-d. le temps de récurrence arrière (*backward recurrence time*). Notons

que $\tilde{Z}_1 = Z_1 + T_{Z_1}$ (voir Figure 2.7 (a)) ;

— dans le cas où $W(0) = 0$, la première période de temps interceptée par le début du panel est une période d'emploi Y_1. Cette dernière est aussi un segment de période pour un temps de récurrence avant (*forward recurrence time*). Nous introduisons alors la notation $\tilde{Y}_1$ pour la période de temps entière et T_{Y_1} pour le segment de la période de temps avant le début du panel (*backward recurrence time*). Notons que $\tilde{Y}_1 = Y_1 + T_{Y_1}$ (voir Figure 2.7 (b)).

Figure 2.7: Schéma pour illustrer les notations $\tilde{Z}_1, \tilde{Y}_1, T_{Z_1}$ et T_{Y_1} requises pour la vraisemblance conditionnelle.



La Figure 2.7 illustre les notations $\tilde{Z}_1, \tilde{Y}_1, T_{Z_1}$ et T_{Y_1} requises pour la vraisemblance conditionnelle. Notons que, si les indices de temps de début indiqués par $(*)$ (en vert) c'est-à-dire

les temps calendaires initiaux (le début des périodes initiales) sont inconnus, il est impossible de connaître $\tilde{Z}_1, \tilde{Y}_1, T_{Z_1}$ et T_{Y_1} et l'approche conditionnelle ne peut pas être utilisée.

Dans l'approche conditionnelle, on conditionne la vraisemblance sur l'événement $\{\tilde{Z}_1\mathbb{1}(W(0) = 1), \tilde{Y}_1\mathbb{1}(W(0) = 0), W(0)\}$. Notons qu'en conditionnant sur $W(0)$, nous conditionnons également sur l'état dans lequel le processus démarre. L'approche conditionnelle est facilement implémentée en remplaçant $f_{C,\text{début}}(z_1)$ par
$f_C(\tilde{z}_1|\tilde{Z}_1 > t_{Z_1})$ et $f_{E,\text{début}}(y_1)$ par $f_E(\tilde{y}_1|\tilde{Y}_1 > t_{Y_1})$ où les minuscules z_1, y_1, t_{Z_1}, t_{Y_1}, $\tilde{z}_1$ et $\tilde{y}_1$ sont les réalisations des variables Z_1, Y_1, T_{Z_1}, T_{Y_1}, $\tilde{Z}_1$ et $\tilde{Y}_1$ respectivement.

Étant donné qu'habituellement, dans les données d'enquête, aucune donnée n'est collectée après la fin du panel, nous utilisons à nouveau la censure à droite pour les segments de période à la fin du panel où dans le cas de $W(T) = 1$, $f_{C,\text{fin}}(z_\tau) = 1 - F_C(z_\tau) = S_C(z_\tau)$ et dans le cas de $W(T) = 0$, $f_{E,\text{fin}}(y_\tau) = 1 - F_E(y_\tau) = S_E(y_\tau)$.

Donc la vraisemblance (2.11), dans ce cas, devient

$$\begin{aligned} l =& [f_C(\tilde{z}_1|\tilde{Z}_1 > t_{Z_1}) f_E(y_1)]^{\mathbb{1}(w(0)=1)} \times [f_E(\tilde{y}_1|\tilde{Y}_1 > t_{Y_1})]^{\mathbb{1}(w(0)=0)} \\ & \times \underbrace{\prod_{j=2}^{\tau-1} [f_C(z_j) f_E(y_j)]}_{\text{si } \tau > 2} \times [S_C(z_\tau)]^{\mathbb{1}(w(T)=1)} \times [f_C(z_\tau) S_E(y_\tau)]^{\mathbb{1}(w(T)=0)}. \end{aligned} \tag{2.13}$$

Dans la conclusion du mémoire, nous discuterons brièvement de ce qui pourrait être fait si les dates de fin (les temps calendaires finaux) des dernières périodes de temps sont collectées après la fin du panel.

Dans le Chapitre 3, nous revisiterons la vraisemblance (2.13) dans le cas d'une chaîne de Markov continue homogène à deux états.

L'encadré suivant présente les avantages et les inconvénients de la méthode de la vraisemblance conditionnelle.

Avantage : Nous utilisons toutes les données disponibles et aucune hypothèse forte telle que la stationnarité n'est requise.
Inconvénient : Dans les données de panel d'enquête, les temps de transition avant le début du panel peuvent être inconnus pour certains ou tous les individus.

2.5.3 Hypothèse de stationnarité

Pour la vraisemblance étudiée ici, nous supposons la stationnarité comme dans Alvarez (2006). Comme indiqué précédemment, nous nous appuyons sur les résultats de la Section 2.3 pour simplifier la présentation. Puis au Chapitre 3, nous présenterons cette vraisemblance pour les distributions exponentielles des temps de séjour dans l'hypothèse d'une chaîne de Markov continue homogène à deux états.

En considérant à nouveau la Figure 2.7, nous voyons que si les dates de début marquées par $(*)$, les temps calendaires initiaux, ne sont pas disponibles alors tout ce que nous avons pour la vraisemblance (pour la contribution du premier renouvellement) est soit le segment z_1 et le temps plein y_1 (en supposant que le processus commence dans l'état C avec $W(0) = 1$) ou bien le segment y_1 avec $z_1 = 0$ (en supposant que le processus commence dans l'état E avec $W(0) = 0$). De plus, dans chaque cas, nous voyons que le segment z_1 (si le processus commence dans l'état C) et le segment y_1 (si le processus est dans l'état E initialement) sont des temps de récurrence avant (*forward recurrence time*) dont la fonction de répartition est donnée par l'expression (2.8). Puisque nous ne conditionnons pas sur l'état initial dans cette vraisemblance, nous devons donc inclure les probabilités de débuter dans les états C et E. Comme le processus de renouvellement alterné est supposé être stationnaire, ces probabilités sont données dans les expressions (2.9) et (2.10).

Nous suivons la présentation de Alvarez (2006) et utilisons une censure à droite indépendante et non informative pour les derniers segments de temps (c'est-à-dire z_τ ou y_τ selon la réalisation de $W(T)$) à la fin du panel. Cependant, les derniers segments de temps du panel sont des temps de récurrence en arrière (*backward recurrence time*) et on pourrait développer une vraisemblance qui les modélise comme tels.

En revenant à notre expression générale (2.11), nous définissons $f_{C,\text{début}}, f_{E,\text{début}}, f_{C,\text{fin}}$ et $f_{E,\text{fin}}$ en supposant la stationnarité et qu'aucune date de transition n'est connue avant le début du panel (les temps calendaires initiaux sont donc supposés inconnus).

En notant les fonctions de survie associées aux fonctions de densité f_C et f_E par S_C et S_E respectivement, les moyennes de Z_j et Y_j par μ_C et μ_E respectivement, nous avons :

— si $W(0) = 1$ (si le processus débute dans l'état C) alors

$$f_{C,\text{début}}(z_1) = \left(\frac{\mathbb{E}[Z_j]}{\mathbb{E}[Z_j] + \mathbb{E}[Y_j]}\right) \times \left(\frac{d}{dz_1}\frac{\int_0^{z_1} S_C(u)du}{\mu_C}\right), \tag{2.14}$$

— si $W(0) = 0$ (si le processus débute dans l'état E) alors

$$f_{E,\text{début}}(y_1) = \left(\frac{\mathbb{E}[Y_j]}{\mathbb{E}[Z_j] + \mathbb{E}[Y_j]}\right) \times \left(\frac{d}{dy_1}\frac{\int_0^{y_1} S_E(u)du}{\mu_E}\right). \tag{2.15}$$

En revenant à l'expression (2.11) la vraisemblance est

$$\begin{aligned} l = &\left[\left(\frac{\mathbb{E}[Z_j]}{\mathbb{E}[Z_j] + \mathbb{E}[Y_j]}\right) \times \left(\frac{d}{dz_1}\frac{\int_0^{z_1} S_C(u)du}{\mu_C}\right) f_E(y_1)\right]^{\mathbb{1}(w(0)=1)} \\ &\times \left[\left(\frac{\mathbb{E}[Y_j]}{\mathbb{E}[Z_j] + \mathbb{E}[Y_j]}\right) \times \left(\frac{d}{dy_1}\frac{\int_0^{y_1} S_E(u)du}{\mu_E}\right)\right]^{\mathbb{1}(w(0)=0)} \\ &\times \underbrace{\prod_{j=2}^{\tau-1}[f_C(z_j)f_E(y_j)]}_{\text{si } \tau > 2} \\ &\times [S_C(z_\tau)]^{\mathbb{1}(w(T)=1)} \times [f_C(z_\tau)S_E(y_\tau)]^{\mathbb{1}(w(T)=0)}. \end{aligned} \tag{2.16}$$

Les avantages et inconvénients de la méthode de l'hypothèse de stationnarité sont présentés dans l'encadré qui suit.

Avantage : Les temps de début avant introduction dans le panel (les temps calendaires initiaux) ne sont pas nécessaires.
Inconvénient : Supposer la stationnarité est une hypothèse forte.

Dans le prochain chapitre, nous revisiterons les vraisemblances (2.12), (2.13) et (2.16), et nous les exprimerons en termes de densités exponentielles pour f_C et f_E en faisant l'hypothèse d'une chaîne de Markov homogène continue à deux états pour notre processus de renouvellement. Dans le présent chapitre, les vraisemblances ont été écrites pour un seul individu. Dans le Chapitre 3, nous généraliserons ces vraisemblances pour n individus.

CHAPITRE 3
VRAISEMBLANCES

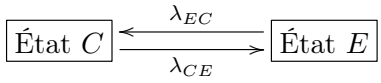
Dans ce chapitre, nous revenons aux vraisemblances dérivées au Chapitre 2 pour les différents cas de données disponibles et appliquons les approches dans le cas d'une chaîne continue de Markov homogène à deux états.
Spécifiquement, nous exprimons f_C et f_E par des densités exponentielles :

$$f_C(z) = \lambda_C \exp(-\lambda_C\, z) \qquad \text{et} \qquad f_E(y) = \lambda_E \exp(-\lambda_E\, y). \tag{3.1}$$

Les estimateurs du maximum de vraisemblance (EMV) sont trouvés pour les paramètres des densités exponentielles f_C et f_E.

Dans le langage généralement utilisé pour la modélisation multi-états, λ_C est l'intensité de transition de C à E et λ_E est l'intensité de transition de E à C. Ainsi, comme indiqué sur la Figure 3.1, on pourrait aussi utiliser la notation λ_{CE} pour λ_C et λ_{EC} pour λ_E. Les intensités de transition sont constantes car nous supposons ici un modèle de Markov homogène continu à deux états avec des densités exponentielles.

Figure 3.1: Intensités de transition ajoutées à la Figure 2.1.



La Figure 3.1 est une reprise de la Figure 2.1 pour laquelle les intensités de transition ont été rajoutées. Dans notre notation, $\lambda_{CE} = \lambda_C$ et $\lambda_{EC} = \lambda_E$

Dans les Sections 3.1, 3.2 et 3.3, nous utilisons les notations suivantes :

τ_i = nombre de renouvellements (X_{ij}) (y compris les périodes de renouvellement incomplets) dans le panel pour l'individu i ;

w_{i0} = valeur réalisée de $W(0)$ pour l'individu i en début de panel ;

w_{iT} = valeur réalisée de $W(T)$ pour l'individu i en fin de panel ;

les valeurs de z_1, y_1, z_τ et y_τ pour l'individu i sont respectivement notées z_{i1}, y_{i1}, $z_{i\tau}$ et $y_{i\tau}$;

$r_1 = \sum_{i=1}^{n} \mathbb{1}(w_{i0} = 1)$ = nombre d'individus commençant le panel dans l'état C ;

$r_0 = \sum_{i=1}^{n} \mathbb{1}(w_{i0} = 0) = n - r_1$ = nombre d'individus commençant le panel dans l'état E ;

$d_1 = \sum_{i=1}^{n} \mathbb{1}(w_{iT} = 1)$ = nombre d'individus terminant le panel dans l'état C ;

$d_0 = \sum_{i=1}^{n} \mathbb{1}(w_{iT} = 0) = n - d_1$ = nombre d'individus terminant le panel dans l'état E ;

$C(T)$ = cumul de tous les temps de chômage, dans l'état C, pour tous les individus dans le panel ;

$E(T)$ = cumul de tous les temps d'emploi, dans l'état E, pour tous les individus dans le panel.

En notant l_i la vraisemblance pour le i^{e} individu, nous avons, pour n individus *iid*,

$$L = \prod_{i=1}^{n} l_i.$$

En prenant le logarithme népérien ln, nous obtenons

$$\ln L = \sum_{i=1}^{n} \ln l_i.$$

Nous retrouvons le ln de l'expression (2.11) en termes des densités exponentielles f_C et f_E (3.1) et les fonctions de survie correspondantes :

$$S_C(z) = \exp(-\lambda_C\, z) \qquad \text{et} \qquad S_E(y) = \exp(-\lambda_E\, y).$$

Nous traitons $f_{C,\text{début}}(z_1)$, $f_{E,\text{début}}(y_1)$, $f_{C,\text{fin}}(z_\tau)$ et $f_{E,\text{fin}}(y_\tau)$ de manière appropriée pour chaque cas. Par ailleurs, puisque nous utilisons la censure à droite dans les trois approches, nous avons

$$f_{C,\text{fin}}(z_\tau) = S_C(z_\tau) \text{ et } f_{E,\text{fin}}(y_\tau) \quad = S_E(y_\tau)$$

pour chaque cas.

3.1 Non prise en compte du premier temps de séjour

3.1.1 Dérivation

Nous ré-exprimons la vraisemblance l_i dans (2.12) pour un individu i en termes de distributions exponentielles. Nous écrivons ensuite $\sum_{i=1}^{n} \ln l_i$ afin d'obtenir la log-vraisemblance $\ln L$ pour n observations *iid*. Notons

$$\begin{aligned} C(T) &= \sum_{i=1}^{n} \sum_{j=2}^{\tau_i} z_{ij}, \\ E(T) &= \sum_{i=1}^{n} y_{i1} \mathbb{1}(w_{i0} = 1) + \sum_{i=1}^{n} \sum_{j=2}^{\tau_i - 1} y_{ij} + \sum_{i=1}^{n} y_{i\tau_i} \mathbb{1}(w_{iT} = 0). \end{aligned}$$

Nous avons

$$\begin{aligned} \ln L = (d_0 + \sum_{i=1}^{n} \tau_i - 2n) \ln \lambda_C + (r_1 + \sum_{i=1}^{n} \tau_i - 2n) \ln \lambda_E \\ - \lambda_C \, C(T) - \lambda_E \, E(T) \, . \end{aligned} \tag{3.2}$$

3.1.2 Optimisation

En maximisant $\ln L$ dans l'expression (3.2), nous obtenons les EMV des paramètres de f_C et f_E :

$$\hat{\lambda}_C = \frac{d_0 + \sum_{i=1}^{n} \tau_i - 2n}{C(T)}$$

et

$$\hat{\lambda}_E = \frac{r_1 + \sum_{i=1}^{n} \tau_i - 2n}{E(T)}.$$

Habituellement, pour un risque constant, ces estimateurs sont le quotient du nombre d'évènements par le temps à risque, Selvin (2008).

3.2 Vraisemblance conditionnelle

3.2.1 Dérivation

Pour l'approche conditionnelle, nous nous référons à l'expression (2.13). Nous introduisons également quelques notations nouvelles :

$\tilde{z}_1$ et $\tilde{y}_1$ sont notés $\tilde{z}_{i1}$ et $\tilde{y}_{i1}$; respectivement, pour l'individu i ;

t_{Z_1} et t_{Y_1} sont notés $t_{Z_{i1}}$ et $t_{Y_{i1}}$ pour l'individu i respectivement.

Nous avons

$$\begin{aligned} f_{C,\text{début}}(z_{i1}) &= f_C(\tilde{z}_{i1} | \tilde{z}_{i1} > t_{Z_{i1}}) \\ &= \frac{f_C(\tilde{z}_{i1}, \tilde{z}_{i1} > t_{Z_{i1}})}{S_C(t_{Z_{i1}})} \\ &= \lambda_C \exp(-\lambda_C(\tilde{z}_{i1} - t_{Z_{i1}})) \\ &= \lambda_C \exp(-\lambda_C \, z_{i1}) \\ &= f_C(z_{i1}). \end{aligned} \tag{3.3}$$

Le résultat ci-dessus est une conséquence du fait que la distribution exponentielle est *sans mémoire*.

De manière analogue, $f_{E,\text{début}}(y_{i1}) = f_E(\tilde{y}_{i1} | \tilde{y}_{i1} > t_{Y_{i1}}) = f_E(y_{i1})$.

Posons

$$C(T) = \sum_{i=1}^{n} z_{i1} \mathbb{1}(w_{i0} = 1) + \sum_{i=1}^{n} \sum_{j=2}^{\tau_i} z_{ij}$$

et

$$E(T) = \sum_{i=1}^{n} \sum_{j=1}^{\tau_i - 1} y_{ij} + \sum_{i=1}^{n} y_{i\tau_i} \mathbb{1}(w_{iT} = 0).$$

Nous obtenons

$$\begin{aligned} \ln L = & \left(r_1 + \sum_{i=1}^{n} \tau_i - 2n + d_0 \right) \ln \lambda_C \\ & + \left(\sum_{i=1}^{n} \tau_i - n \right) \ln \lambda_E - \lambda_C \, C(T) - \lambda_E \, E(T). \end{aligned} \tag{3.4}$$

3.2.2 Optimisation

En maximisant $\ln L$ dans l'expression (3.4), nous avons

$$\hat{\lambda}_C = \frac{r_1 + d_0 + \sum_{i=1}^{n} \tau_i - 2n}{C(T)}$$

et

$$\hat{\lambda}_E = \frac{\sum_{i=1}^{n} \tau_i - n}{E(T)}.$$

Bien que cette approche soit différente de celle présentée dans la Section 3.1, nous obtenons des estimateurs dont les formules ont la même généralité, soit le quotient du nombre d'évènements par le temps à risque.

3.3 Hypothèse de stationnarité

3.3.1 Dérivation

Pour l'approche stationnaire, nous nous référons à l'expression (2.16).

Des équations (2.14) et (2.15) nous obtenons respectivement,

$$f_{C,\text{début}}(z_1) = \left(\frac{\lambda_E}{\lambda_C + \lambda_E}\right) \lambda_C \exp(-\lambda_C\ z_1) \tag{3.5}$$

et

$$f_{E,\text{début}}(y_1) = \left(\frac{\lambda_C}{\lambda_C + \lambda_E}\right) \lambda_E \exp(-\lambda_E\ y_1). \tag{3.6}$$

Posons

$$C(T) = \sum_{i=1}^{n} z_{i1} \mathbb{1}(w_{i0} = 1) + \sum_{i=1}^{n} \sum_{j=2}^{\tau_i} z_{ij}$$

et

$$E(T) = \sum_{i=1}^{n} \sum_{j=1}^{\tau_i - 1} y_{ij} + \sum_{i=1}^{n} y_{i\tau_i} \mathbb{1}(w_{iT} = 0).$$

Avec ces différentes notations, (2.16) devient :

$$\begin{aligned} \ln L = &-n \ln(\lambda_C + \lambda_E) + (\sum_{i=1}^{n} \tau_i - n + d_0) \ln \lambda_C \\ &+ (\sum_{i=1}^{n} \tau_i - n + r_1) \ln \lambda_E - \lambda_C\ C(T) - \lambda_E\ E(T). \end{aligned} \tag{3.7}$$

3.3.2 Optimisation

Posons

$$
\begin{aligned}
U(T) &= 2\sum_{i=1}^{n}\tau_i - 3n + d_0 + r_1 \ , \\
\alpha(T) &= C(T)\left[E(T) - C(T)\right] \ , \\
\beta(T) &= n\ E(T) + U(T)C(T) - \left[E(T) - C(T)\right]\left(\sum_{i=1}^{n}\tau_i - n + d_0\right) \\
&= C(T)\left(3\sum_{i=1}^{n}\tau_i - 4n + 2d_0 + r_1\right) + E(T)\left(2n - \sum_{i=1}^{n}\tau_i - d_0\right) \ , \\
&\quad et \\
\gamma(T) &= -U(T)\left[\sum_{i=1}^{n}\tau_i - n + d_0\right] \ .
\end{aligned}
$$

Avec ces notations, l'optimisation de l'expression de $\ln L$ dans (3.7) conduit à la forme quadratique (3.8) en λ_C :

$$\alpha(T)\ \lambda_C^2 + \beta(T)\ \lambda_C + \gamma(T) = 0 \tag{3.8}$$

dont les racines positives sont les EMV recherchés.

En maximisant $\ln L$ dans l'expression (3.7), nous avons

si $C(T) \neq E(T)$, $\alpha(T)$ serait non nul et la solution positive de (3.8) est

$$\hat{\lambda}_C = \frac{-\beta(T) + \sqrt{(\beta(T))^2 - 4\ \alpha(T)\ \gamma(T)}}{2\ \alpha(T)};$$

si $C(T) = E(T)$ alors $\alpha(T)$ serait nul et

$$\hat{\lambda}_C = \frac{U(T)\left(\sum_{i=1}^{n}\tau_i - n + d_0\right)}{E(T)\ (n + U(T))}.$$

Et dans les deux cas, on a

$$\hat{\lambda}_E = \frac{U(T) - \hat{\lambda}_C \, C(T)}{E(T)}.$$

Dans le Chapitre 4, nous présenterons des simulations qui permettront d'étudier les distributions d'échantillonnage des EMV (pour différentes tailles d'échantillon et largeurs de panel) pour les trois approches présentées dans ce chapitre.

CHAPITRE 4
SIMULATIONS ET RÉSULTATS

4.1 Introduction

Dans ce chapitre, nous présentons les résultats de différentes simulations pour illustrer les approches présentées dans le Chapitre 2 et appliquées au Chapitre 3 à une chaîne de Markov continue homogène. Dans l'hypothèse d'une chaîne de Markov continue homogène, nous utilisons des distributions exponentielles pour les fonctions de densité f_C et f_E. La recherche est faite pour différentes valeurs de T, la largeur de la fenêtre d'observation, tout en faisant varier la taille n de l'échantillon afin de comparer la puissance des différents modèles.

4.2 Algorithme de simulation

Après examen de la base de données de l'EDTR, il est apparu que les données de cette enquête sont relatives à celles d'un modèle de renouvellement alterné retardé.
Pour créer un processus de renouvellement alterné retardé, nous simulons un processus de renouvellement alterné pur commençant à un instant initial t', puis collectons plus tard (à un instant défini comme origine $t = 0$ et désignant dans notre cas le début de l'enquête) les données sur une fenêtre, un intervalle ou un panel de longueur T. Lorsque l'on souhaite simuler un processus stationnaire, on laisse le processus évoluer un certain temps avant de collecter les données sur le panel de longueur T. Voir la Section 2.2 en guise de rappel.

Pour vérifier l'hypothèse de stationnarité, nous pouvons utiliser les relations (2.7), (2.8), (2.9) et (2.10). À partir de la relation (2.8) par exemple, nous pouvons vérifier l'hypothèse de stationnarité en comparant la distribution observée des temps de récurrence avant (*Forward recurrence times*) à celle des temps de récurrence arrière (*Backward recurrence times*). On

s'attend à obtenir des distributions identiques.
Dans notre cas de figure, nous avons utilisé les relations (2.9) et (2.10). Nous avons donc comparé, à $t = 0$, les proportions observées des individus en état de chômage au début du panel avec la probabilité d'être en chômage et les équivalents pour les individus en état d'emploi pour différentes tailles d'échantillon et de largeur de panel selon qu'on observe l'hypothèse de stationnarité ou non. Les résultats de chaque catégorie, répliquée 100 fois, sont présentés dans le Tableau A.4 de l'Annexe A.5.

Pour simuler le processus de renouvellement alterné pur, nous commençons par générer une variable $U \sim U(0, 1)$ afin de savoir par quel type d'état commencer le processus. Si la réalisation de cette uniforme est inférieure à un seuil (disons le taux de chômage actuel), nous générons une période de temps de chômage à l'aide de la fonction de densité f_C. Sinon, nous générons une période de temps d'emploi à l'aide de la fonction de densité f_E.

Supposons que la première période de temps générée, pour un individu , est un temps de chômage. Si cette période de temps est supérieure à un seuil (disons *terminus*[1]) que nous nous sommes fixé, on arrête. Sinon, nous générons une période de temps d'emploi qu'on cumule à la période de chômage précédente. Si le cumul de ces deux périodes est supérieur à *terminus*, on arrête. Sinon, on génère une troisième période de temps, qui est une deuxième période de chômage, que nous cumulons aux deux périodes précédentes. Si le cumul est supérieur au seuil *terminus*, nous arrêtons le processus. Sinon, nous générons une quatrième période de temps, qui sera la deuxième période de travail. On vérifie le cumul, et ainsi de suite. Une fois que le cumul est supérieur ou égal à *terminus*, la génération des temps pour cet individu prend fin. Nous répétons cette procédure autant de fois nécessaire pour atteindre la taille d'échantillon désirée.

1. Le seuil *terminus* constitue la borne supérieure de la fenêtre. Il correspond donc à la somme de la borne inférieure du panel et de sa largeur. Dans le présent cas, $terminus = 0 + T$.

En faisant l'hypothèse d'une chaîne de Markov continue homogène, les périodes de temps de chômage et de travail sont, dans ce qui suit, générées par des lois exponentielles de paramètres respectifs λ_C et λ_E. Ces paramètres seront ensuite estimés.
Pour cela, en premier lieu, nous faisons varier la largeur T de la fenêtre d'observation et déterminons les estimateurs selon les approches présentées dans le Chapitre 2. En second lieu, nous faisons varier la taille n de l'échantillon. En itérant ce processus, nous comparons les trois approches en évaluant, par exemple, leurs erreurs quadratiques moyennes (EQM) dans ce chapitre, et présentons leurs biais ainsi que les écarts-types en Annexe A.5.

Dans ce qui suit, nous avons fixé $\lambda_C = 4$, $\lambda_E = 0.2$, puis avons fait varier n la taille de l'échantillon dans un cas, et T la largeur du panel dans l'autre cas. Le tout a été répliqué, pour approche Monte Carlo, $B = 100$ fois. La taille n de l'échantillon a pris diverses valeurs $n \in \{10, 30, 100\}$ alors que $T \in \{13, 26, 52\}$ dont l'unité de mesure peut, par exemple, être en semaines correspondant alors respectivement à un trimestre, un semestre et une année d'observation. Les résultats des simulations pour chaque approche (*Suppression*, *Conditionnelle* et *Stationnarité*) sont présentés dans différents tableaux et figures aux sections suivantes et à la Section A.5 de l'annexe.

4.3 Résultats de simulation en fixant la taille n de l'échantillon et en faisant varier la durée T de la fenêtre d'observation

Dans la présente section, pour une taille d'échantillon fixée, nous avons fait varier la largeur de la fenêtre d'observation $T \in \{13, 26, 52\}$, puis nous avons estimé les paramètres λ_C et λ_E pour chacune des approches.

En ayant fixé $n = 10$ et faisant varier la largeur T du panel, nous constatons que pour toutes les valeurs de T ($T = 13, T = 26$ et $T = 52$), les estimateurs sont légèrement biaisés. Voir les Tableaux A.5, A.6 et A.7. Cependant le biais, dans notre étude de simulation, a

tendance à s'amenuiser lorsque la taille T du panel devient grande. L'écart-type de chaque EMV tend à devenir plus petit lorsque la taille de la fenêtre devient grande. Dans le cas de l'hypothèse de stationnarité, les trois approches ont des écarts-types sensiblement égaux (voir les Figures A.5 à A.13). Lorsque l'hypothèse de stationnarité n'est pas vérifiée, l'approche *Stationnarité* fournit l'EMV ayant le plus petit écart-type mais avec des biais plus grands en valeur absolue ; l'approche *Suppression* fournit celui ayant le plus grand. Cependant l'écart-type, dans chaque approche (*Suppression, Conditionnelle et Stationnarité*) et dans chaque situation (avec ou sans hypothèse de stationnarité), diminue lorsque la largeur T de la fenêtre augmente. Cela s'observe aussi lorsque la taille de l'échantillon augmente. Les deux dernières observations semblent plus cohérentes puisqu'avec n ou T plus grand, on observe beaucoup plus de périodes de temps (chômage et emploi) et les EMV sont plus précis.
Par ailleurs, en comparant les EQM (Figures 4.1 à 4.3), on voit que l'approche *Stationnaire* produit les estimateurs avec les plus faibles EQM lorsque l'hypothèse de stationnarité est vérifiée tandis que l'approche *Conditionnelle* donne les seconds meilleurs estimateurs. Cependant, plus la taille de la fenêtre d'observation est élargie, plus les approches ont tendance à converger car les EQM ont tendance à être égales.

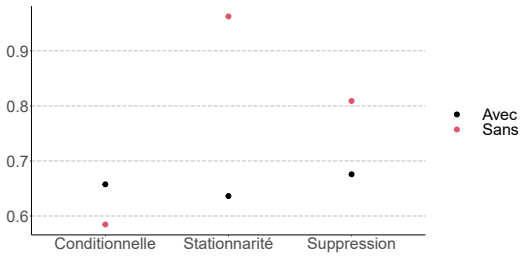
Pour $n = 30$ et $n = 100$, les résultats de l'EQM sont similaires à ceux observés pour $n = 10$: l'approche *Stationnaire* est celle produisant les estimateurs avec les plus petites erreurs quadratiques moyennes.

Conclusion partielle : En somme, en faisant varier de façon progressive la taille de la fenêtre $\{13, 26, 52\}$, il ressort que les approches *Stationnaire* et *Conditionnelle* fournissent des résultats plus précis lorsque l'hypothèse de stationnarité est vérifiée comparées à l'approche *Suppression*. Dans la situation où cette hypothèse n'est pas vérifiée, l'approche *Conditionnelle* performe mieux en terme d'EQM que l'approche *Suppression* suivie de l'approche

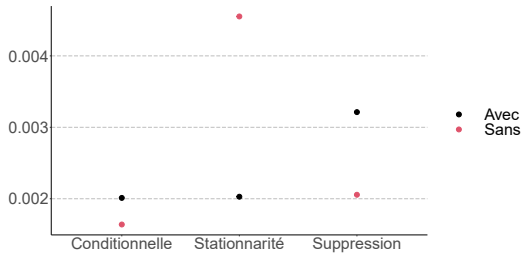
Stationnaire qui donne les pires résultats. Cela semble « plus logique » puisque l'approche *Stationnaire* a été développée pour donner de meilleurs résultats lorsqu'elle est vérifiée. L'approche *Suppression* donne de moins bons résultats que les deux autres dans le cas où la stationnarité est vérifiée mais produit de bien meilleurs résultats que l'approche *Stationnaire* dans le cas où la stationnarité n'est pas vérifiée. Par contre, tous les résultats, dans toutes les approches, ont tendance à se rapprocher lorsque la taille du panel, c'est-à-dire la largeur de la fenêtre, augmente. Cela a beaucoup de sens parce que, à l'inverse, plus la largeur du panel est petite, plus grande sera la proportion des temps de séjour incomplets, et les trois méthodes s'accorderont moins pour traiter ces temps de séjour. L'approche *Conditionnelle* est plus robuste : elle performe relativement mieux, en présence d'hypothèse de stationnarité ou non. Les résultats relatifs aux EQM pour chaque approche (*Suppression*, *Conditionnelle* et *Stationnaire*) sont présentés dans les figures suivantes de la présente Section 4.3 alors que ceux correspondant aux biais et aux écarts-types le sont en annexe. Les valeurs réelles des paramètres ainsi que les valeurs moyennes de leurs estimations par approche Monte Carlo sont présentées dans la Section A.5 de l'annexe.

Figure 4.1: EQM de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 10$, $T = 13$, $T = 26$ et $T = 52$

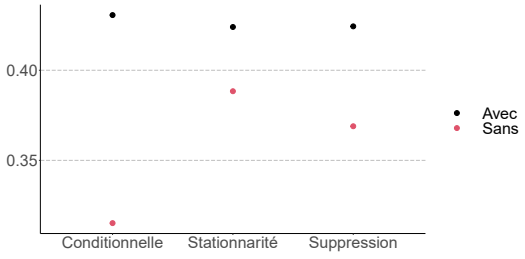
(a) EQM de $\hat{\lambda}_C$ pour $n = 10$ et $T = 13$



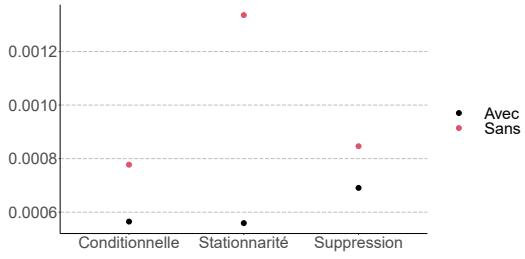
(b) EQM de $\hat{\lambda}_E$ pour $n = 10$ et $T = 13$



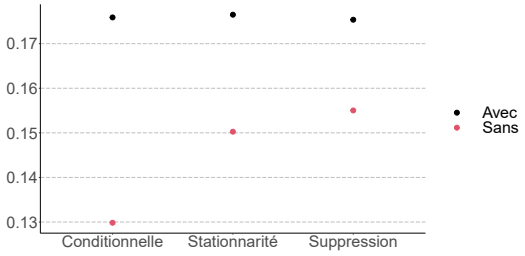
(c) EQM de $\hat{\lambda}_C$ pour $n = 10$ et $T = 26$



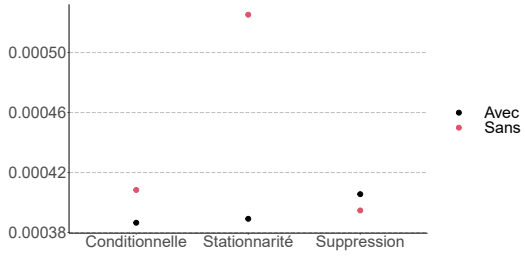
(d) EQM de $\hat{\lambda}_E$ pour $n = 10$ et $T = 26$



(e) EQM de $\hat{\lambda}_C$ pour $n = 10$ et $T = 52$



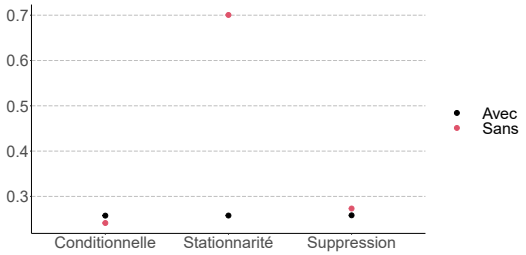
(f) EQM de $\hat{\lambda}_E$ pour $n = 10$ et $T = 52$



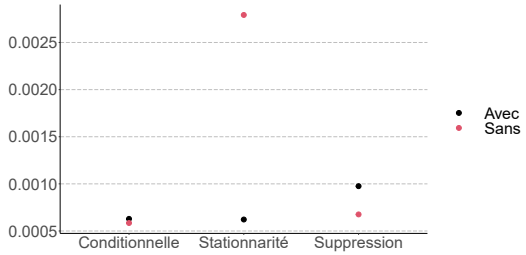
Notes de légende : Dans ces figures et les suivantes, « Avec » signifie que la stationnarité est atteinte alors qu'elle ne l'est pas pour « Sans »

Figure 4.2: EQM de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 30$, $T = 13$, $T = 26$ et $T = 52$

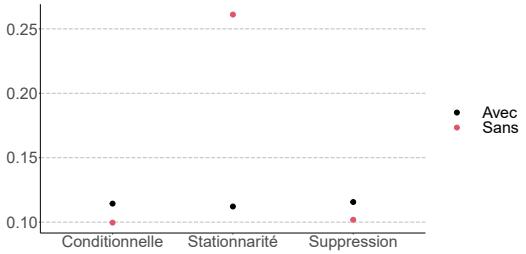
(a) EQM de $\hat{\lambda}_C$ pour $n = 30$ et $T = 13$



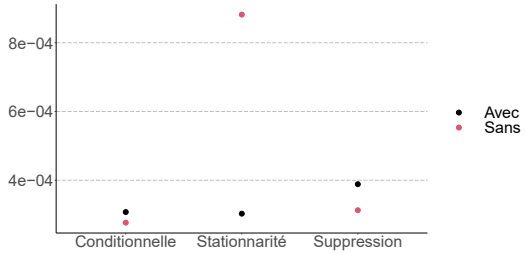
(b) EQM de $\hat{\lambda}_E$ pour $n = 30$ et $T = 13$



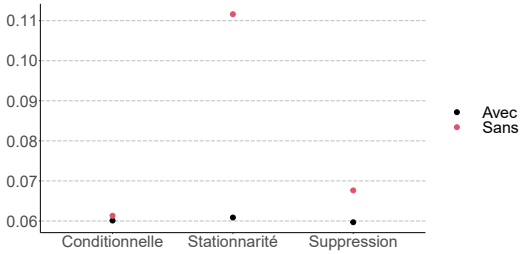
(c) EQM de $\hat{\lambda}_C$ pour $n = 30$ et $T = 26$



(d) EQM de $\hat{\lambda}_E$ pour $n = 30$ et $T = 26$



(e) EQM de $\hat{\lambda}_C$ pour $n = 30$ et $T = 52$



(f) EQM de $\hat{\lambda}_E$ pour $n = 30$ et $T = 52$

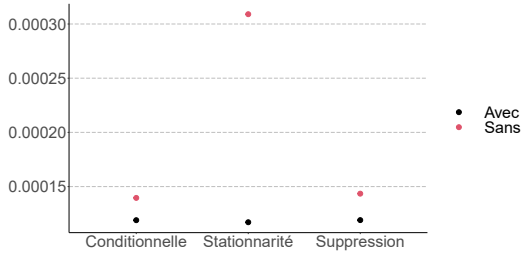
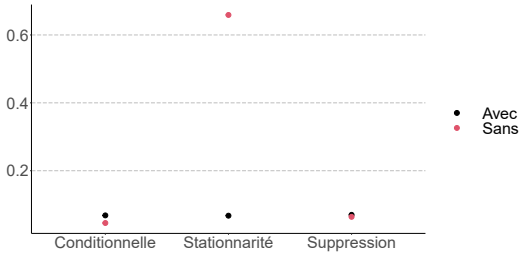
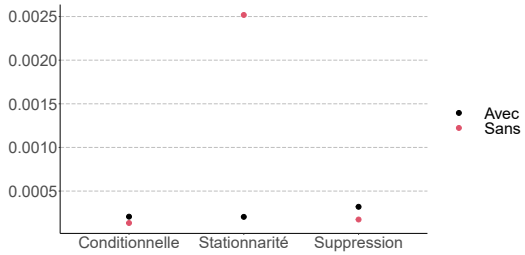


Figure 4.3: EQM de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 100$, $T = 13$, $T = 26$ et $T = 52$

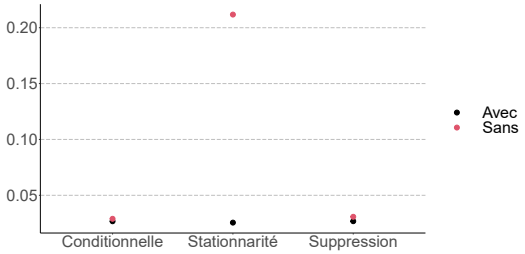
(a) EQM de $\hat{\lambda}_C$ pour $n = 100$ et $T = 13$



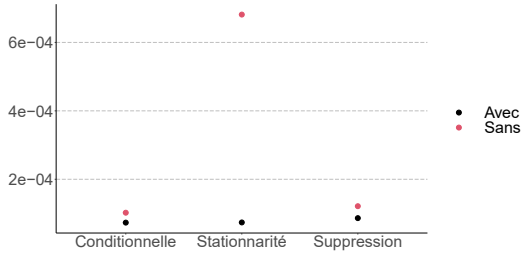
(b) EQM de $\hat{\lambda}_E$ pour $n = 100$ et $T = 13$



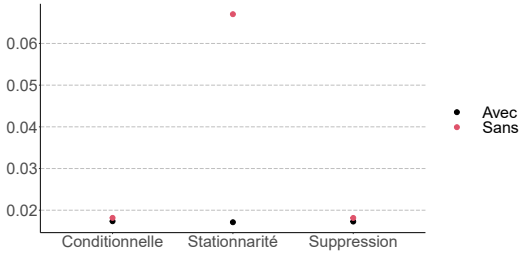
(c) EQM de $\hat{\lambda}_C$ pour $n = 100$ et $T = 26$



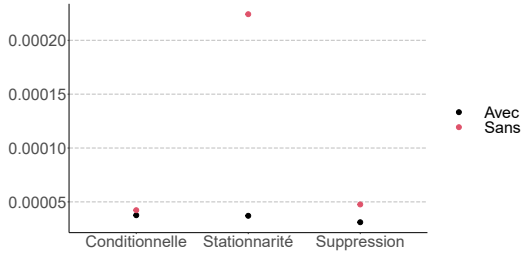
(d) EQM de $\hat{\lambda}_E$ pour $n = 100$ et $T = 26$



(e) EQM de $\hat{\lambda}_C$ pour $n = 100$ et $T = 52$



(f) EQM de $\hat{\lambda}_E$ pour $n = 100$ et $T = 52$



4.4 Résultats de simulation en fixant la largeur T du panel et en faisant varier la taille n de l'échantillon

Comme précédemment, nous avons déterminé les estimateurs des paramètres en faisant varier la taille n de l'échantillon, tout en fixant la largeur T de la fenêtre du panel. Nous avons, une fois de plus, fixé $\lambda_C = 4$, $\lambda_E = 0.2$. Le tout a été répliqué, pour approche Monte Carlo, $B = 100$ fois. L'intervalle de la fenêtre d'observation T a pris diverses valeurs $T \in \{13, 26, 52\}$.

Pour $T = 13$, tout en faisant varier la taille n de l'échantillon, nous avons obtenu que, pour toutes les tailles d'échantillon n $(10, 30$ et $100)$, l'approche *Stationnaire* produit les meilleurs estimateurs bien que l'approche *Conditionnelle* en fasse autant puisque les EQM sont très proches lorsque l'hypothèse de stationnarité est vérifiée. Dans le cas où l'hypothèse de stationnarité n'est pas vérifiée, l'approche *Suppression* est celle qui fournit les meilleurs estimateurs selon l'EQM.

Pour toutes les valeurs de n, lorsque $T = 26$, en l'absence d'hypothèse de stationnarité, l'approche *Conditionnelle* performe mieux, suivie des approches *Suppression* puis *Stationnaire*. En présence d'hypothèse de stationnarité, c'est l'approche *Stationnaire* qui performe le mieux. L'approche *Suppression* est la pire des trois dans ce cas.
Lorsque $T = 52$, les résultats sont tous de même ordre de grandeur bien que la préférence puisse être accordée à l'approche *Stationnaire* lorsque l'hypothèse de stationnarité est vérifiée et lorsqu'elle ne l'est pas, l'approche *Suppression* est préférée.

Conclusion partielle : Selon les résultats de simulation obtenus et présentés aux Figures 4.1 à 4.3 ; en présence de stationnarité, les approches de *Stationnaire* et *Conditionnelle* produisent des estimateurs plus précis que l'approche *Suppression*. Par contre, lorsque l'hypothèse de stationnarité n'est pas vérifiée, les approches *Conditionnelle* et *Suppression* performent mieux que l'approche *Stationnaire*. Cependant, lorsque la largeur du panel croît, toutes ces approches donnent des estimateurs de même ordre de grandeur. On peut, une fois de plus, considérer l'approche *Conditionnelle* car elle est moins contraignante (elle ne permet pas d'hypothèse forte de stationnarité et ne « jette » pas de données) et est aussi précise en présence ou non d'hypothèse de stationnarité.

Conclusion générale du chapitre : Dans ce chapitre, nous avons simulé des données de panel d'enquête dans un processus de renouvellement alterné retardé. Pour différents arguments de taille d'échantillon et de largeur de la fenêtre, nous avons déterminé les EMV dans chacune des approches étudiées dans le cadre de cette recherche, notamment :

— l'approche *Suppression*, dans laquelle les périodes sectionnées par le début du panel sont « jetées » et celles sectionnées par la fin du panel sont traitées par la censure aléatoire non-informative ;

— l'approche par vraisemblance conditionnelle où l'on suppose avoir les informations sur les dates d'entrée dans la période avant introduction dans le panel. Ces périodes interceptant le début du panel étant ainsi entièrement connues, cette approche conditionne sur l'état d'entrée dans le panel pour traiter les premières périodes. Les dernières périodes sont encore traitées avec la censure aléatoire non-informative ;

— la dernière approche est celle faisant l'hypothèse de stationnarité. Dans celle-ci, il n'est pas nécessaire de connaître entièrement la première période de temps interceptant le début du panel. En utilisant les résultats pour un processus de renouvellement alterné stationnaire, les premières périodes de temps sont traitées. Les dernières périodes sont encore une fois traitées avec la censure aléatoire non-informative.

Les résultats des simulations, répliquées $N = 100$ fois, permettent de conclure que lorsque l'hypothèse de stationnarité est vérifiée [2], l'approche *Stationnaire* performe mieux que l'approche *Conditionnelle* et l'approche *Suppression* vient en dernier. En l'absence d'hypothèse de stationnarité, les approches *Conditionnelle* et *Suppression* donnent, dans cet ordre, de meilleurs résultats que l'approche *Stationnaire*. Si la taille d'échantillon ou la largeur du panel est grande, ces différentes approches ont tendance à donner des résultats similaires (c'est-à-dire aussi bien pour T que pour n). En fait, plus la largeur T du panel ou la taille n de l'échantillon est grande, plus il y a d'observations et observer un temps de séjour (de chômage ou d'emploi) pour un individu donné n'est pas différent d'en observer un pour un autre individu. De plus, un grand nombre de temps de séjour diminuera l'impact des périodes tronquées par le début ou sectionnées par la fin du panel.

2. Les proportions des individus commençant le panel en situation de chômage ou d'emploi sont présentées au Tableau A.4 à la Section A.5 des résultats de l'annexe.

CHAPITRE 5
CONCLUSION ET TRAVAUX FUTURS

Motivé par des données d'enquête par panel, telle que l'enquête sur la dynamique du travail et du revenu (EDTR) de Statistique Canada, l'objectif de ce mémoire a été d'utiliser un modèle de renouvellement alterné pour apporter des traitements aux périodes de temps tronquées par le début de panel et sectionnées par la fin dudit panel. Pour ce faire, nous avons présenté, au Chapitre 1, la base de données de l'EDTR ainsi que ses complexités dont certaines ont conduit à la problématique ayant fait l'objet de ce travail. Le Chapitre 2 a servi à présenter la théorie utilisée pour le traitement de ces périodes de temps « problématiques » (périodes de temps tronquées et/ou sectionnées par le panel). Dans le Chapitre 3, nous avons abordé trois approches pour le traitement desdites périodes dans un tel contexte. Une étude de simulation faite au Chapitre 4 nous a permis de confronter chacune des approches explorées en présence ou non de l'hypothèse de stationnarité, par comparaison de l'erreur quadratique moyenne (EQM) des divers estimateurs obtenus au chapitre précédent.

Les données de panel proviennent d'enquêtes longitudinales. Comme exemple, on peut citer l'EDTR de Statistique Canada. Plusieurs complexités concernant cette base de données ont été relevées au Chapitre 1 de ce document. Parmi ces complexités, une particularité a retenu notre attention : comment traiter les périodes de temps sectionnées par la « fenêtre » que constitue le panel ? Bien que l'EDTR ait été pris comme référence, les résultats de cette recherche peuvent s'appliquer à toutes les enquêtes par panel, notamment, le Penn World Table (Feenstra *et al.*, 2015), le British Household Panel Study (Taylor *et al.*, 1993) et le Panel Study of Income, Dynamics Institute for Social Research (2021).

Dans le Chapitre 2, nous avons discuté de la façon dont les modèles multi-états peuvent être utilisés comme une approche naturelle pour traiter de telles données. Dans le cas

de l'EDTR, un processus de renouvellement alterné est un bon choix pour modéliser les périodes d'emploi et de chômage. Dans ce mémoire, nous nous sommes concentrés sur les périodes d'emploi et de chômage qui sont coupées par le début et la fin du panel, conduisant à des données incomplètes.

Au Chapitre 3, nous avons présenté trois approches pour traiter les données manquantes. Celles-ci sont décrites ci-dessous :

1. première approche (*Suppression*). Elle consiste à simplement laisser tomber toutes les périodes incomplètes au début du panel. Étant donné que les périodes incomplètes sont exclues de l'analyse, les dates de début avant le commencement du panel (les temps calendaires initiaux) ne sont pas requis. Aucune hypothèse de modélisation supplémentaire n'est aussi imposée ;
2. deuxième approche (*Conditionnelle*). L'approche classique de l'utilisation d'une vraisemblance conditionnelle. L'application de cette méthode nécessite de connaître les temps calendaires initiaux, les temps de début des périodes avant introduction dans le panel. Mais aucune hypothèse de modélisation supplémentaire n'est indispensable. Le fait que plusieurs temps de transition avant le début du panel soient inconnus pour certains ou pour tous les individus dans le cas de l'EDTR est un désavantage pour cette approche ;
3. troisième approche (*Stationnaire*). Étant donné que les dates de début des périodes avant introduction dans le panel sont souvent inconnues pour de nombreux individus, nous avons considéré une approche qui suppose que le processus de renouvellement alterné d'être en emploi et sans-emploi (chômage) a atteint la stationnarité. En faisant cette hypothèse, cela nous donne plus de flexibilité dans le traitement des périodes interrompues par le début du panel et nous pouvons utiliser ces périodes non complètement observées pour faire une inférence même si les dates de début avant le commencement du panel sont manquantes. Cependant, supposer la stationnarité est

une hypothèse forte.

Étant donné que seuls quelques temps de transition avant et après le panel peuvent être connus, une attention particulière a été accordée aux informations requises pour chacun des trois cas.

Dans le Chapitre 4, nous avons simulé des données d'un processus de renouvellement alterné retardé, en utilisant une chaîne continue de Markov homogène. Ces données ont permis d'évaluer les performances de chacune des trois méthodes aussi bien en situation de stationnarité qu'en situation de non hypothèse de stationnarité. Nous constatons, pour différentes valeurs de T puis de n, qu'en présence d'hypothèse de stationnarité, les approches *Stationnaire* et *Conditionnelle* performent mieux avec une plus faible erreur quadratique moyenne (EQM). Dans ce contexte, l'approche *Suppression* donne les EMV avec de moins « bons » résultats. Par contre en absence d'hypothèse de stationnarité, les approches *Conditionnelle* et *Suppression* produisent les EMV de meilleure qualité. Toutefois, l'approche *Conditionnelle* semble plus robuste que les deux autres.

Nous concluons en discutant des extensions possibles à ce travail. Comme indiqué dans la Section 1.3, il existe de nombreuses questions intéressantes à traiter lors de l'application d'un modèle de renouvellement alterné (ou d'un modèle similaire) utilisant des données de panel d'enquête. Comme de nombreux problèmes ont déjà été soulevés dans la Section 1.3, nous ne les répétons pas ici. Étant donné que l'EDTR est un ensemble de données complexe, nous n'avons pas essayé de les analyser avec un processus de renouvellement alterné dans le cadre de ce mémoire. Nous avons plutôt résumé différentes approches qui ont été utilisées pour faire face aux données manquantes des premières et dernières périodes de temps présentes dans la base de données de ladite enquête. Nous portons également une attention particulière aux données exactes disponibles : à savoir si des données ont été collectées avant le début du panel. Ci-dessous dans les paragraphes 4, 5 et 6, nous suggérons d'autres

approches pour gérer ces données manquantes.

Rappelons que nous avons adopté trois approches de la littérature : Kovačević et Roberts (2007) pour l'approche *Suppression*, Cook et Lawless (2018) pour *Conditionnelle* ainsi que Alvarez (2006) pour *Stationnaire*. Voici d'autres variantes et extensions des approches que nous aurions pu adapter :

4. concernant l'approche *Suppression*, il y a deux extensions qui pourraient être intéressantes :
   (a) si on laisse tomber la première période, on peut en faire de même pour la dernière période incomplète. Cependant, comme la fin du panel devrait introduire une censure à droite de type 1 qui est indépendante et non informative, l'abandon de la dernière période ne devrait conduire qu'à « jeter » plus de données et à faire une mauvaise inférence ;
   (b) puisque la dernière période est traitée en utilisant une censure à droite indépendante, on pourrait adopter une approche similaire avec la première période incomplète et utiliser la fonction de survie évaluée à la longueur de segment observée. On ne « jetterait » pas ainsi les premières périodes de temps. Ce qui pourrait accroître la précision des EMV dans ce cas de figure ;
5. relativement à l'approche *Conditionnelle*, si les temps calendaires finaux des périodes après la fin du panel sont observés alors on pourrait aussi adopter une approche *Conditionnelle* avec les périodes sectionnées à la fin du panel. À notre connaissance, cependant, il est peu probable que ces données soient recueillies dans un panel d'enquête comme l'EDTR ;
6. à propos de l'approche *Stationnaire*, il y a deux extensions apparentes en supposant la stationnarité :
   (a) supposons qu'aucune donnée n'est collectée avant et après le panel (les temps calendaires initiaux et finaux ne seraient donc pas connus). Étant donné que

l'article original Alvarez (2006) suppose que le premier segment de temps de la période observé dans le panel suit une distribution de temps de récurrence avant (*forward recurrence time*), il est logique de faire une hypothèse similaire concernant la dernière période de temps dans le panel, c'est-à-dire écrire une vraisemblance qui modélise le segment observé de la dernière période en utilisant un temps de récurrence arrière (*backward recurrence time*). Cette approche combinerait donc l'hypothèse de stationnarité en début et en fin de panel en utilisant la fonction de densité du temps de récurrence arrière en fin de panel ;

(b) si les données avant l'introduction et après la sortie du panel sont collectées alors les temps calendaires initiaux et finaux seraient connus. Ainsi sans supposer la stationnarité, on utiliserait l'approche *Conditionnelle*. Cependant, une extension possible dans cette situation, serait de faire l'hypothèse de stationnarité et de modéliser ces temps en utilisant la densité (2.7) exprimée de manière appropriée pour chaque type de période. Cette approche est inspirée de Addona (2005).

Le Chapitre 2 présente les vraisemblances pour des distributions non spécifiées. Plus loin dans les Chapitres 3 et 4, nous n'utilisons que des densités exponentielles. Une extension immédiate à ceci serait de considérer une autre distribution telle que la distribution de Weibull. Contrairement à la distribution exponentielle, la distribution de Weibull n'est pas sans mémoire et les travaux d'optimisation n'auraient pas été faits manuellement mais auraient nécessité des programmes d'optimisation.

Par ailleurs, supposer que les périodes de temps sont, pour une unité i donnée, indépendantes n'est pas une hypothèse réaliste. Il est plus vraisemblable, par exemple, de penser que les périodes de temps d'emploi sont dépendantes de celles de chômage et vice-versa. On pourrait donc utiliser certains modèles plus proches de la réalité, tels que les modèles de fragilité, pour tenir compte de cette dépendance.

La non-prise en compte des premiers temps de séjour dans l'approche *Suppression* entraîne des difficultés venant du nombre de renouvellement X_j. Cette approche ne serait pas en adéquation avec les cas $\tau = 1$ et $\tau = 2$, où l'on n'observe qu'un ou deux renouvellements, pouvant occasionner des biais de sélection du fait qu'on pourrait ne pas considérer les temps les plus longs. Ces cas pourraient donc être traités dans la théorie et dans la simulation par d'autres approches (*Conditionnelle* et *Stationnaire*, par exemple). En annexe à la Section A.4, nous discutons de ces cas.

Bien que nous n'ayons considéré qu'un processus de renouvellement alterné, les approches pour traiter les périodes problématiques (en début et en fin de panel) présentées dans ce mémoire peuvent être utilisées pour d'autres modèles tels que des modèles multi-états plus généraux et des modèles de fragilité. Par ailleurs considérer un modèle à deux états (chômage et emploi), simplifie fortement la réalité. On pourrait donc utiliser un modèle à trois états ou plus (pour tenir compte des individus qui décident de prendre leur retraite). Les autres extensions générales liées au sujet traité dans ce mémoire pourraient être :

— de tenir compte de la pondération fournie par Statistique Canada dans l'analyse pour extrapolation à la population-cible en faisant une application aux données de l'EDTR ;

— inclure les covariables dans le modèle ;

— traiter les données manquantes des temps de transitions se trouvant à l'intérieur du panel.

ANNEXE A

A.1 Données de l'EPA

Dans cette section, nous décrivons brièvement la méthodologie de collecte des données de l'EPA dont la base de sondage sert dans l'élaboration des échantillons de l'EDTR. Différentes méthodes de collecte sont utilisées, chacune avec ses avantages et ses inconvénients selon le coût, le temps d'exécution et le taux de réponse. Elles dépendent aussi de plusieurs critères liés au ménage sélectionné, à savoir :

- les entrevues sur place assistées par ordinateur (IPAO) ;
- les entrevues téléphoniques assistées par ordinateur (ITAO) ;
- les entrevues Web assistées par ordinateur (IWAO) aussi dénommées interviews par questionnaires électroniques (QE).

Pour la collecte des données, on distingue deux types de ménage : les ménages qui sont à leur premier mois de participation à l'enquête (appelés les « naissances ») et les ménages faisant partie de l'enquête depuis plus de deux mois : entre deux et six mois. Ces derniers sont désignés par le terme « subséquents ».

Les ménages naissants sont joints par ITAO, si un numéro de téléphone résidentiel fonctionnel est obtenu. Sinon, la collecte pour ces ménages est faite par IPAO. Les subséquents sont sondés par QE s'ils satisfont plusieurs critères, notamment si à la fin de l'entrevue pour laquelle ils étaient naissants, ils ont fourni une adresse de courriel valide. Dans le cas où le répondant principal du ménage ne fournit pas d'adresse de courriel ou ne souhaite pas être joint par courriel, l'option proposée est de contacter le ménage par ITAO les mois suivants si un numéro de téléphone valide est fourni. Si le répondant principal n'opte pas pour cette proposition alors le ménage est contacté par IPAO.

Chaque mois, la méthode d'interview est réévaluée afin de faire passer les ménages joints par IPAO en ITAO ou QE et ceux joints par ITAO en QE, dans le but de réduire les coûts de collecte. Depuis la mise en place du QE, 20% des ménages subséquents sont sondés par QE, 76% le sont par ITAO et 4% par IPAO. Ces proportions étaient de 96% pour ITAO et 4% pour IPAO avant la mise en place du QE, (Statistique Canada, 2017).

Pour réduire le fardeau des répondants et les coûts de collecte tout en maintenant des taux de réponse élevés et une excellente qualité des données, l'EPA utilise plusieurs approches. Pour les aînés [1], par exemple, les réponses de l'entrevue initiale sont réutilisées. La rotation de l'échantillon fait aussi partie des approches utilisées par l'EPA dans cette optique.

A.2 Pondération de l'EDTR

Pour inférer les résultats de sondage au niveau de la population, un ensemble de poids est produit. Dans les Sections A.2.1, A.2.2 et A.2.3, nous discutons des trois types de poids utilisés pour l'estimation. Les poids Bootstrap ne sont pas discutés.

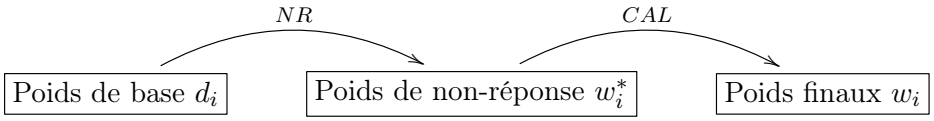
Pour chaque panel, il est produit de façon indépendante, après plusieurs transformations, les poids (longitudinaux, transversaux et combinés) finaux. Une illustration simplifiée est faite à la Figure A.1. Les poids initiaux d_i [2] s'interprètent comme étant le nombre d'unités dans la population que ladite unité représente, en plus d'elle-même.
Les poids de base sont ensuite ajustés pour tenir compte de la non-réponse (NR) et du calage [3] (CAL). Pour plus de détails, on consultera Dufour *et al.* (2001) et LaRoche (2007).

1. Les personnes de 70 ans et plus

2. Il s'agit ici des poids de base. Le poids de base d_i est le poids initial reçu et correspondant à l'inverse de la probabilité de sélection π_i de l'unité i dans l'échantillon S : $d_i = \pi_i^{-1}$.

3. Le calage est le processus d'ajustement des poids, dans notre contexte des poids pour la non-réponse, afin de faire coïncider les estimations avec des totaux connus de la population.

Figure A.1: Processus d'ajustement des poids



Dans le cas de l'EDTR, trois ensembles de poids finaux sont produits : les poids longitudinaux, les poids transversaux et les poids longitudinaux avec panels combinés qui sont associés à l'échantillon combiné sur une période de 3 ans lorsque les panels se chevauchent.

La pondération longitudinale sert dans la production d'estimations représentatives de la population au moment de la sélection de l'échantillon longitudinal tandis que la pondération transversale sert dans la production d'estimations représentatives de la population au 31 décembre de l'année de référence.

On peut noter, en guise d'illustration, que les estimations produites avec les poids longitudinaux du quatrième panel de l'EDTR sont représentatives de la population canadienne en âge de travailler au 1er janvier 2002 alors que celles produites avec les poids transversaux de la première année du même panel le sont pour la population générale au 31 décembre 2002.

Nous présenterons, certes, différents types de poids qui apparaissent dans la production de divers estimateurs issus de l'EDTR mais il est fort intéressant de remarquer que c'est l'échantillon longitudinal, et partant, les poids longitudinaux qui suscitent notre intérêt.

A.2.1 Pondération longitudinale

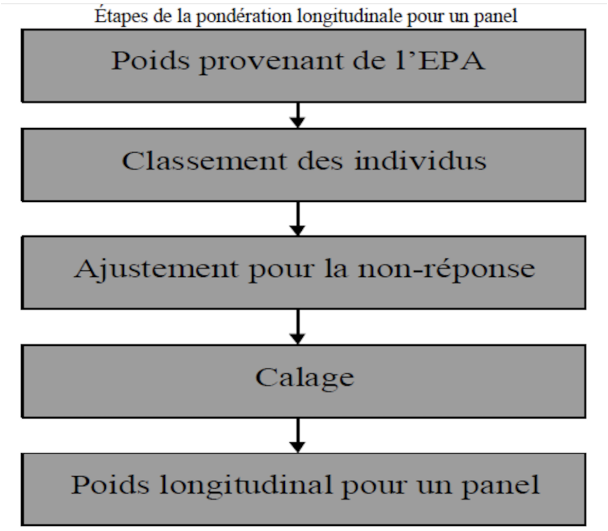
La pondération longitudinale consiste en la création d'un système de poids qui, appliqué aux caractéristiques d'intérêt collectées, permet de produire des estimations représentatives

de la population au moment de la sélection de l'échantillon longitudinal. Elle s'effectue selon cinq étapes que sont :

1. la détermination des poids initiaux ou poids de base ;
2. la classification des individus longitudinaux ;
3. la modélisation et l'ajustement pour la non-réponse ;
4. l'ajustement pour les valeurs influentes ;
5. le calage (pour ajuster les estimations de l'enquête à des totaux connus de la population).

Pour plus d'informations sur ces cinq étapes, voir LaRoche (2007).

Figure A.2: Étapes de la pondération longitudinale pour un panel - EDTR (Naud, 2004, p.8)



La Figure A.2 présente les différentes étapes dans la production de poids longitudinaux pour un panel telles que décrites dans les lignes précédentes.

A.2.2 Pondération transversale

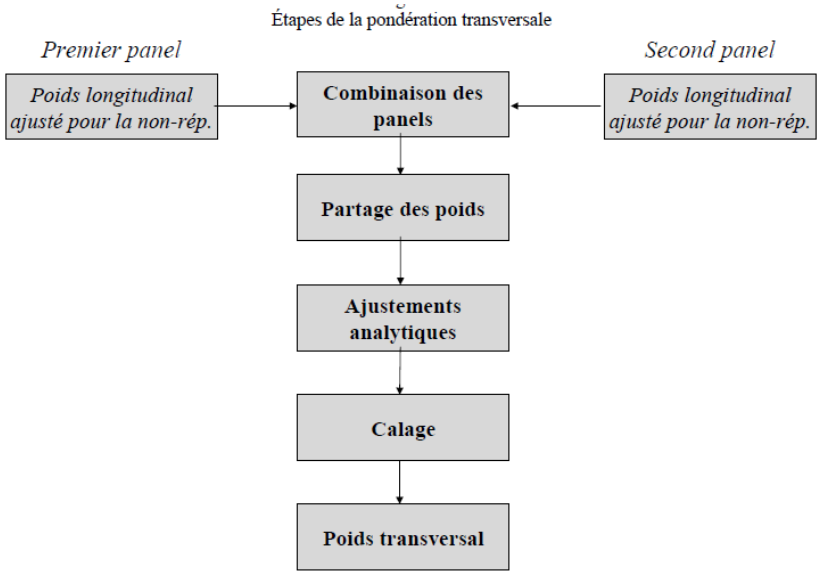
Dans la section précédente, nous avons présenté les étapes de la production de poids longitudinaux. Ces poids permettent l'élaboration des estimateurs représentatifs de la population lors de la sélection de l'échantillon longitudinal.

Dans la présente section, nous présentons les poids transversaux qui servent dans la production d'estimations représentatives de la population au 31 décembre de l'année de référence. Contrairement à la pondération longitudinale, il y a deux types de poids transversaux : le poids transversal intégré et le poids transversal individuel. Le poids transversal intégré est identique pour tous les individus d'un même ménage et sert dans la production d'estimateurs au niveau du ménage. Le poids transversal individuel, quant à lui, peut être différent d'un individu à l'autre y compris au sein d'un ménage particulier, et sert dans l'analyse au niveau des individus.

Par un processus semblable à celui de la Section A.2.1, les poids transversaux sont produits selon les étapes suivantes et dont les détails sont présentés dans LaRoche (2007) :

1. détermination des individus éligibles à la pondération transversale ;
2. ajustement pour la non-réponse ;
3. ajustement pour la migration inter-provinciale ;
4. application des facteurs d'allocation des panels ;
5. partage des poids ;
6. ajustement pour les valeurs influentes ;
7. calage sur marges.

Figure A.3: Étapes de la pondération transversale - EDTR (Naud, 2004, p.9)



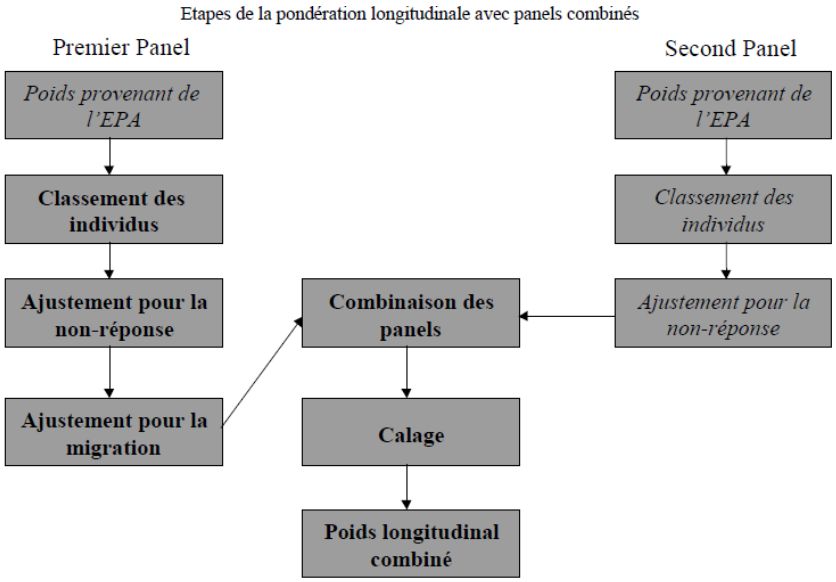
La Figure A.3 donne une vue des différentes étapes nécessaires dans l'obtention de poids transversaux pour un panel.

A.2.3 Pondération longitudinale avec panels combinés

Depuis le 1er janvier 1996, deux panels se chevauchent dans l'EDTR. Pour tirer profit de ces deux panels à la fois, le poids longitudinal combiné a été introduit. Ce type de poids permet une analyse longitudinale plus précise mais reste limité à une durée de trois ans, correspondant à la durée maximale du chevauchement des deux panels consécutifs (par exemple, de 2005 à 2007 dans le cas des panels 4 et 5 ; voir la Figure 1.2). La méthodologie de la pondération longitudinale combinée s'inspire des pondérations longitudinale et transversale de l'EDTR. Elle permet d'ajuster l'échantillon du panel le plus âgé et de le combiner au plus récent de façon à ce que le nouvel échantillon représente la population

cible au moment de l'introduction du second panel. Naud (2004) fournit plus de détails sur les étapes de la pondération longitudinale avec panels combinés.

Figure A.4: Étapes de la pondération longitudinale avec panels combinés - EDTR (Naud, 2004, p.13)



La Figure A.4, présente les étapes nécessaires dans l'obtention de poids longitudinaux combinés pour deux panels combinés.

A.3 Données de transition manquantes selon le nombre de périodes de chômage observées.

Dans ce mémoire, l'accent a été mis sur les périodes de temps interceptées par le début et la fin du panel mais il ne faut pas perdre de vue les autres complexités liées aux périodes de temps à l'intérieur du panel. Au nombre de celles-ci, on peut citer les dates de début et de fin manquantes à l'intérieur du panel pour certaines périodes.

Tableau A.1: Nombre de périodes de chômage dans le panel 4 de l'EDTR selon que la date de début ou la date de fin soit observée ou non

	Date de début				
	observée		non observée		
Périodes de chômage	Date de fin observée	Date de fin non observée	Date de fin observée	Date de fin non observée	Total
1	4 727	5 484	2 895	3 217	16 323
2	3 203	1 712	548	926	6 389
3	1 659	826	229	159	2 873
4	801	411	96	51	1 359
5 et plus	536	404	48	26	1 014
Total	10 926	8 837	3 816	4 379	27 958

.

Le Tableau A.1 donne un aperçu de la distribution des périodes de chômage dans le panel 4 de l'EDTR selon la date de début et la date de fin.

De ce tableau, on peut lire que sur 27 958 périodes de chômage, environ :

— 39% (10 926 périodes) sont des périodes complètes : avec dates de début et dates de fin observées ;

— 29% (3 816 + 4 379 = 8 195 périodes) ont des dates de début non observées ;

— 47% (8 837 + 4 379 = 13 216) de ces périodes ont des dates de fin non observées ;

— 32% (8 837 périodes) ont des dates de début observées et des dates de fin non observées ;

— 14% (3 816 périodes) ont des dates de début non observées et des dates de fin observées ;

— 16% (4 379 périodes) ont des dates de début et dates de fin non observées.

A.4 Traitement des cas $\tau = 1$ et $\tau = 2$

Le fait de négliger les premières périodes de temps dans l'approche Suppression mène à des problèmes liés au nombre de renouvellement X_j. En effet, cette approche ne pourrait être implémentée pour un individu n'ayant expérimenté qu'une seule période de temps ; cette période devant être supprimée selon la procédure inhérente à ladite approche. Nous rejetons dans cette situation de 50% à 100% des données de cet individu. Par ailleurs, dans la situation où l'individu expérimente deux renouvellements X_j, ne pas prendre en compte la première période de temps revient presque à rejeter 25% à 50% des données de cet individu. S'il arrive que la période de temps à rejeter occupe une grande partie du panel, la rejeter serait préoccupant.

Dans les prochaines sections, nous traitons néanmoins les cas où l'on observe un ou deux renouvellements ($\tau = 1$ et $\tau = 2$).

• Le cas $\tau = 1$. Dans ce cas de figure, l'individu expérimente un seul renouvellement. Cela correspond à la situation où l'on observe dans le panel pour cet individu.

Tableau A.2: Différents cas de figure pour $\tau = 1$.

Périodes de temps	Commentaires
Une unique période de chômage	z_1 = segment sectionné au début et à la fin y_1 = non observé
Une unique période d'emploi	$z_1 = 0$ y_1 = segment sectionné au début et à la fin
Une période de chômage suivie d'une période d'emploi	z_1 =segment sectionné au début y_1 = segment sectionné à la fin

Dans cette éventualité, la vraisemblance pour le cas $\tau = 1$ pour un individu i notée L_i^1 sera définie par :

$$\begin{aligned} L_i^1 :=& [f_{C,\text{début}}(z_{i1})]^{\mathbb{1}(w_i(0)=1)} \times [f_{E,\text{fin}}(y_{i1})]^{\mathbb{1}(w_i(T)=0)} \\ & \times [f_{C,\text{début}}(z_{i1})]^{\mathbb{1}(w_i(0)=1)\mathbb{1}(w_i(T)=1)} \\ & \times [f_{E,\text{début}}(y_{i1})]^{\mathbb{1}(w_i(0)=0)\mathbb{1}(w_i(T)=0)}. \end{aligned} \tag{A.1}$$

- Le cas $\tau = 2$. Il correspond à la situation où l'individu expérimente.

Tableau A.3: Différents cas de figure pour $\tau = 2$.

Périodes de temps	Commentaires
Une période de chômage suivie d'une période d'emploi et d'une autre période de chômage	z_1 = segment sectionné au début, y_1 = complet, z_2 = segment sectionné à la fin et y_2 = non observé
Une période de chômage suivie d'une période d'emploi et une autre période de chômage suivie d'une autre période d'emploi	z_1 = segment sectionné au début, y_1 = complet, z_2 = complet et y_2 = segment sectionné à la fin
Une période d'emploi suivie d'une période de chômage	$z_1 = 0$, y_1 = segment sectionné au début z_2 = segment sectionné à la fin et y_2 = non observé
Une période d'emploi suivie d'une période de chômage et d'une autre période d'emploi	$z_1 = 0$, y_1 = segment sectionné au début z_2 = complet et y_2 = segment sectionné à la fin

Dans le Tableau A.3, les périodes complètes sont celles qui sont entièrement observées.

En notant L_i^2 la vraisemblance associée à $\tau = 2$ pour un individu i, nous aurons :

$$\begin{aligned} L_i^2 :=& [f_{C,\text{début}}(z_{i1}) f_E(y_{i1})]^{\mathbb{1}(w_i(0)=1)} \times [f_{E,\text{début}}(y_{i1})]^{\mathbb{1}(w_i(0)=0)} \\ & \times [f_{C,\text{fin}}(z_{i2})]^{\mathbb{1}(w_i(T)=1)} \\ & \times [f_C(z_{i2}) f_{E,\text{fin}}(y_{i2})]^{\mathbb{1}(w_i(T)=0)}. \end{aligned} \tag{A.2}$$

A.4.1 Approche Suppression (Non-prise en compte du premier temps de séjour) : traitement des cas $\tau = 1$ et $\tau = 2$

Dans cette section, nous adaptons la non-prise en compte du premier temps de séjour avec les cas $\tau = 1$ et $\tau = 2$.

Pour le cas $\tau = 1$, seule l'éventualité $X_1 = Z_1 + Y_1$ sera effective. Dans cette configuration, la période Y_1 ne sera que la seule période considérée. Ainsi l'équation (A.1) deviendra

$$L_i^1 = [S_E(y_{i1})]^{\mathbb{1}(w_i(T)=0)}, \tag{A.3}$$

qui ne peut être optimisée en λ_E ;

pour le cas $\tau = 2$, dans une approche équivalente du cas précédent, l'équation (A.2) aboutira à

$$\begin{aligned} L_i^2 =& [f_E(y_{i1})]^{\mathbb{1}(w_i(0)=1)} \times [S_C(z_{i2})]^{\mathbb{1}(w_i(T)=1)} \\ & \times [f_C(z_{i2}) S_E(y_{i2})]^{\mathbb{1}(w_i(T)=0)}. \end{aligned} \tag{A.4}$$

Avec les densités exponentielles, pour n individus de ce type, nous obtenons :

$$\begin{aligned} \ln L_2 &= \sum_{i=1}^{n} \ln L_i^2 \\ &= d_0 \ln \lambda_C + r_1 \ln \lambda_E - \lambda_C \, C(T) - \lambda_E \, E(T), \end{aligned}$$

où

$$\begin{aligned} C(T) &= \sum_{i=1}^{n} z_{i2}, \\ E(T) &= \sum_{i=1}^{n} y_{i1} \mathbb{1}(w_i(0) = 1) + \sum_{i=1}^{n} y_{i2} \mathbb{1}(w_i(T) = 0). \end{aligned}$$

L'optimisation donne

$$\hat{\lambda}_C = \frac{d_0}{C(T)}$$

et

$$\hat{\lambda}_E = \frac{r_1}{E(T)}.$$

A.4.2 Approche Conditionnelle : traitement des cas $\tau = 1$ et $\tau = 2$

Pour $\tau = 1$, les équations (A.1) et (3.3) donneront :

$$\begin{aligned} L_i^1 = [f_C(z_{i1})]^{\mathbb{1}(w_i(0)=1)}[S_E(y_{i1})]^{\mathbb{1}(w_i(T)=0)} \times [f_C(z_{i1})]^{\mathbb{1}(w_i(0)=1)\mathbb{1}(w_i(T)=1)} \\ \times [f_E(y_{i1})]^{\mathbb{1}(w_i(0)=0)\mathbb{1}(w_i(T)=0)}. \end{aligned} \tag{A.5}$$

Pour n individus de ce type, on a plutôt

$$\begin{aligned} \ln L_1 &:= \sum_{i=1}^{n} \ln L_i^1 \\ &= (r_1 + \sum_{i=1}^{n} (\mathbb{1}(w_i(0) = 1)\mathbb{1}(w_i(T) = 1))) \ln \lambda_C \\ &\quad + (\sum_{i=1}^{n} (\mathbb{1}(w_i(0) = 0)\mathbb{1}(w_i(T) = 0))) \ln \lambda_E - \lambda_C \, C(T) - \lambda_E \, E(T) \end{aligned} \tag{A.6}$$

avec

$$C(T) = \sum_{i=1}^{n} z_{i1} \left(\mathbb{1}(w_i(0) = 1) + \mathbb{1}(w_i(0) = 1)\mathbb{1}(w_i(T) = 1)\right),$$

$$E(T) = \sum_{i=1}^{n} y_{i1} \left(\mathbb{1}(w_i(T) = 0) + \mathbb{1}(w_i(0) = 0)\mathbb{1}(w_i(T) = 0)\right).$$

L'optimisation donne

$$\hat{\lambda}_C = \frac{r_1 + \sum_{i=1}^{n} (\mathbb{1}(w_i(0) = 1)\mathbb{1}(w_i(T) = 1))}{C(T)}$$

et

$$\hat{\lambda}_E = \frac{\sum_{i=1}^{n} (\mathbb{1}(w_i(0) = 0)\mathbb{1}(w_i(T) = 0))}{E(T)}.$$

Pour l'approche conditionnelle, nous avons montré que
$f_{C,\text{début}} = f_C$ et $f_{E,\text{début}} = f_E$ (voir équation (3.3)). Alors lorsque $\tau = 2$ l'équation (A.2) donne

$$\begin{aligned} L_i^2 =& [f_C(z_{i1}) f_E(y_{i1})]^{\mathbb{1}(w_i(0)=1)} \times [f_E(y_{i1})]^{\mathbb{1}(w_i(0)=0)} \\ & \times [S_C(z_{i2})]^{\mathbb{1}(w_i(T)=1)} \times [f_C(z_{i2}) S_E(y_{i2})]^{\mathbb{1}(w_i(T)=0)}. \end{aligned} \tag{A.7}$$

Et ainsi on a

$$\begin{aligned} \ln L^2 &= \sum_{i=1}^{n} \ln L_i^2 \\ &= (r_1 + d_0) \ln \lambda_C + \ln \lambda_E - \lambda_C \, C(T) - \lambda_E \, E(T) \end{aligned}$$

avec

$$\begin{aligned} C(T) &= \sum_{i=1}^{n} z_{i1} \mathbb{1}(w_i(0) = 1) + \sum_{i=1}^{n} z_{i2}, \\ E(T) &= \sum_{i=1}^{n} y_{i1} + \sum_{i=1}^{n} y_{i2} \mathbb{1}(w_i(T) = 0). \end{aligned}$$

Les résultats de l'optimisation sont

$$\hat{\lambda}_C = \frac{r_1 + d_0}{C(T)}$$

et

$$\hat{\lambda}_E = \frac{1}{E(T)}.$$

A.4.3 Stationnarité : traitement des cas $\tau = 1$ et $\tau = 2$

Dans cette section, nous faisons référence aux équations (3.5) et (3.6) pour déterminer les vraisemblances dans l'hypothèse de stationnarité pour les cas $\tau = 1$ et $\tau = 2$. Dans le cas $\tau = 1$, l'équation (A.1) devient

$$\begin{aligned}
L_i^1 = & \left[\frac{\lambda_E}{\lambda_C + \lambda_E}\lambda_C \exp\left(-\lambda_C\ z_{i1}\right)\right]^{\mathbb{1}(w_i(0)=1)} \left[\exp\left(-\lambda_E\ y_{i1}\right)\right]^{\mathbb{1}(w_i(T)=0)} \\
& \times \left[\frac{\lambda_E}{\lambda_C + \lambda_E}\lambda_C \exp\left(-\lambda_C\ z_{i1}\right)\right]^{\mathbb{1}(w_i(0)=1)\mathbb{1}(w_i(T)=1)} \\
& \times \left[\frac{\lambda_C}{\lambda_C + \lambda_E}\lambda_E \exp\left(-\lambda_E\ y_{i1}\right)\right]^{\mathbb{1}(w_i(0)=0)\mathbb{1}(w_i(T)=0)} \\
= & \lambda_C^{\mathbb{1}(w_i(0)=1)+\mathbb{1}(w_i(0)=1)\mathbb{1}(w_i(T)=1)+\mathbb{1}(w_i(0)=0)\mathbb{1}(w_i(T)=0)} \\
& \times \lambda_E^{\mathbb{1}(w_i(0)=1)+\mathbb{1}(w_i(0)=1)\mathbb{1}(w_i(T)=1)+\mathbb{1}(w_i(0)=0)\mathbb{1}(w_i(T)=0)} \\
& \times \left(\frac{1}{\lambda_C + \lambda_E}\right)^{\mathbb{1}(w_i(0)=1)+\mathbb{1}(w_i(0)=1)\mathbb{1}(w_i(T)=1)+\mathbb{1}(w_i(0)=0)\mathbb{1}(w_i(T)=0)} \\
& \times \left[\exp\left(-\lambda_C\ z_{i1}\ (\mathbb{1}(w_i(0)=1) + \mathbb{1}(w_i(0)=1)\mathbb{1}(w_i(T)=1))\right)\right] \\
& \times \left[\exp\left(-\lambda_E\ y_{i1}\ (\mathbb{1}(w_i(T)=0) + \mathbb{1}(w_i(0)=0)\mathbb{1}(w_i(T)=0))\right)\right].
\end{aligned} \tag{A.8}$$

Pour n individus *iid* de ce type, on a plutôt

$$\begin{aligned}\ln L^1 &= \sum_{i=1}^{n} \ln L_i^1 \\ &= \kappa(T) \ln \lambda_C + \kappa(T) \ln \lambda_E - \kappa(T) \ln (\lambda_C + \lambda_E) - \lambda_C \, C(T) - \lambda_E\end{aligned}$$

avec

$$\begin{aligned}\kappa(T) &= r_1 + \sum_{i=1}^{n} \mathbb{1}(w_i(0) = 1)\mathbb{1}(w_i(T) = 1) + \sum_{i=1}^{n} \mathbb{1}(w_i(0) = 0)\mathbb{1}(w_i(T) = 0), \\ C(T) &= \sum_{i=1}^{n} z_{i1}(\mathbb{1}(w_i(0) = 1) + \mathbb{1}(w_i(0) = 1)\mathbb{1}(w_i(T) = 1), \\ E(T) &= \sum_{i=1}^{n} y_{i1}(\mathbb{1}(w_i(T) = 0) + \mathbb{1}(w_i(0) = 0)\mathbb{1}(w_i(T) = 0).\end{aligned}$$

L'optimisation aboutit à

si $C(T) = E(T)$,

$$\hat{\lambda}_C = \frac{\kappa(T)}{2 \, C(T)};$$

si $C(T) \neq E(T)$,

$$\hat{\lambda}_C = \frac{\kappa(T) \times (C(T) - \sqrt{(C(T) \, (2 \, C(T) - E(T)))})}{C(T) \, (E(T) - C(T))}.$$

Dans les deux cas, on a

$$\hat{\lambda}_E = \frac{\kappa(T) - \hat{\lambda}_C \, C(T)}{E(T)}.$$

Pour le cas $\tau = 2$, l'équation (A.2) donne

$$\begin{aligned}L_i^2 = &\left[\left(\frac{\lambda_E}{\lambda_C + \lambda_E}\right) \lambda_C \exp(-\lambda_C \, z_{i1}) \, \lambda_E \exp(-\lambda_E \, y_{i1})\right]^{\mathbb{1}(w_i(0)=1)} \\ &\times \left[\left(\frac{\lambda_C}{\lambda_C + \lambda_E}\right) \lambda_E \exp(-\lambda_E \, y_{i1})\right]^{\mathbb{1}(w_i(0)=0)} \\ &\times [\exp(-\lambda_C \, z_{i2})]^{\mathbb{1}(w_i(T)=1)} \times [\lambda_C \exp(-\lambda_C \, z_{i2}) \, \exp(-\lambda_E \, y_{i2})]^{\mathbb{1}(w_i(T)=0)}.\end{aligned}$$

Avec n individus, on a plutôt

$$\begin{aligned}\ln L &= \sum_{i=1}^{n} \ln l_i \\ &= (n+d_0)\ln\lambda_C + (n+r_1)\ln\lambda_E - n\ln(\lambda_C+\lambda_E) - \lambda_C\, C(T) - \lambda_E\, E(T)\end{aligned}$$

où

$$\begin{aligned}C(T) &= \sum_{i=1}^{n} z_{i1}\mathbb{1}(w_i(0)=1) + \sum_{i=1}^{n} z_{i2}, \\ E(T) &= \sum_{i=1}^{n} y_{i1} + \sum_{i=1}^{n} y_{i2}\mathbb{1}(w_i(T)=0).\end{aligned}$$

Les résultats de l'optimisation donnent

si $C(T) = E(T)$

$$\begin{aligned}\hat{\lambda}_C &= \frac{(n+d_0)(n+d_0+r_1)}{C(T)\,(2n+d_0+r_1)}, \\ \hat{\lambda}_E &= \frac{(n+d_0+r_1) - \hat{\lambda}_C\, C(T)}{C(T)};\end{aligned}$$

si $C(T) < E(T)$

$$\begin{aligned}\hat{\lambda}_C &= \frac{-\beta(T) + \sqrt{(\beta(T))^2 - 4\alpha(T)\,\gamma(T)}}{2\alpha(T)}, \\ \hat{\lambda}_E &= \frac{(n+d_0+r_1) - \hat{\lambda}_C\, C(T)}{E(T)};\end{aligned}$$

si $C(T) > E(T)$

$$\begin{aligned}\hat{\lambda}_C &= \frac{-\beta(T) - \sqrt{(\beta(T))^2 - 4\alpha(T)\,\gamma(T)}}{2a(T)}, \\ \hat{\lambda}_E &= \frac{(n+d_0+r_1) - \hat{\lambda}_C\, C(T)}{E(T)}\end{aligned}$$

avec

$$\begin{aligned}\alpha(T) &= C(T)[C(T) - E(T)], \\ \beta(T) &= (2n + 2d_0 + r_1)\, C(T) - d_0\, E(T), \\ \gamma(T) &= -(n+d_0)(n+d_0+r_1)\,.\end{aligned}$$

A.5 Résultats

Dans les prochaines sections, nous présentons les résultats des simulations faites dans le Chapitre 4 pour lesquelles l'hypothèse d'une chaîne continue de Markov homogène a été faite et l'hypothèse de stationnarité est vérifiée (voir le Tableau A.4).

La Section A.5.1 présente les résultats pour une taille d'échantillon n fixée tout en faisant varier la largeur T du panel. Ces résultats sont similaires lorsqu'on fixe la largeur du panel et qu'on fait varier les tailles d'échantillon. Ces derniers n'ont donc pas été présentés.

Nous avons comparé en début de panel, à $t = 0$, les proportions observées des individus en état de chômage (respectivement en état d'emploi) avec la probabilité théorique d'être en chômage (respectivement en emploi). Les résultats sont présentés dans le Tableau A.4.

Tableau A.4: Proportions observées des individus ayant commencé le panel en chômage et en emploi.

Stationnarité	Largeur fenêtre T	Taille échantillon n	Proportions observées		Proportions théoriques	
			Chômage	Emploi	Chômage	Emploi
Avec	13	10	0.048	0.952	0.0476	0.9524
Sans	13	10	0.680	0.320	0.0476	0.9524
Avec	26	10	0.048	0.952	0.0476	0.9524
Sans	26	10	0.675	0.325	0.0476	0.9524
Avec	52	10	0.046	0.954	0.0476	0.9524
Sans	52	10	0.691	0.309	0.0476	0.9524
Avec	13	30	0.046	0.953	0.0476	0.9524
Sans	13	30	0.670	0.330	0.0476	0.9524
Avec	26	30	0.046	0.954	0.0476	0.9524
Sans	26	30	0.660	0.340	0.0476	0.9524
Avec	52	30	0.047	0.953	0.0476	0.9524
Sans	52	30	0.670	0.330	0.0476	0.9524
Avec	13	100	0.047	0.953	0.0476	0.9524
Sans	13	100	0.668	0.332	0.0476	0.9524
Avec	26	100	0.049	0.951	0.0476	0.9524
Sans	26	100	0.667	0.333	0.0476	0.9524
Avec	52	100	0.047	0.953	0.0476	0.9524
Sans	52	100	0.682	0.318	0.0476	0.9524

Le Tableau A.4 présente, en début de panel, à $t = 0$, les proportions observées des individus en état de chômage avec la probabilité théorique d'être en chômage et les équivalents pour les individus en état d'emploi. Ces probabilités théoriques sont $p_C = \dfrac{\lambda_E}{\lambda_C + \lambda_E}$ et $p_E = \dfrac{\lambda_C}{\lambda_C + \lambda_E} = 1 - p_C$ et valent $p_C = 0.0476$ et $p_E = 0.9524$ respectivement.

A.5.1 Résultats de simulation en fixant la taille n de l'échantillon et en faisant varier la durée T de la fenêtre d'observation

Dans la suite, les valeurs présentées dans les différents tableaux sont les moyennes des EMV pour les trois approches obtenues par la méthode de Monte-Carlo avec une taille échantillonnale $B = 100$.

Les valeurs obtenues dans les 2^{e} et 3^{e} colonnes du Tableau A.5 (ainsi que les tableaux équivalents) sont les moyennes des résultats de simulations répliquées 100 fois.

Tableau A.5: Résultats pour $n = 10$ et $T = 13$

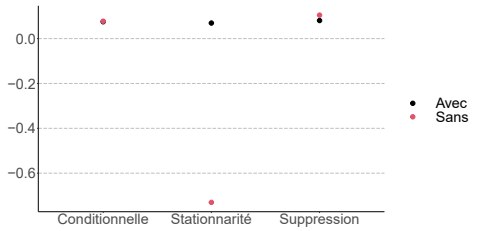
Paramètre	Valeurs moyennes des résultats de simulations répliquées 100 fois	
	Avec hypothèse de stationnarité	Sans hypothèse de stationnarité
$\hat{\lambda}_C$ Suppression	4.081	4.105
$\hat{\lambda}_E$ Suppression	0.190	0.198
$\hat{\lambda}_C$ Conditionnelle	4.075	4.078
$\hat{\lambda}_E$ Conditionnelle	0.195	0.201
$\hat{\lambda}_C$ Stationnaire	4.069	3.270
$\hat{\lambda}_E$ Stationnaire	0.195	0.251
λ_C Théorique	4.000	4.000
λ_E Théorique	0.200	0.200

Notes de légende : Dans les figures qui suivent, « Avec » signifie que la stationnarité est atteinte alors qu'elle ne l'est pas pour « Sans ».

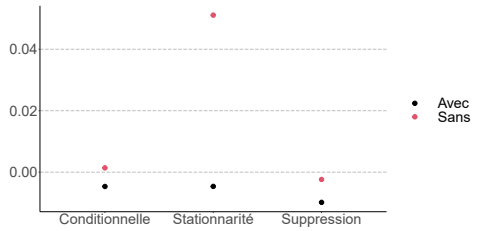
Sur la Figure A.5 (a), les biais de $\hat{\lambda}_C$ en présence ou non d'hypothèse de stationnarité coïncident pour les approches Conditionnelle et Suppression. Le biais est faible pour les deux estimateurs dans les approches Conditionnelle et Suppression aussi bien lorsque la stationnarité est vérifiée (« Avec ») que lorsqu'elle ne l'est pas (« Sans »). L'approche Stationnarité est presque sans biais dans le cas « Avec » tandis que dans le cas « Sans » $\hat{\lambda}_C$ a un biais négatif et un biais positif pour $\hat{\lambda}_E$. Dans l'ensemble des simulations, nous observons que l'écart-type de $\hat{\lambda}_E$ est très inférieur à celui de $\hat{\lambda}_C$. Et ces valeurs tendent vers 0 plus la

Figure A.5: Biais et écarts-types de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 10$ et $T = 13$

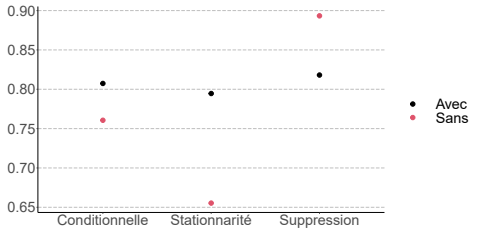
(a) Biais de $\hat{\lambda}_C$ pour $n = 10$ et $T = 13$



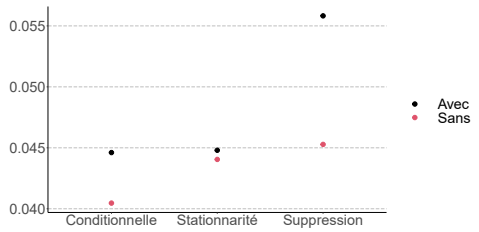
(b) Biais de $\hat{\lambda}_E$ pour $n = 10$ et $T = 13$



(c) Écart-type de $\hat{\lambda}_C$ pour $n = 10$ et $T = 13$



(d) Écart-type de $\hat{\lambda}_E$ pour $n = 10$ et $T = 13$



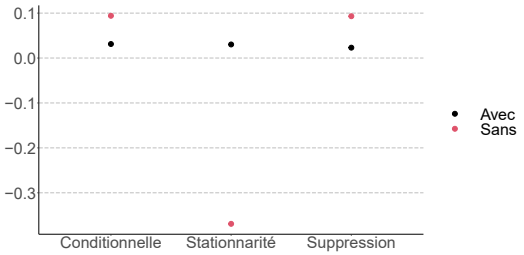
largeur de T ou de n s'accroît. Comme nous l'avons mentionné au Chapitre 4, cela fait du sens vu l'hypothèse d'homogénéité faite en ce qui concerne les individus. Par conséquent, il n'y a aucune différence entre observer un temps de séjour pour un individu donné ou observer celui d'un autre. Plus T ou n sera grand, plus il y aura de données et moins grand sera l'écart-type.

Tableau A.6: Résultats pour $n = 10$ et $T = 26$

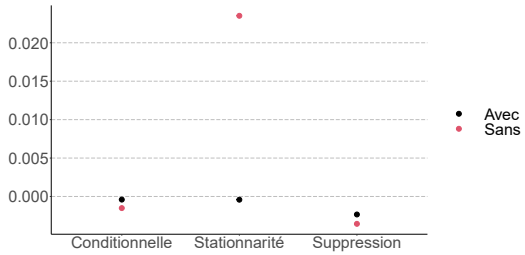
Paramètre	Valeurs moyennes des résultats de simulations répliquées 100 fois	
	Avec hypothèse de stationnarité	Sans hypothèse de stationnarité
$\hat{\lambda}_C$ Suppression	4.023	4.093
$\hat{\lambda}_E$ Suppression	0.198	0.196
$\hat{\lambda}_C$ Conditionnelle	4.031	4.094
$\hat{\lambda}_E$ Conditionnelle	0.200	0.198
$\hat{\lambda}_C$ Stationnaire	4.030	3.631
$\hat{\lambda}_E$ Stationnaire	0.200	0.224
λ_C Théorique	4.000	4.000
λ_E Théorique	0.200	0.200

Figure A.6: Biais et écarts-types de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 10$ et $T = 26$

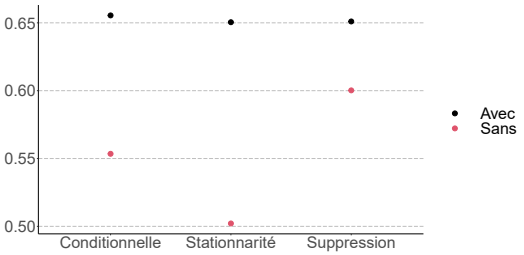
(a) Biais de $\hat{\lambda}_C$ pour $n = 10$ et $T = 26$



(b) Biais de $\hat{\lambda}_E$ pour $n = 10$ et $T = 26$



(c) Écart-type de $\hat{\lambda}_C$ pour $n = 10$ et $T = 26$



(d) Écart-type de $\hat{\lambda}_E$ pour $n = 10$ et $T = 26$

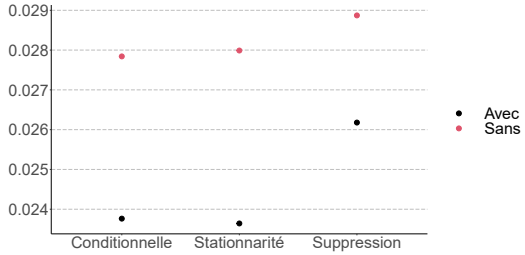
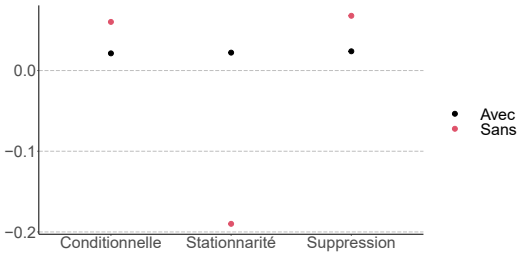


Tableau A.7: Résultats pour $n = 10$ et $T = 52$

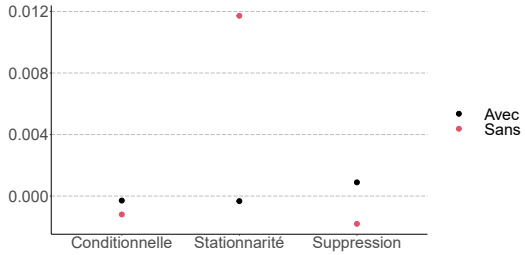
Paramètre	Valeurs moyennes des résultats de simulations répliquées 100 fois	
	Avec hypothèse de stationnarité	Sans hypothèse de stationnarité
$\hat{\lambda}_C$ Suppression	4.024	4.068
$\hat{\lambda}_E$ Suppression	0.201	0.198
$\hat{\lambda}_C$ Conditionnelle	4.021	4.060
$\hat{\lambda}_E$ Conditionnelle	0.200	0.199
$\hat{\lambda}_C$ Stationnaire	4.022	3.810
$\hat{\lambda}_E$ Stationnaire	0.200	0.212
λ_C Théorique	4.000	4.000
λ_E Théorique	0.200	0.200

Figure A.7: Biais et écarts-types de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 10$ et $T = 52$

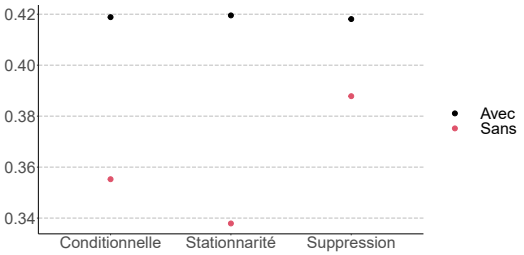
(a) Biais de $\hat{\lambda}_C$ pour $n = 10$ et $T = 52$



(b) Biais de $\hat{\lambda}_E$ pour $n = 10$ et $T = 52$



(c) Écart-type de $\hat{\lambda}_C$ pour $n = 10$ et $T = 52$



(d) Écart-type de $\hat{\lambda}_E$ pour $n = 10$ et $T = 52$

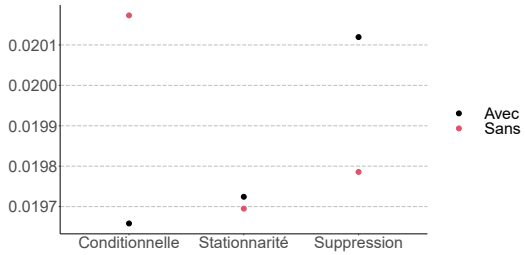
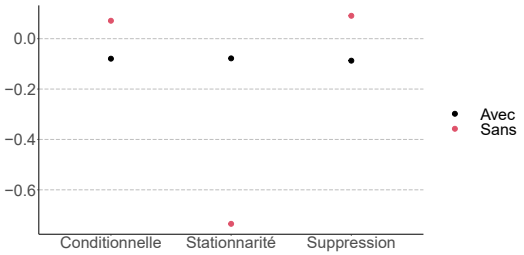


Tableau A.8: Résultats pour $n = 30$ et $T = 13$

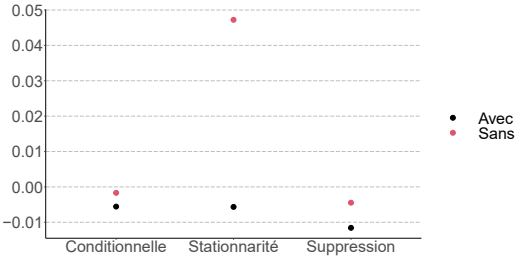
Paramètre	Valeurs moyennes des résultats de simulations répliquées 100 fois	
	Avec hypothèse de stationnarité	Sans hypothèse de stationnarité
$\hat{\lambda}_C$ Suppression	3.913	4.091
$\hat{\lambda}_E$ Suppression	0.188	0.196
$\hat{\lambda}_C$ Conditionnelle	3.921	4.071
$\hat{\lambda}_E$ Conditionnelle	0.194	0.198
$\hat{\lambda}_C$ Stationnaire	3.922	3.265
$\hat{\lambda}_E$ Stationnaire	0.194	0.247
λ_C Théorique	4.000	4.000
λ_E Théorique	0.200	0.200

Figure A.8: Biais et écarts-types de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 30$ et $T = 13$

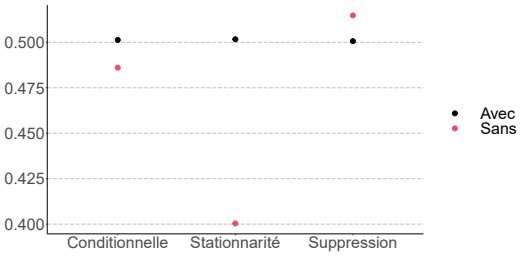
(a) Biais de $\hat{\lambda}_C$ pour $n = 30$ et $T = 13$



(b) Biais de $\hat{\lambda}_E$ pour $n = 30$ et $T = 13$



(c) Écart-type de $\hat{\lambda}_C$ pour $n = 30$ et $T = 13$



(d) Écart-type de $\hat{\lambda}_E$ pour $n = 30$ et $T = 13$

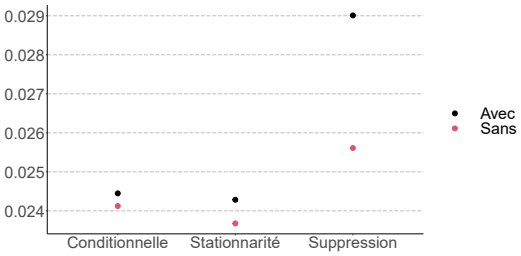
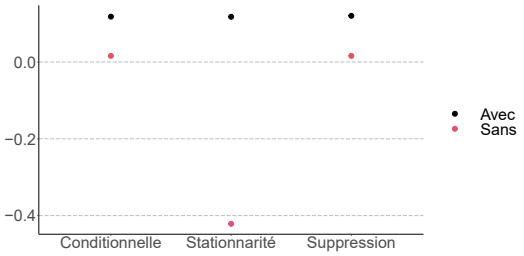


Tableau A.9: Résultats pour $n = 30$ et $T = 26$

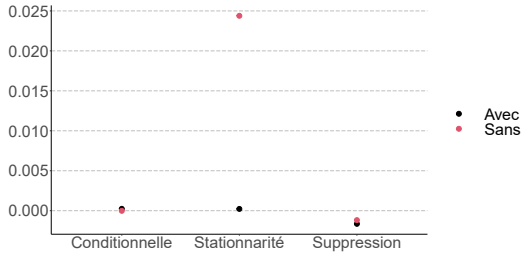
Paramètre	Valeurs moyennes des résultats de simulations répliquées 100 fois	
	Avec hypothèse de stationnarité	Sans hypothèse de stationnarité
$\hat{\lambda}_C$ Suppression	4.120	4.016
$\hat{\lambda}_E$ Suppression	0.198	0.199
$\hat{\lambda}_C$ Conditionnelle	4.118	4.016
$\hat{\lambda}_E$ Conditionnelle	0.200	0.200
$\hat{\lambda}_C$ Stationnaire	4.118	3.579
$\hat{\lambda}_E$ Stationnaire	0.200	0.224
λ_C Théorique	4.000	4.000
λ_E Théorique	0.200	0.200

Figure A.9: Biais et écarts-types de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 30$ et $T = 26$

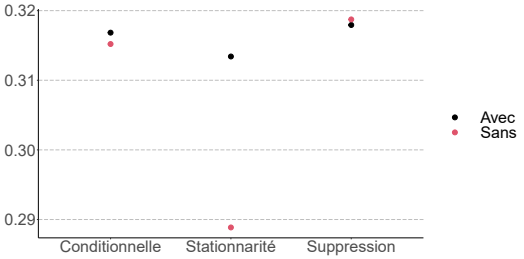
(a) Biais de $\hat{\lambda}_C$ pour $n = 30$ et $T = 26$



(b) Biais de $\hat{\lambda}_E$ pour $n = 30$ et $T = 26$



(c) Écart-type de $\hat{\lambda}_C$ pour $n = 30$ et $T = 26$



(d) Écart-type de $\hat{\lambda}_E$ pour $n = 30$ et $T = 26$

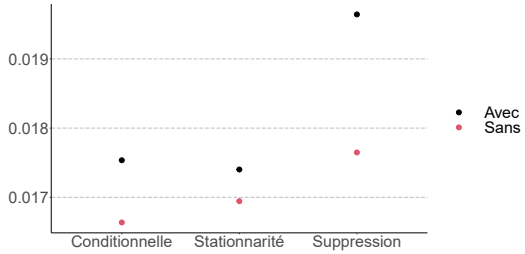
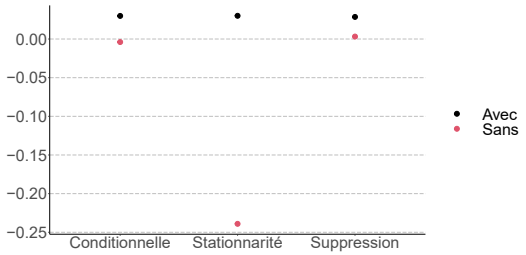


Tableau A.10: Résultats pour $n = 30$ et $T = 52$

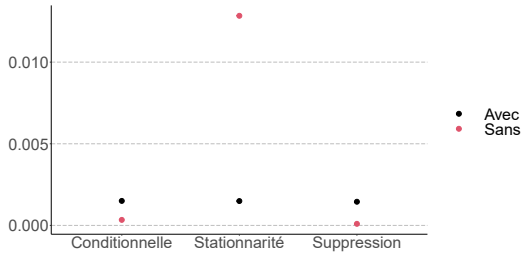
Paramètre	Valeurs moyennes des résultats de simulations répliquées 100 fois	
	Avec hypothèse de stationnarité	Sans hypothèse de stationnarité
$\hat{\lambda}_C$ Suppression	4.029	4.003
$\hat{\lambda}_E$ Suppression	0.201	0.200
$\hat{\lambda}_C$ Conditionnelle	4.030	3.996
$\hat{\lambda}_E$ Conditionnelle	0.202	0.200
$\hat{\lambda}_C$ Stationnaire	4.030	3.761
$\hat{\lambda}_E$ Stationnaire	0.201	0.213
λ_C Théorique	4.000	4.000
λ_E Théorique	0.200	0.200

Figure A.10: Biais et écarts-types de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 30$ et $T = 52$

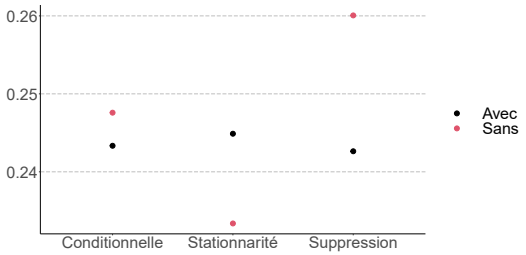
(a) Biais de $\hat{\lambda}_C$ pour $n = 30$ et $T = 52$



(b) Biais de $\hat{\lambda}_E$ pour $n = 30$ et $T = 52$



(c) Écart-type de $\hat{\lambda}_C$ pour $n = 30$ et $T = 52$



(d) Écart-type de $\hat{\lambda}_E$ pour $n = 30$ et $T = 52$

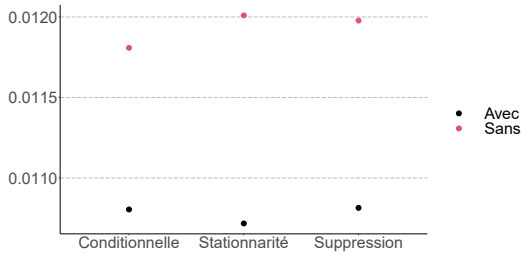
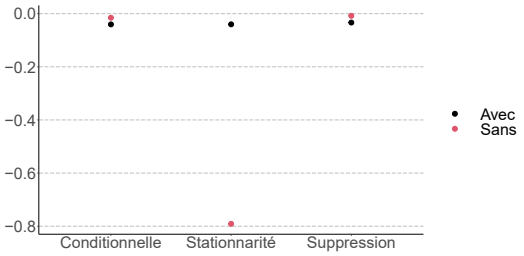


Tableau A.11: Résultats pour $n = 100$ et $T = 13$

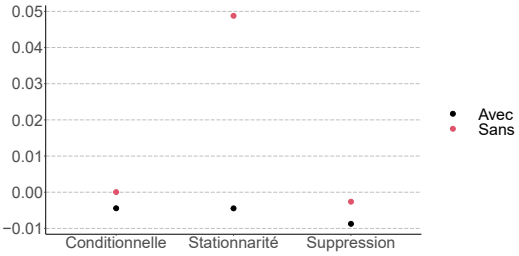
Paramètre	Valeurs moyennes des résultats de simulations répliquées 100 fois	
	Avec hypothèse de stationnarité	Sans hypothèse de stationnarité
$\hat{\lambda}_C$ Suppression	3.967	3.992
$\hat{\lambda}_E$ Suppression	0.191	0.197
$\hat{\lambda}_C$ Conditionnelle	3.960	3.985
$\hat{\lambda}_E$ Conditionnelle	0.196	0.200
$\hat{\lambda}_C$ Stationnaire	3.960	3.209
$\hat{\lambda}_E$ Stationnaire	0.196	0.249
λ_C Théorique	4.000	4.000
λ_E Théorique	0.200	0.200

Figure A.11: Biais et écarts-types de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 100$ et $T = 13$

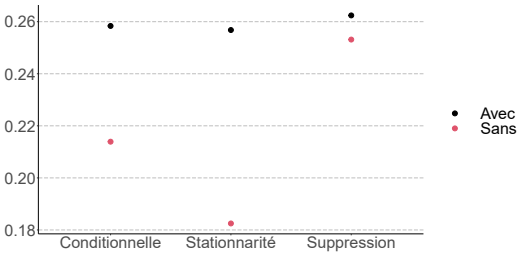
(a) Biais de $\hat{\lambda}_C$ pour $n = 100$ et $T = 13$



(b) Biais de $\hat{\lambda}_E$ pour $n = 100$ et $T = 13$



(c) Écart-type de $\hat{\lambda}_C$ pour $n = 100$ et $T = 13$



(d) Écart-type de $\hat{\lambda}_E$ pour $n = 100$ et $T = 13$

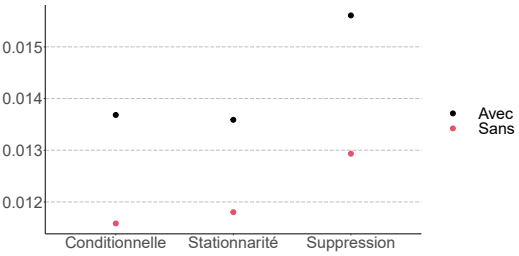
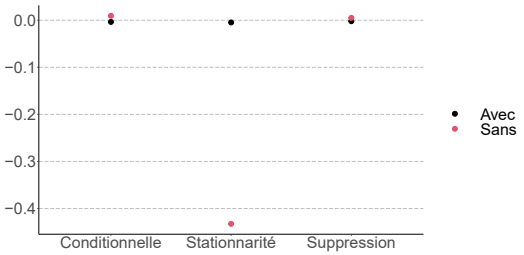


Tableau A.12: Résultats pour $n = 100$ et $T = 26$

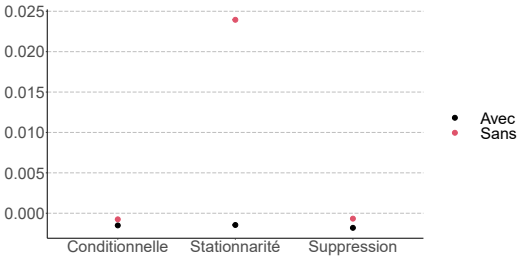
Paramètre	Valeurs moyennes des résultats de simulations répliquées 100 fois	
	Avec hypothèse de stationnarité	Sans hypothèse de stationnarité
$\hat{\lambda}_C$ Suppression	3.998	4.005
$\hat{\lambda}_E$ Suppression	0.198	0.199
$\hat{\lambda}_C$ Conditionnelle	3.996	4.009
$\hat{\lambda}_E$ Conditionnelle	0.198	0.199
$\hat{\lambda}_C$ Stationnaire	3.995	3.567
$\hat{\lambda}_E$ Stationnaire	0.199	0.224
λ_C Théorique	4.000	4.000
λ_E Théorique	0.200	0.200

Figure A.12: Biais et écarts-types de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 100$ et $T = 26$

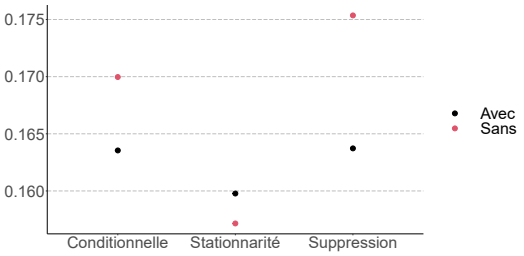
(a) Biais de $\hat{\lambda}_C$ pour $n = 100$ et $T = 26$



(b) Biais de $\hat{\lambda}_E$ pour $n = 100$ et $T = 26$



(c) Écart-type de $\hat{\lambda}_C$ pour $n = 100$ et $T = 26$



(d) Écart-type de $\hat{\lambda}_E$ pour $n = 100$ et $T = 26$

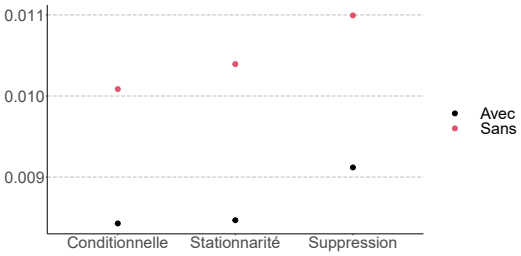
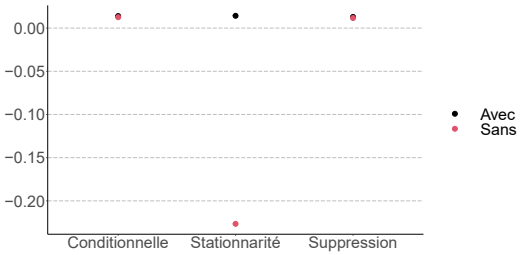


Tableau A.13: Résultats pour $n = 100$ et $T = 52$

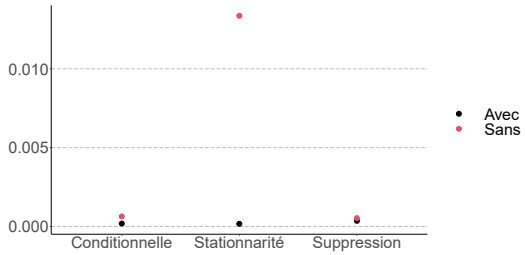
Paramètre	Valeurs moyennes des résultats de simulations répliquées 100 fois	
	Avec hypothèse de stationnarité	Sans hypothèse de stationnarité
$\hat{\lambda}_C$ Suppression	4.013	4.012
$\hat{\lambda}_E$ Suppression	0.200	0.201
$\hat{\lambda}_C$ Conditionnelle	4.014	4.013
$\hat{\lambda}_E$ Conditionnelle	0.200	0.201
$\hat{\lambda}_C$ Stationnaire	4.014	3.773
$\hat{\lambda}_E$ Stationnaire	0.200	0.213
λ_C Théorique	4.000	4.000
λ_E Théorique	0.200	0.200

Figure A.13: Biais et écarts-types de $\hat{\lambda}_C$ et $\hat{\lambda}_E$ pour $n = 100$ et $T = 52$

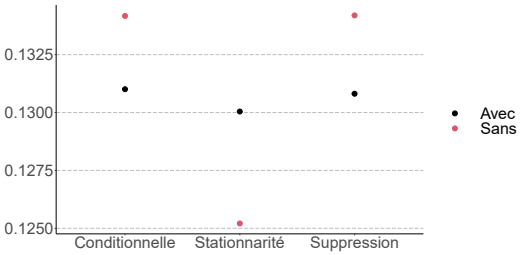
(a) Biais de $\hat{\lambda}_C$ pour $n = 100$ et $T = 52$



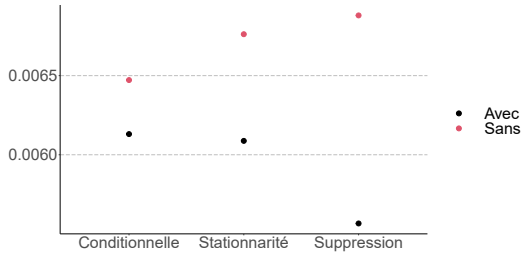
(b) Biais de $\hat{\lambda}_E$ pour $n = 100$ et $T = 52$



(c) Écart-type de $\hat{\lambda}_C$ pour $n = 100$ et $T = 52$



(d) Écart-type de $\hat{\lambda}_E$ pour $n = 100$ et $T = 52$



RÉFÉRENCES

Addona, V. (2005). *Stationarity in a prevalent cohort study with follow-up.* (Thèse de doctorat). McGill University.

Alvarez, E. E. (2006). Maximum likelihood estimation in alternating renewal processes under window censoring. *Stochastic Models*, *22*(1), 55–76.

Arango-Castillo, L., Atherton, J. et Froda, S. (2019). Unemployment durations in fixed panels : a comparison using complex survey data in the presence of missing information. Submitted.

Ardilly, P. (2006). *Les techniques de sondage.* Editions Technip.

Cook, R. J. et Lawless, J. F. (2018). *Multistate models for the analysis of life history data.* CRC Press.

Cox, D. R. et Isham, V. (2000). *Point processes.* Chapman & Hall.

Dufour, J., Gagnon, F., Morin, Y., Renaud, M. et Särndal, C. (2001). Mieux comprendre la transformation des poids à l'aide d'une mesure de changement. *Techniques d'enquête.*

Feenstra, R. C., Inklaar, R. et Timmer, M. P. (2015). The next generation of the Penn World Table. *American Economic Review*, *105*(10), 3150–82.

Grimmett, G. et Stirzaker, D. (2001). *Probability and random processes.* Oxford University Press.

Institute for Social Research (2021). PSID Main Interview User Manual : Release 2021.

Kalbfleisch, J. D. et Prentice, R. L. (2011). *The statistical analysis of failure time data.* John Wiley & Sons.

Kovačević, M. S. et Roberts, G. (2007). Modelling durations of multiple spells from longitudinal survey data. *Survey Methodology*, *33*(1), 13–22.

LaRoche, S. (2007). *Pondérations longitudinale et transversale de l'enquête sur la dynamique du travail et du revenu. Année de référence 2003.* Rapport technique, Statistique Canada, Ottawa.

Lawless, J. (2003). *Statistical models and methods for lifetime data.* Wiley.

Lohr, S. L. (2009). *Sampling : design and analysis.* Cengage Learning.

Naud, J.-F. (2004). *Pondération longitudinale avec panels combinés, Enquête sur la dynamique du travail et du revenu.* Statistique Canada, Direction de la méthodologie, Division des méthodes d'enquêtes sociales.

Ross, S. M. (1996). *Stochastic processes.* Wiley.

Selvin, S. (2008). *Survival analysis for epidemiologic and medical research.* Cambridge University Press.

Statistique Canada. (2003). *Méthodes et pratiques d'enquête*, volume No 12-587-X, ISBN 978-1-100-95206-2. Statistique Canada, Direction de la méthodologie.

Statistique Canada. (2017). *Méthodologie de l'enquête sur la population active du Canada.* Statistique Canada, Direction de la méthodologie.

Statistique Canada. (2018). *Guide de l'enquête sur la population active.* Statistique Canada, Direction de la méthodologie.

Statistique Canada (Mis en ligne le 3 mai 2019, version mise à jour le 17 juin 2019, consulté le 3 août 2019). « le programme des centres de données de recherche (cdr) ». https ://www.statcan.gc.ca/fra/cdr/index.

Taylor, M. F., Brice, J., Buck, N. et Prentice-Lane, E. (1993). *British Household Panel Survey user manual : Volume A : Introduction, technical report and appendices.* University of Essex Colchester.