

SIMUL

SIMUL : MODÈLE DYNAMIQUE EN FORME RÉDUITE

**DOCUMENT TECHNIQUE
TECHNICAL DOCUMENT**

Nicholas-James Clavet, Jean-Yves Duclos et Bernard Fortin

Décembre 2014

La Chaire de recherche Industrielle Alliance sur les enjeux économiques des changements démographiques est une chaire bi-institutionnelle qui s'appuie sur un partenariat avec les organisations suivantes :

- **Centre interuniversitaire de recherche en analyse des organisations (CIRANO)**
 - **Industrielle Alliance, Assurance et services financiers inc.**
 - **Régie des rentes du Québec**

Les opinions et analyses contenues dans ce document ne peuvent en aucun cas être attribuées aux partenaires ni à la Chaire elle-même et elles n'engagent que leurs auteurs.

Opinions and analyses contained in this document cannot be attributed to the Chair or its partners and are the sole responsibility of the authors.



SIMUL : Modèle dynamique en forme réduite

NICHOLAS-JAMES CLAVET
JEAN-YVES DUCLOS
BERNARD FORTIN

CIRPÉE
UNIVERSITÉ LAVAL

DÉCEMBRE 2014

Résumé

Ce document présente les résultats préliminaires d'une première version du Modèle de Micro-Simulation de l'Université Laval (SimUL), portant sur l'évolution de divers indicateurs sociaux et de mesures de niveaux de vie pour la période 2004-2030. Cette évolution est largement tributaire des changements démographiques en cours et à prévoir, tels le vieillissement de la population, l'augmentation de l'espérance de vie, l'augmentation de la proportion d'immigrants dans la population et l'augmentation prévisible du niveau de scolarité. En prenant compte de l'ensemble de ces facteurs, les prévisions de SimUL montrent entre autres que les revenus de travail en termes réels des individus devraient augmenter significativement entre 2004 et 2030, même dans un scénario où la productivité du capital humain est maintenue fixe pour un niveau de scolarité donné. De plus, les résultats prévoient des changements dans la participation au marché du travail, notamment une augmentation du nombre d'heures travaillées à temps plein des femmes. La prolongation de la durée des études limitera par ailleurs la participation au marché du travail des plus jeunes (16-24 ans), mais fera augmenter celle des personnes en âge de prendre leur retraite (50 ans et plus). Une part importante des changements dans les indicateurs socio-économiques est ainsi expliquée par l'augmentation du niveau de scolarité des individus.

Mots clé : Micro-simulation dynamique ; changements démographiques ; revenus ; impacts fiscaux des changements démographiques.

Table des matières

1	Introduction	2
1.1	SimUL	2
1.2	Premier aperçu	3
1.3	Plan du rapport	4
2	Revue des modèles de micro-simulation	4
3	SimUL : un modèle de micro-simulation dynamique	7
3.1	La modularisation	7
3.2	L'estimation des probabilités de transition	8
3.3	La calibration	8
4	Approches de modélisation	9
5	Sources de données pour le modèle SimUL	12
6	Le modèle de simulation de l'Université Laval (SimUL)	15
6.1	L'immigration	15
6.2	L'émigration	16
6.3	La mortalité	16
6.4	Les départs du ménage	17
6.5	La formation et la dissolution de couples	17
6.6	La natalité	18
6.7	L'éducation	19
6.8	Le marché du travail	19
6.9	Les revenus de placements	20
6.10	Les revenus de retraite	21
7	Prévisions de SimUL	22
8	Simulations supplémentaires	43
8.1	Effet de l'augmentation du niveau de scolarité sur le revenu moyen	43
8.2	Effet de l'augmentation de l'âge sur le revenu moyen	44
8.3	Effets de l'immigration	45
9	Conclusion	46
A	Construction des bases de données	51
A.1	Données initiales	51
A.2	Données des immigrants internationaux	58
B	Annexe technique	59
B.1	Ajustement des poids échantillonnaires	59
B.2	Méthode d'appariement	61
B.3	Dérivation du risque de sortie pour le logit	62

C	Les variables explicatives des modèles	63
C.1	Le modèle de départ de ménage	63
C.2	Le modèle de formation de couple	64
C.3	Le modèle de dissolution de couple	65
C.4	Le modèle de naissance	66
C.5	Le modèle de scolarité	66
C.6	Le modèle de marché du travail	67
C.7	Le modèle de RRQ (1 si l'individu en reçoit, 0 sinon)	68
C.8	Le modèle de de prestation de RRQ)	68
C.9	Le modèle de pension privée (1 si l'individu en reçoit, 0 sinon)	69
C.10	Le modèle de niveau de pension privée	70
C.11	Le modèle de différence de pension privée	71
C.12	Le modèle de revenus de placements	71

Résumé du modèle

SIMUL est un modèle de microsimulation dynamique en forme réduite de l'économie québécoise. Il s'appuie à la fois sur un large échantillon représentatif transversal de la population québécoise, sur plusieurs enquêtes rétrospectives et longitudinales et sur des données prospectives sur la composition future de la population québécoise. Le modèle permet de dresser un portrait de l'évolution socio-démographique et socio-économique du Québec, et à terme de l'ensemble du Canada, pour les prochaines décennies.

L'approche par microsimulation, en contraste avec l'approche par macrosimulation, permet d'analyser les actions et les interactions au niveau individuel afin d'obtenir des distributions complètes (courantes et futures) de variables socio-économiques d'intérêt et de groupe d'intérêt (tel que les retraités et les non-retraités, les hommes et les femmes, les personnes en couple et les célibataires). Cette approche permet également d'obtenir des macro-résultats par agrégation.

Chaque individu au sein du modèle SIMUL suit un parcours de vie caractérisé par des événements démographiques, sociaux et économiques tel que la natalité, la mortalité, l'immigration, l'émigration, la formation et la séparation des familles, l'éducation, l'épargne, l'emploi et les revenus de retraite. La fréquence et le type de ces événements dépendent des caractéristiques de l'individu et de celles des autres membres de son ménage. Il est dit également que le modèle est dynamique puisqu'il met à jour l'ensemble de ces caractéristiques individuelles et familiales sur plusieurs périodes. Pour ce faire, les caractéristiques ci-dessus sont modélisées de manière séquentielle à l'intérieur d'une période.

Ainsi, la finalité du modèle se trouve à la fois dans sa nature comptable (des finances privées et publiques) et distributive. Dans cette perspective, SIMUL permet de prévoir les recettes fiscales ainsi que les dépenses gouvernementale en prenant en compte la répartition des variables socio-économiques au sein de la population québécoise. Enfin, SIMUL peut servir à évaluer l'impact d'éléments existants ou prospectifs de l'environnement économique ou encore d'une politique publique sur la répartition courante et future des variables socio-économiques ainsi que sur les finances publiques.

1 Introduction

Ce rapport présente les résultats préliminaires d’une première version du Modèle de Micro-Simulation de l’Université Laval (SimUL). Ces résultats tracent l’évolution de divers indicateurs sociaux et de mesures de niveaux de vie au Québec pour la période 2004-2030. Cela inclut des estimés de l’évolution des revenus à travers les 20 prochaines années. Le modèle permet aussi de saisir l’importance des facteurs qui influenceront l’évolution future des revenus et des indicateurs sociaux. Le modèle utilisé permet de plus de simuler en quoi cette évolution serait modifiée par des changements de la structure socio-démographique de la population ou par des chocs éventuels dans les conditions du marché du travail.

1.1 SimUL

Le Modèle de Micro-Simulation de l’Université Laval (SimUL) que nous utilisons aux fins des simulations a été développé au département d’économique de l’Université Laval par les auteurs du présent rapport. Ce modèle repose sur un large échantillon représentatif d’individus et de ménages. Il est à la fois dynamique (en temps discret) et stochastique. Les variables endogènes prédites se fondent en grande partie sur des probabilités de transition qui dépendent elles-mêmes de variables endogènes prédéterminées (passées) et de variables exogènes. Il s’agit (pour le moment) essentiellement d’un modèle en forme réduite et non d’un modèle structurel fondé sur la modélisation de comportements d’optimisation des individus et des ménages.

Les fondements empiriques de SimUL reposent largement sur l’usage d’enquêtes rétrospectives et longitudinales sur les transitions des ménages de manière à estimer les paramètres de projections de la population québécoise. Des données prospectives sur la composition future de la population (*e.g.*, prévisions démographiques de l’Institut de la Statistique du Québec (ISQ)) sont également employées dans le but d’harmoniser les cohortes avec les prévisions officielles. Une méthodologie propre aux modèles discrets de micro-simulation est utilisée pour obtenir des prévisions sur la population québécoise (Cassells et al. 2006, Zaidi and Rake 2001 et Morrison and Dussault 2000a). Les phénomènes démographiques et économiques principaux (mortalité, immigration, émigration, formation et dissolution des familles, départ du foyer familial, scolarité, revenu d’emploi, participation au marché du travail et épargne) sont modélisés de manière séquentielle à l’intérieur d’une période (une année). Des techniques telles que les régressions linéaires, les régressions logistiques et l’imputation stochastique sont employées pour déterminer le moment d’arrivée d’un événement.

SimUL comprend aussi certains éléments de type macroéconomique qui peuvent permettre de simuler de manière prospective l’évolution différentielle des revenus selon différentes hypothèses sur la croissance de la productivité et des salaires moyens, sur les changements socio-démographiques, sur les taux d’emploi par groupe âge-sexe, sur l’immigration nette, le taux de rendement des actifs, pour n’en nommer que quelques-unes.

La plupart des modèles utilisés à des fins de prévisions reposent sur des *macro-simulations*. Ces modèles prédisent, à partir des tendances observées dans le passé, l’évolution de mesures agrégées, comme le PIB, l’espérance de vie moyenne, les impôts totaux, *etc.* Ils tentent ainsi d’établir des tendances agrégées au sein d’une population.

L’analyse unique de prévisions agrégées ignore cependant l’étude d’un certain nombre

de questions importantes, en particulier celles reliées à l'impact désagréé des changements démographiques sur la population, les revenus et les finances publiques. Les modèles de micro-simulation dynamiques permettent au contraire de modéliser les distributions futures complètes de variables socio-économiques d'intérêt et de les séparer en autant de classes désirées (comme les retraités et les non-retraités, les hommes et les femmes, les personnes en couple et les célibataires). Ces calculs sont essentiels pour étudier des phénomènes telles que la pauvreté et l'inégalité, ces dimensions du bien-être ne pouvant être captées par des informations portant uniquement sur l'individu représentatif. Ils permettent aussi de prévoir correctement les recettes fiscales provenant de l'impôt sur le revenu ou d'autres formes d'impôt ou de transfert non-linéaire. Les modèles de micro-simulation dynamiques peuvent aussi servir à évaluer comment des changements de politiques peuvent influencer les répartitions de revenus et affecter les finances publiques (comme la taille et la distribution des recettes fiscales ainsi que la viabilité des régimes de pension publiques).

1.2 Premier aperçu

Cette première version de SimUL est en mesure d'effectuer des prévisions sur plusieurs variables. Le modèle présente ainsi les contributions de la natalité, la mortalité, l'immigration et l'émigration à la croissance à prévoir de la population jusqu'en 2030. On prévoit que le nombre de décès rattrapera en 2030 le nombre de naissances, la croissance totale de la population demeurant toutefois positive dû à l'immigration. La population du Québec devrait ainsi atteindre 9 millions d'habitants en 2030. On voit aussi comment l'immigration affectera la distribution de l'âge, de l'éducation et des revenus dans la population. La majorité des immigrants arrivent au Québec avant 40 ans, ce qui contribue à faire augmenter le nombre de travailleurs par rapport au nombre de retraités. L'effet n'est cependant pas suffisant pour empêcher le ratio de dépendance d'augmenter significativement jusqu'en 2030, passant de moins de 0,5 en 2010 à plus de 0,7 en 2030.

SimUL prévoit également que l'augmentation continue du niveau d'éducation des individus dans le temps augmentera significativement leur niveau de vie, même si la productivité par niveau de scolarité est maintenue fixe (en supposant par exemple un rapport capital humain/capital physique constant, ainsi qu'une technologie constante). C'est ainsi que l'on prévoit que, de 2010 à 2030, le revenu de travail moyen (incluant les revenus nuls) passera de 34492\$ à 39110\$ en dollars constant de 2004, soit une augmentation d'environ 13%. Les prévisions suggèrent aussi que les individus dans la cinquantaine, qui deviendront de plus en plus éduqués, auront un taux de participation de plus en plus élevé. L'éducation aura cependant comme effet de diminuer le taux d'emploi des jeunes adultes, car ils prolongeront la durée de leurs études. On prévoit aussi que la proportion des femmes qui travailleront à temps plein augmentera significativement, en raison entre autres des effets d'âge et d'éducation qu'amèneront les changements socio-démographiques à prévoir. Globalement, on comptera, selon les prévisions de SimUL, 15 non-travailleurs (enfants, retraités, chômeurs et inactifs) pour chaque 10 travailleurs en 2030 contre 12 non-travailleurs pour 10 travailleurs en 2010.

1.3 Plan du rapport

Dans la section 2, certains modèles de micro-simulation dynamiques à travers le monde sont présentés et des comparaisons entre SimUL et les autres modèles de micro-simulation canadiens sont effectuées. La section 3 décrit le fonctionnement de notre la micro-simulation dynamique. Différentes approches de modélisation sont exposées dans la section 4. La section 5 présente les données utilisées et la section 6 décrit plus précisément le modèle SimUL. La section 7 décrit les prévisions principales.

2 Revue des modèles de micro-simulation

Il est possible de répertorier au moins 27 modèles de micro-simulation dynamique à travers le monde (Spielauer, 2007). Plusieurs pays développés en possèdent au moins un (Australie, Canada, États-Unis, Irlande, Italie, France, Japon, Nouvelle-Zélande, Norvège, Royaume-Unis et Suède). Chaque modèle possède des caractéristiques particulières selon ses objectifs et sa durée d'existence. Ainsi, les modèles de micro-simulation appréhendent plus ou moins précisément les principaux phénomènes démographiques et économiques suivants : immigration, mortalité, fécondité, émigration, formation et dissolution des familles, scolarité, participation au marché du travail, revenu, habitation, autres dépenses, actifs et dettes de la famille, pension de retraite, transferts gouvernementaux, fiscalité, santé et soins des personnes âgées (Harding et al. 2009, Meyerson et al. 2009, Statistique Canada 2009, Smith et al. 2007, Cassells et al. 2006, Flood et al. 2005, Favreault and Smith 2004, Robert-Bobée 2001, Morrison and Dussault 2000a et Fredriksen 1998a). La plupart des modèles, incluant SimUL, reposent sur l'usage d'un large échantillon représentatif d'individus et de ménages et s'inspirent d'un point de vue méthodologique de Harding (1993) et de Gupta and Kapur (2000). Malgré tout, chaque modèle de micro-simulation possède des caractéristiques particulières, qui reflètent les priorités de ses concepteurs.

Un des premiers modèles de micro-simulation à avoir été développé est intitulé DYNASIM. La toute première version fut développée entre 1969 et 1976. Cependant, la plus récente version a été complétée en 2004 (DYNASIM3). Son objectif principal est de traiter de la problématique des pensions de retraite et du vieillissement de la population. Il est possible de le décomposer en trois sous-modèles. Le premier modèle « Family and Earnings History » (FEH) simule année par année les événements au niveau démographique et des comportements d'offre de travail, alors que le modèle « Jobs and Benefits » (JBH) simule personne par personne les emplois, le secteur industriel, la couverture des pensions de retraite, la retraite, les prestations de sécurité sociale et les prestations de pension de retraite privée. Finalement, le dernier sous-modèle, « The Cross-Sectional Imputation Model » (CSIM), impute de manière statique des indicateurs de santé et de richesse. La conceptualisation de DYNASIM, au niveau de la décision de la prise de retraite, des types de régimes de retraite privés et des transferts gouvernementaux reliés à la retraite, est détaillée. Ceci lui permet de faire des prévisions précises sur l'évolution des revenus de retraite.

APPSIM (Harding et al., 2009) est un modèle de microsimulation du *National Centre for Social and Economic Modelling* (NATSEM). Il a pour but de modéliser des modifications aux politiques fiscales et d'analyser les effets du vieillissement de la population sur les

revenus du gouvernement et les dépenses en santé. Pour ce faire, ce modèle peut compter sur onze composantes : Démographie, Formation de ménage et transition, Éducation et formation, Main d'oeuvre, Revenus, Habitation, Autres revenus et dépenses, Actifs financiers (dette), Sécurité sociale, Fiscalité, Santé et Soins des personnes âgées. Une particularité de ce modèle vient de sa méthode d'imputation des immigrants, car les concepteurs ont porté une attention particulière aux types de visa d'immigration (famille, travailleur spécialisé, humanitaire). DYNAMOD (Kelly, 2003) est pour sa part un modèle de microsimulation plus âgé du NATSEM. Ce modèle se démarque des autres par ses fonctions de survie et sa précision mensuelle. Comme le modèle évalue à chaque mois certaines variables, les fonctions de survie accélèrent la simulation en calculant la date des transitions démographiques seulement quelques fois durant la vie d'une personne.

Le modèle MOSART (Fredriksen, 1998b) a été développé par Statistiques Norvège et il a pour but d'évaluer les effets des réformes fiscales et du vieillissement de la population. La base de données initiale de ce modèle provient de 12 échantillons de 1 % du recensement de Norvège. Cependant, MOSART se démarque de la plupart des autres modèles grâce à la fusion de cette première source d'information avec des données administratives longitudinales qui renferment un foule d'information sur les citoyens norvégiens. Ce modèle possède des modules lui permettant de simuler la migration, la mortalité, la fécondité, la formation et la séparation de couple, la scolarité, la participation au marché du travail et la sécurité sociale. Ainsi, une des applications pratiques du modèle MOSART est d'effectuer des prévisions sur le taux de contribution d'équilibre au régime de pension publique norvégien.

SESIM (Flood et al., 2005) est un modèle de microsimulation du Ministère des Finances de Suède. Comme la Norvège, ce pays possède des données administratives longitudinales sur lequel le modèle est basé. L'objectif initial de SESIM était l'évaluation du financement du système d'éducation suédois. Par la suite, le modèle a été adapté à l'étude de questions reliées au vieillissement de la population et aux revenus de retraite. Ainsi, ce modèle est tout aussi capable d'estimer la valeur des subventions des services publics que d'évaluer les prestations de retraite des émigrants suédois. Le modèle possède également une composante spciale pour l'habitation. À l'aide de celle-ci, on détermine si un enfant de parents âgés malades réside suffisamment près de ceux-ci pour être un fournisseur potentiel de soins à domicile. Finalement, SESIM comprend une modélisation détaillée de la décision d'offre de travail, de la prise de retraite et de l'accumulation de richesse. Dans le cas du Canada, il n'existe à notre connaissance qu'un seul modèle économique de micro-simulation dynamique en opération. Néanmoins, il a été possible de répertorier un autre modèle canadien dont le projet a été abandonné. Le modèle encore en usage est intitulé LifePaths et il provient de Statistique Canada.¹ Celui qui a été délaissé se nomme DYNACAN et il avait été développé par Ressources humaines et Développement des compétences Canada (RHDCC) (Morrison, 2007).

LifePaths est construit sur la base de *personnes dominantes* tirées du recensement et leur recrée un historique synthétique. Les informations sur le conjoint ou les enfants d'une personne dominante sont déterminées par les équations du modèle. Le parcours de vie des personnes dominantes et synthétiques est aussi simulé à l'aide des équations du modèle. Celles-ci modélisent les phénomènes socio-économiques suivants : migration, mortalité, sco-

1. Voir Statistique Canada (2009) pour une description complète du modèle LifePaths.

larité, fécondité, formation et séparation de ménage, départ du foyer familial, institutionnalisation des personnes âgées, participation au marché du travail, congé de maternité et revenu d'emploi. Lifepaths sert aussi de point de départ pour des modèles plus spécialisés de Statistiques Canada.

SimUL utilise plutôt un échantillon représentatif de familles pour lequel nous disposons d'informations sur tous les membres de la famille. SimUL utilise ensuite les équations du modèle pour faire évoluer ces individus et ces familles à travers le temps. Une deuxième différence majeure entre LifePaths et SimUL est que LifePaths utilise des fonctions de risque en temps continu, alors que SimUL se déroule en temps discret. L'usage d'une modélisation en temps continu permet d'éviter certains problèmes de simulation, mais est plus exigeant au niveau des données. Aussi, le fait d'utiliser des données synthétiques dans LifePaths permet une plus grande variabilité dans les données. Celle-ci est toutefois entièrement tributaire du modèle spécifié. LifePaths se doit ainsi d'être bien spécifié pour reproduire la variabilité à laquelle on s'attend dans la population. Dans SIMUL, la base de données initiale est tirée d'un échantillon de ménages. La représentativité de cet échantillon, tiré d'observations réelles, assure statistiquement la fiabilité de la population de départ. La précision statistique de cet échantillon dépend toutefois du nombre d'observations disponibles au départ ainsi que de la validité du modèle utilisé pour faire « vieillir » la population. DYNACAN est une refonte par RHDCC du modèle CORSIM, qui était originellement conçu pour les États-Unis. Ce modèle est de type populationnel dans le sens qu'on considère que les observations du modèle sont représentatives de la totalité de la population ((Morrison and Dussault, 2000b)). Ainsi, les interactions entre les individus sont modélisées à l'aide des observations déjà présentes dans le modèle. À ce niveau, SimUL est semblable à DYNACAN. De plus, DYNACAN utilise, tout comme SimUL, une base de données représentative de la population d'intérêt. Celle-ci est le Recensement de 1971. L'objectif de DYNACAN était d'évaluer les effets de modifications potentielles au régime canadien de pension de retraite (RCP). Ainsi, il modélise autant les cotisations que les prestations du RCP. Comme le Québec gère lui-même son régime de pension de retraite, une attention particulière est apportée à la migration interprovinciale. De plus, le modèle possède des algorithmes permettant de réconcilier les données observées aux données prédites comme source de validation. Des efforts ont été faits pour prendre en compte l'immigration illégale et le retour des émigrants. Une deuxième méthode de validation est la calibration de DYNACAN avec un modèle actuariel traditionnel (semi-agrégé) nommé ACTUCAN.

LifePaths et DYNACAN ne semblent pas tenir compte de l'ensemble des spécificités de la fiscalité du Québec. En effet, pour Lifepaths, seul les paramètres du régime fiscal fédéral figurent dans le calcul des montants d'impôt à payer. Or, le Québec possède une assiette fiscale et une table d'imposition qui lui sont propres. Certaines autres particularités du Québec telles que les services de garde et les montants d'aide sociale (ou d'assistance-emploi) ne semblent pas non plus avoir été prises en compte dans LifePaths et DYNACAN. SimUL comble certaines de ces lacunes au niveau de la fiscalité et de certains transferts au Québec. À plus long terme, SimUL pourra aussi mieux prévoir l'effet sur les individus de changements dans la fiscalité et les transferts gouvernementaux par l'introduction d'une modélisation structurelle de certains comportements socio-économiques.

3 SimUL : un modèle de micro-simulation dynamique

SimUL est un modèle de micro-simulation dynamique. Le modèle prend entre autres compte de la complexité de la fiscalité et des transferts sociaux pour simuler l'évolution des revenus sur un large échantillon représentatif d'individus et de ménages. Contrairement aux modèles de macro-simulation habituellement utilisés à des fins de prévisions, SimUL modélise non seulement l'évolution des agrégats, mais aussi celle de chaque individu dans la base de données utilisée. On obtient ainsi des prévisions des distributions complètes des variables d'intérêt et des estimés plus fiables des agrégats.

La logique de SimUL et de plusieurs autres modèles de micro-simulation est simple. Les étapes (simplifiées) d'une simulation typique sont les suivantes :

1. On a d'abord recours à une base de données représentative de la population d'intérêt. À partir de cette base ou d'autres enquêtes, on calcule les probabilités que certains événements surviennent (mariage, décès, obtention d'un diplôme, *etc.*) en fonction des caractéristiques des individus.
2. On fait ensuite vieillir les individus d'une période. On simule les changements dans leurs caractéristiques socio-économiques ainsi que les événements qui les affectent selon les probabilités calculées à l'étape précédente.
3. À l'aide des nouvelles caractéristiques des individus et de leurs familles, on répète l'étape précédente jusqu'à ce qu'on ait simulé le nombre de périodes désiré.

Ainsi, l'exercice de simulation met à jour les caractéristiques individuelles à chaque période, traçant la distribution des variables d'intérêt à travers le temps. Si la logique conceptuelle de la micro-simulation est simple, l'exercice pratique de programmation rigoureuse et efficace peut être ardu et complexe. Les paragraphes suivants décrivent les étapes de développement du modèle de manière plus détaillée.

3.1 La modularisation

La micro-simulation est en mesure de prendre en compte un nombre important de variables interagissant les unes avec les autres de manière potentiellement complexe. La littérature sur les micro-simulations porte ainsi une attention particulière à la modularisation, une méthode qui s'avère efficace pour accroître la compréhension, la flexibilité et la rapidité des simulations.²

L'idée générale est de programmer chaque événement séparément dans un « module ». ³ Le modèle est ainsi plus flexible : si on veut modifier certains paramètres associés à un événement (*e.g.*, les probabilités en fonction des caractéristiques, les variables exogènes, *etc.*), on modifie un seul module plutôt que le modèle au complet. Un autre avantage est que, si certaines catégories peuvent être simulées indépendamment des autres, on peut s'attarder à seulement certaines catégories lors d'une simulation. Par exemple, un chercheur qui

2. Voir Cassells et al. (2006) pour une discussion plus approfondie sur l'importance de bien structurer son modèle.

3. Il peut arriver que l'on programme deux événements dans un même module. Par exemple, le décès d'une mère monoparentale entraîne l'adoption de l'orphelin dans une autre famille. Ce lien de cause à effet rend donc simple la programmation conjointe de la mortalité et de l'adoption.

s'intéresse uniquement aux conséquences d'une politique sur les variables démographiques peut simuler le modèle en utilisant uniquement cette catégorie de modules. Il évite ainsi de s'encombrer de ce qui ne lui est pas nécessaire et la simulation en est plus rapide. Chaque module de SimUL est présenté plus en détails plus loin dans le texte.

3.2 L'estimation des probabilités de transition

Chaque module a comme fonction de simuler un événement selon les probabilités qu'il survienne. Ces probabilités sont fonction des caractéristiques de l'individu et, s'il y a lieu, de certaines variables exogènes (*e.g.* l'âge d'accès à la rente de retraite, le taux d'imposition sur le revenu, *etc*). On doit généralement utiliser plusieurs bases de données représentatives de la population d'intérêt pour estimer ces probabilités.

Dans le cas de SimUL, la base de données initiale (BDMSPS) est représentative de la population du Québec, mais contient relativement peu d'informations permettant d'estimer les probabilités des événements. C'est pourquoi nous utilisons l'Enquête Sociale Générale (ESG) et l'Enquête Dynamique sur le Travail et le Revenu (EDTR) pour estimer la plupart des probabilités nécessaires à la microsimulation. Comme l'ESG contient des informations rétrospectives sur les individus, il est possible d'estimer dans cette enquête des fonctions de risque en temps discret à l'aide d'un logit.⁴ Malheureusement dans l'EDTR, la durée des panels n'est pas suffisante pour l'utilisation d'une telle technique. Ainsi, pour les événements estimés à partir de cette dernière enquête, on procède à l'estimation des proportions de personnes dans un état précis. Ceci est encore une fois évalué à l'aide d'un logit, mais avec une structure de données différente.⁵ Les paramètres de ces estimation sont ensuite utilisés pour calculer les probabilités de transition de chaque individu dans l'échantillon initial.

Notons que certains modules n'utilisent pas de modèle économétrique pour estimer les probabilités *agrégées* de transition car des prévisions exogènes sont disponibles. Cela est le cas pour la mortalité, l'immigration et l'émigration, qui sont, pour l'instant, dans SimUL entièrement déterminées d'un point de vue agrégé (mais pas individuellement) par les prévisions de l'ISQ. Les naissances, comme nous le verrons plus loin, sont déterminées par une combinaison de probabilités estimées et de prévisions exogènes.

3.3 La calibration

Comme les individus dans la simulation sont soumis à une probabilité d'être affectés par un certain événement, on ne peut savoir *a priori* le nombre d'individus qui seront affectés par cet événement dans la simulation. Le nombre d'individus affectés par un événement pour une période donnée suit une distribution précise mais sa réalisation reste incertaine, laissant à « l'aléa de Monte Carlo » le soin de fixer ce nombre. Il est alors peu probable que le nombre d'événements d'un phénomène socio-économique corresponde à son espérance mathématique pour chaque période, surtout si on simule un grand nombre de périodes. Cette section décrit les techniques de calibration qui permettent d'éliminer l'incertitude

4. Voir en annexe B.3 pour une dérivation de la vraisemblance.

5. Dans le cas d'un modèle logit binaire, la probabilité de l'individu i d'être dans un état précis ($y_i = 1$) en fonction de ses caractéristiques $\mathbf{x}_i$ et des coefficients estimés $\hat{\beta}$ se calcule selon $P(y_i = 1|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i\hat{\beta}}}{1+e^{\mathbf{x}_i\hat{\beta}}}$.

quant à la réalisation d'une variable.⁶

La technique à suivre est de calibrer le nombre exact d'événements pour une période donnée.⁷ On peut premièrement utiliser une calibration à partir de prévisions agrégées exogènes au modèle. Par exemple, supposons qu'on ait une population de 1000 femmes de 20 ans et que 100 d'entre elles soient dans notre échantillon. Si les prédictions indiquent que, dans 5 ans, 50 femmes de 25 ans donneront naissance à un enfant (c'est-à-dire 5% de la population), on impose alors dans notre simulation qu'il y aura exactement 5 femmes de l'échantillon ($5\% \times 100 = 5$) qui accoucheront à cette période. On sélectionne ensuite les femmes qui accoucheront en utilisant les probabilités propres à chaque femme calculées plus haut.

Dans le cas où on n'a pas accès à des prévisions exogènes, on peut tout de même calibrer le nombre d'événements de façon endogène au modèle. Toujours en considérant l'exemple précédent, on considère d'abord les probabilités des 100 femmes de notre échantillon de donner naissance à un enfant. À partir de ces probabilités, on peut calculer l'espérance du nombre de naissances pour ce groupe de femmes. Cette méthode de calibration impose que le nombre de naissances à chaque période corresponde à l'espérance mathématique du nombre d'événements, conditionnellement aux caractéristiques des individus. La calibration peut ainsi être considérée comme une méthode permettant d'imposer la loi des grands nombres quand un sous-groupe de la population concerné est trop petit pour que celle-ci s'applique.

En calibrant ainsi chacun (ou la plupart) des événements, les résultats d'une simulation sont plus près de l'espérance mathématique des variables. Ceci est utile lorsqu'on veut comparer les résultats agrégés d'une simulation à d'autres sources de données (observées ou provenant d'autres types de modèle) sans avoir à effectuer un grand nombre de simulations différentes.⁸ Il en suit que le seul processus aléatoire dans la simulation est la sélection des individus qui sont affectés par ces événements.

4 Approches de modélisation

Développer un modèle de micro-simulation nécessite d'abord de choisir parmi une multitude de méthodes. Cette section présente les éléments à considérer, en plus des avantages et inconvénients des différentes approches de modélisation. Un premier point est la décision d'ouvrir ou de fermer un modèle. Celui-ci sera considéré fermé si on contraint les observations impliquées dans la formation de ménage et l'immigration à provenir uniquement de la base de données principale. Un modèle sera ouvert si de nouvelles personnes sont intégrées dans le modèle lors de ces événements. L'avantage de ce dernier type de modèle est qu'il est plus facile de traiter les observations. Par exemple, il n'est pas nécessaire de faire d'appariement entre elles lors du processus de formation de couple. Dans le cas de l'immigration, on n'a qu'à piger les immigrants dans une banque de données externe ou

6. Un autre raison d'utiliser la calibration est qu'il est difficile d'estimer des probabilités de transition purgées des effets de cohorte et de temps. Voir Kelly and Percival (2009) pour une plus grande discussion à ce sujet.

7. Une autre technique aurait été de calibrer les paramètres des équations de transition plutôt que de calibrer le nombre d'événements.

8. Par simulation différente on réfère au fait d'effectuer une simulation en changeant seulement de point de départ des algorithmes aléatoires.

synthétique. Le défaut majeur de cette approche est néanmoins que des incohérences surviennent entre les observations initiales et les nouvelles. En effet, l'historique des premières est le fruit du modèle dynamique alors que les observations externes sont le fruit d'une imputation statique (Imhoff and Post, 1997).

Nous avons choisi de faire de SimUL un modèle semi-fermé : les couples se forment entre deux individus déjà présents dans la base de données principale et les immigrants sont sélectionnés dans une base de données créée à partir du recensement de 2001. Malgré les défauts de l'ouverture d'un modèle, il a été décidé d'ouvrir SimUL au niveau de l'immigration, car la variabilité des immigrants dans la base de données principale a été jugée insuffisante. De plus, l'évolution des caractéristiques des immigrants internationaux est difficile à anticiper et il est improbable qu'elle corresponde à l'évolution des immigrants déjà présents dans le modèle. Ainsi, la banque de données des immigrants provient d'un échantillon de personnes qui ont immigré au Canada durant les 5 années précédant le Recensement de 2001.

Un deuxième élément à considérer est le choix entre le temps continu et le temps discret. Lorsque le temps est modélisé de manière discrète, on dispose d'informations sur un individu après chaque période sans connaître l'ordre des événements qui s'y sont passés. En temps continu, on a une connaissance plus précise du moment des événements, ce qui permet de mieux gérer les risques compétitifs (Imhoff and Post, 1997).⁹ Toutefois, la modélisation en temps discret est habituellement adoptée, puisque le temps continu nécessite une fréquence de cueillette de données plus grande que celle généralement disponible.¹⁰ Comme les données utilisées pour SimUL sont annuelles, nous adoptons une modélisation à temps discret.

Les risques compétitifs et les événements multiples sont une autre considération d'importance qui découle directement du choix du temps discret. Comme il a été discuté plus haut, le temps discret ne permet pas de savoir ce qui se passe à l'intérieur d'une période. Cependant, l'ordre des événements aura un impact sur les agrégats de la simulation. Il existe trois manières de traiter de cette question (Imhoff and Post, 1997). La première est d'estimer les probabilités composées en spécifiant l'ordre dans lequel les événements se produisent. Une autre solution est de rendre aléatoire l'ordre des événements. L'événement ayant le plus de probabilité de survenir aura plus de chance de se réaliser avant l'autre. La dernière est de déterminer dans une première étape s'il y a changement d'état ou non et si c'est le cas, on détermine quel événement se produit en premier. Généralement, la première méthode est utilisée pour tenir compte des risques compétitifs. C'est pourquoi, les modèles sont habituellement divisés en modules pour chaque bloc d'événements simulés : démographie, éducation, marché du travail, épargne, dépenses du ménage, impôts et etc.), ceux-ci suivant un ordre arbitraire. C'est effectivement le cas pour SimUL ; l'ordre des événements sera présenté plus loin.

Un autre sujet délicat est le traitement des événements impliquant plusieurs individus. Un premier exemple est la séparation d'un couple. Il existe quatre méthodes générales pour modéliser un tel événement (Imhoff and Post, 1997). La première est de considérer que

9. Par exemple, en temps continu, lorsque deux événements (naissance et/ou mort de la mère) sont prévus se produire durant une même période, il est possible de distinguer si une femme enceinte donnera naissance à son enfant avant de mourir ou si elle décèdera en premier.

10. En temps continu, les données se doivent d'être recueillies à chaque mois plutôt qu'à chaque année, par exemple.

la femme prend la décision de la séparation. Ainsi, seules les caractéristiques de la femme seront prises en compte dans l'équation évaluant les probabilités de séparation d'un couple. Une seconde voie possible est d'attribuer à l'homme la décision de séparation, et seules ses caractéristiques sont alors utilisées dans ce cas. Une troisième approche est d'inclure dans l'équation de séparation les caractéristiques des deux partenaires. La dernière méthode est d'employer une équation de séparation pour chacun des sexes et d'obtenir la probabilité espérée de séparation en prenant la moyenne des probabilités des deux équations. C'est cette dernière méthode qui est retenue pour SimUL.

Le deuxième événement d'importance entre deux personnes est la formation d'un couple. Une première approche est de créer de toutes pièces un partenaire à une personne. Cependant, lorsque le modèle est fermé, il est nécessaire de faire l'appariement entre deux observations de la base de données. La première des deux méthodes possibles est l'approche déterministe. On choisit d'abord une ou plusieurs caractéristiques qui détermineront comment se ressemblent les deux partenaires potentiels (par exemple la différence d'âge, d'éducation, etc.) On choisit ensuite une fonction des variables choisies, par exemple la différence d'âge au carré, que l'on voudra minimiser. Finalement, pour chaque homme (ou chaque femme), on trouve la femme (ou l'homme) qui minimise cette fonction. La deuxième méthode est l'approche probabiliste. Comme dans le cas précédent, on choisit une fonction des variables choisies. La différence est que l'on génère ensuite des probabilités d'union entre les partenaires potentiels en fonction de cette fonction. On sélectionne finalement le couple selon les probabilités calculées. (Bouffard et al., 2001) comparent les deux méthodes en vérifiant comment les résultats des modèles CORSIM et DYNANCAN se comparent aux données du recensement. Ils trouvent que l'approche déterministe entraîne une trop forte proportion « d'unions parfaites » en termes de différences d'âge et de revenus, alors que l'approche probabiliste reproduit mieux l'ensemble de la distribution d'intérêt. C'est pourquoi SimUL modélise la formation de couples par l'approche probabiliste. La méthode sera décrite plus précisément plus loin.

Une cinquième considération est le niveau de conceptualisation et d'intégration des modèles. Un conseil généralement donné dans la littérature est de garder son modèle de microsimulation simple mais flexible (Zaidi and Rake, 2001). Cassells et al. (2006) mentionnent à plusieurs reprises qu'il est important de développer un modèle de microsimulation en partant du plus simple vers le plus complexe, tout ceci devant être fait en gardant bien en tête les objectifs du projet. Trois principales voies de complexification peuvent être empruntées. La première est d'accroître les phénomènes modélisés (état de santé, mobilité régionale, pension de retraite privée). Une deuxième est d'améliorer la conceptualisation des événements en transformant des équations réduites en équations à raisonnements logiques (comportement de maximisation). La dernière méthode est de développer un modèle de prévision macro tenant compte de la structure du modèle micro. Effectivement, des liens micro-macro sont régulièrement créés de manière à limiter les sources de variances (échantillonnale et aléatoire) du modèle micro. SimUL est encore au début de son développement et est donc encore un modèle de forme réduite relativement simple. Nous envisageons élaborer les aspects décrits ci-dessus dans le futur.

Le dernier enjeu de taille est celui du type de données sur lesquelles baser les simulations. Les deux principales qualités recherchées dans une base de données sont la représentativité et la quantité d'information. Malheureusement, ces caractéristiques ne sont

pas complémentaires et il est nécessaire de faire un compromis entre elles. Habituellement, la représentativité est primée puisque les biais à ce niveau ne sont qu’accentués avec les simulations (Cassells et al., 2006). C’est pourquoi les recensements servent régulièrement de base de données initiales, malgré leur plus faible quantité d’information. De leur côté, les données d’enquêtes ont une plus faible représentativité, mais elles possèdent de plus grandes quantités d’information sur des sujets précis. De plus, certaines d’entre elles sont dynamiques et permettent de mieux analyser les comportements sur lesquels porte l’enquête. Les données d’enquête sont donc à privilégier si la modélisation détaillée des comportements est plus importante que les seules prévisions de type démographique. Deux autres sources de données plus marginales pouvant être utilisées sont les données administratives et les données synthétiques. Malgré leur richesse d’information, les données administratives ont une utilisation restreinte, puisqu’elles s’appliquent généralement à des sous-groupes précis de la population. Pour leur part, les données synthétiques peuvent être utilisées lorsque qu’aucunes données appropriées n’existent. Toutefois, elles sont peu reconnues auprès du public et des décideurs politiques (Cassells et al., 2006). SimUL utilise une combinaison de données représentatives et d’enquêtes dynamiques. La section suivante décrit ces données plus précisément.

5 Sources de données pour le modèle SimUL

La première source de données utilisée est la Base de Données de Micro-Simulation de Politiques Sociales (BDMSPS) de Statistique Canada. Cette base de données est spécialement conçue pour effectuer des simulations comptables des régimes fiscaux au Canada. La BDMSPS est une combinaison de quatre sources de données : l’Enquête sur la dynamique du travail et du revenu (EDTR) de Statistique Canada, un échantillon de données tirées des déclarations d’impôt des particuliers (connu sous le nom de Livre vert) de l’Agence de revenu du Canada, des données historiques de demandes de prestations d’assurance-emploi et, finalement, l’Enquête sur les dépenses des ménages (EDM) de Statistique Canada. De plus, les données de la BDMSPS sont corrigées pour assurer que les agrégats et les distributions des variables sont représentatifs de la population canadienne, les corrections étant principalement effectuées à l’aide du recensement. Les données de la BDMSPS pour le Québec servent d’échantillon initial dans SimUL. En d’autres mots, nous les utilisons comme échantillon représentatif de la population québécoise pour la première période du modèle (2004).

La modélisation de phénomènes aussi variés que ceux modélisés dans SimUL nécessite une quantité importante d’informations qui ne se retrouvent pas dans la BDMSPS. Le problème majeur est de pouvoir simuler la variabilité temporelle des états et des comportements des individus et des familles. Pour cela, il faut disposer de données panel ou encore de données prospectives. Les données panel sont à privilégier puisqu’elles apportent une perspective dynamique aux données en décrivant l’évolution de mêmes individus au cours de plusieurs années. Des enquêtes rétrospectives et longitudinales sur les transitions des ménages sont donc utilisées pour estimer les paramètres permettant d’élaborer des projections de la population québécoise dans l’avenir. Des données prospectives sur la composition future de la population (ISQ, 2009) sont également employées dans le but de réconcilier les

agrégats des cohortes avec ceux des prévisions officielles.

L'Enquête sociale générale (ESG) sur la famille (années 2001 et 2006) et l'Enquête sur la dynamique du travail et du revenu (EDTR) sont utilisées dans l'estimation des phénomènes socio-économiques. Ces enquêtes permettent d'estimer les probabilités que les variables de notre base initiale changent et dans quelle mesure elles évoluent pour chaque période jusqu'à la fin de la période de simulation. Les cycles 15 et 20 de l'ESG fournissent des informations rétrospectives sur les comportements de fécondité,¹¹ de formation des ménages¹² et de scolarité¹³ permettant de dresser un profil historique des individus à cet égard. Comme le thème de l'ESG change à chaque année, mais qu'il est récurrent aux environs de 5 ans, on retrouve des informations sur les comportements de fécondité, de relation conjugale et de scolarité seulement dans les cycles 10, 15 et 20. Le cycle 10 n'a pas été retenu, puisqu'il est le cycle de l'enquête avec les données les plus anciennes et que présentement nous considérons que le nombre d'observations disponibles dans les cycles 15 et 20 suffit. Quant à l'EDTR, elle contient des données de panel sur des variables économiques, comme les revenus de travail, de placements, de pension privée, du Régime de Rentes du Québec (RRQ), *etc.* Pour disposer des données panel de l'EDTR, il est nécessaire d'avoir accès aux données restreintes de celle-ci. En effet, les données publiques de l'EDTR ne présentent que des coupes transversales. C'est ainsi que nous avons fait appel au CIQSS-Laval pour obtenir l'accès restreint aux deux bases de données (ESG et EDTR).¹⁴ Le CIQSS-Laval constitue l'antenne à l'Université Laval du Centre interuniversitaire québécois de statistiques sociales (CIQSS). Le CIQSS offre à l'ensemble des chercheurs du Québec l'accès aux données détaillées des enquêtes de Statistique Canada et de l'Institut de la statistique du Québec.

Nous présentons ici les statistiques descriptives concernant les diverses sources de revenus possibles pour différentes classes d'âges. Le tableau 1 présente premièrement les revenus initiaux (de la BDMSPP) pour l'année 2004 alors que le tableau 2 présente les mêmes statistiques à partir des données de l'EDTR.

Les revenus de travail sont les revenus les plus importants pour la plupart des classes d'âges. Ils atteignent un sommet pour les individus de 25 à 54 ans, car ceux-ci sont davantage en âge de travailler. Les revenus de placements augmentent progressivement avec l'âge. On constate que, bien qu'il ne s'agisse pas officiellement de revenus de retraite, il s'agit d'une source de revenus non-négligeables pour les personnes âgées. Les revenus de pensions privés deviennent une des sources de revenus principales à partir de 55 ans. Notons que, à partir de 65 ans, les données ne nous permettent pas de distinguer entre les revenus de pensions privées et les retraits de REER. Avant 65 ans, les retraits de REER sont mesurés séparément dans les bases de données, et on ne peut savoir s'ils sont utilisés comme des revenus de retraite ou à d'autres fins. Les retraits de REER avant 65 ans sont toutefois peu importants. Il est possible de recevoir des prestations de la Régie des Rentes du Québec

11. Âge du répondant à la naissance de chaque enfant.

12. Nombre d'union, le type d'union (mariage ou union libre.) et âge du répondant au commencement et à la fin de chaque union

13. Âge de graduation du secondaire, âge de début et de fin d'études pour l'obtention d'un diplôme post-secondaire et âge de fin des études.

14. Même si l'ESG est accessible au niveau publique, certaines variables supplémentaires sont disponibles au niveau restreint.

TABLE 1 – Statistiques descriptives - BDMSPS 2004

Revenu	16 - 24 ans		25 - 54 ans		55 - 64 ans		65 ans et plus	
	Moyenne	Écart-type	Moyenne	Écart-type	Moyenne	Écart-type	Moyenne	Écart-type
Travail	9680	11097	29950	46996	14602	37072	1328	17490
Placements	34	317	1093	23859	1780	17466	2436	8500
Pension	0	0	136	2659	6598	13178	6451	11143
REER	13	188	265	2361	584	2849	—*	—
RRQ	0	0	0	0	1801	3059	4298	2806
SV & SRG	0	0	0	0	0	0	7091	2239
	Obs	1586	Obs	5546	Obs	1523	Obs	1862
	Poids	87536	Poids	3387739	Poids	876997	Poids	1013056

Les revenus sont arrondis au dollar près.

* Les retraits de REER après 65 ans sont considérés comme des revenus de pension.

TABLE 2 – Statistiques descriptives - EDTR

Revenu	16 - 24 ans		25 - 54 ans		55 - 64 ans		65 ans et plus	
	Moyenne	Écart-type	Moyenne	Écart-type	Moyenne	Écart-type	Moyenne	Écart-type
Travail	8181	10998	24627	26543	12279	22348	639	5248
Placements	78	826	790	7364	2424	12989	2463	7254
Pension	0	0	133	1866	5379	15224	6907	15224
REER	5	91	246	2985	503	2633	—*	—
RRQ	0	0	0	0	1967	11498	4668	3090
SV & SRG	0	0	0	0	0	0	6414	2359
	Obs	3826	Obs	14992	Obs	4003	Obs	4792
	Poids	1877555	Poids	8058562	Poids	213036	Poids	2151270

Les revenus sont arrondis au dollar près.

* Les retraits de REER après 65 ans sont considérés comme des revenus de pension.

(RRQ) à partir de 60 ans. Les revenus moyens provenant de la RRQ augmentent avec l'âge en raison de l'augmentation avec l'âge de la proportion de retraités. Les prestations de sécurité de la vieillesse (SV) et de supplément de revenu garanti (SRG) sont versées à tous les individus de 65 ans ou plus selon leur niveau de revenu. La grande majorité des individus reçoivent la pleine prestation de sécurité de la vieillesse (environ 5600\$ en 2004) et les moins riches reçoivent également une prestation de supplément de revenu garanti.

On remarque que les statistiques pour certaines variables diffèrent légèrement entre les deux bases de données. Une cause probable est que les données de l'EDTR sont observées seulement pour les individus ayant accepté de répondre à l'enquête. On peut penser que les individus à revenus élevés, ou encore ceux à revenus particulièrement faibles, ont un taux de réponse plus faible pour ce genre de questions. La BDMSPS utilise les données de l'EDTR pour modéliser la plupart de ses revenus. Cependant, les revenus plus élevés sont imputés à partir des données du Livre vert de l'Agence de revenu du Canada. Il est donc possible que les extrémités des distributions de revenus soient mal représentées dans l'EDTR. Ceci pourrait être potentiellement problématique car nous utilisons les tendances observées dans l'EDTR pour nos prévisions. Cependant, et comme il le sera mentionné plus loin, nous tenons compte de certains biais de sélection potentiels dans ces estimations de revenus.

6 Le modèle de simulation de l'Université Laval (SimUL)

SimUL, comme la majorité des modèles de micro-simulation dynamique, est un modèle à temps discret¹⁵. Ceci implique que l'on doit choisir plus ou moins arbitrairement l'ordre dans lequel les individus passent à travers les modules à l'intérieur d'une période.

La Figure 1 présente l'ordre choisi dans SimUL. Une fois que les individus ont traversé ces modules, les données sont sauvegardées, les individus vieillissent d'un an, et une nouvelle simulation est lancée. Des données sont ainsi obtenues pour chaque période. Cette section présente brièvement chaque module de SimUL dans l'ordre indiqué dans la Figure.

6.1 L'immigration

Nous construisons une base de données d'immigrants internationaux potentiels à partir de laquelle nous sélectionnons les ménages qui immigreront à chaque période. Cette base est construite à partir du Recensement de 2001. Elle contient uniquement les ménages qui ont immigré dans les 5 années qui ont précédé le recensement. Comme nous voulons modéliser les périodes futures, il est davantage approprié d'utiliser des données sur les immigrants les plus récents, puisque les futurs immigrants risquent de leur ressembler davantage.

Pour chaque période, nous utilisons d'abord le nombre d'immigrants prévus sur la base de cibles d'immigration exogènes au modèle qui proviennent de l'ISQ, *i.e.*, les cibles prévues pour l'immigration internationale au Québec pour chaque année jusqu'à 2031. (Notons que

15. L'approche alternative est la modélisation en temps continu. Celle-ci a l'avantage de donner plus précisément le moment où l'événement survient à l'intérieur d'une année, de sorte qu'on n'a pas besoin d'imposer arbitrairement l'ordre des événements. Cependant, la plupart des micro-simulations n'utilisent pas cette approche, car elle demande une fréquence de cueillette de données plus grande que celles habituellement disponibles.

l’immigration interprovinciale est modélisée plus loin dans le module « émigration », où est considérée l’émigration interprovinciale nette.)

On sélectionne ensuite aléatoirement les ménages qui immigreront à partir de la matrice d’immigrants potentiels jusqu’à ce que la cible de l’ISQ soit atteinte¹⁶. Pour la période suivante, les immigrants continuent d’être sélectionnés à partir de la base de données initiale d’immigrants. Il s’agit donc d’une sélection aléatoire avec remise. On attribue ainsi à tous les individus des ménages sélectionnés le statut d’immigrant. Les futurs enfants de ces immigrants, qui naîtront au Québec dans des périodes ultérieures, ne seront toutefois pas considérés comme des immigrants.

6.2 L’émigration

Le modèle distingue entre l’émigration internationale et l’émigration interprovinciale. Pour l’émigration interprovinciale, on tient seulement compte de l’émigration interprovinciale nette (émigration moins immigration interprovinciale) qui, selon les prévisions de l’ISQ, sera positive pour toutes les périodes à simuler. (L’émigration sera ainsi plus importante que l’immigration.)

Pour chaque période, nous déterminons d’abord le nombre prévu d’émigrants pour chaque catégorie d’émigration à partir des cibles d’émigration de l’ISQ. Le nombre d’émigrants est donc exogène au modèle. Nous sélectionnons ensuite aléatoirement les ménages qui émigrent jusqu’à ce que la cible soit atteinte. Pour l’instant, chaque ménage a la même chance d’émigrer. Les ménages (et leurs membres) sélectionnés pour émigrer sont simplement retirés de la base de données.

6.3 La mortalité

L’ISQ fournit des taux prévus de mortalité spécifiques au sexe pour chaque âge (de 0 à 119) et ce, pour chaque période de la simulation. À l’aide de ces taux, nous calibrons le nombre de décès pour chaque période en multipliant le nombre d’individus dans chaque groupe (spécifique à la période, à l’âge et au sexe) par le taux de mortalité du groupe.

Le nombre total de décès est ainsi exogène au modèle. Nous supposons que, pour un âge et un genre donné, chaque individu a la même probabilité de mourir durant une période donnée. Nous sélectionnons les observations pour chaque groupe jusqu’à ce que la cible de mortalité soit atteinte.

Les individus qui décèdent sont retirés de la base de données. Nous ajustons ensuite les caractéristiques des conjoints des individus décédées (indicateur de relation conjugale, temps de l’union, temps depuis le début du célibat, *etc*), et les caractéristiques de leur ménage (nombre de personnes, nombre d’enfants, *etc*.) Les orphelins majeurs (18 ans et plus) fondent un nouveau ménage. Pour l’instant, l’adoption n’est pas modélisée et les orphelins mineurs sont supprimés de la simulation, leur nombre étant négligeable.

16. La probabilité de chaque ménage d’être sélectionné est corrigée pour son poids statistique. Cette correction est également effectuée pour tous les autres événements.

6.4 Les départs du ménage

À chaque période, chaque individu qui a au moins 16 ans et qui habite toujours chez ses parents est soumis à une probabilité de quitter ce ménage. Nous devons calculer cette probabilité pour chaque individu. Notons que nous considérons uniquement les départs pour fonder un ménage seul ; le choix de quitter le ménage pour former un couple est modélisé dans le module de formation de couple. Nous utilisons d’abord les données de l’enquête sociale générale (ESG) de 2001 et de 2006 pour estimer un modèle logit dans lequel la variable dépendante est la décision de quitter le ménage (1 pour oui, 0 pour non). Les variables explicatives utilisées sont présentées dans l’annexe C.

Les paramètres de ce modèle de départs sont donc estimés à partir des données de l’ESG et sont ensuite importés dans SimUL pour déterminer les probabilités des jeunes individus de quitter leur ménage. Nous calibrons le nombre de départs de ménages de façon endogène au modèle. Plutôt que de laisser les probabilités déterminer librement le nombre total de départs pour une année, nous calibrons ce nombre en multipliant la probabilité de départ de chaque observation par son poids statistique, puis en prenant la somme de ces valeurs. Une fois cette cible de départs établie, nous sélectionnons les observations en fonction de leur probabilité de quitter le ménage jusqu’à ce que la cible soit atteinte.

La décision d’un individu de quitter le ménage de ses parents influencera ensuite sa probabilité de poursuivre ses études. Pour les parents de l’individu, le départ de leur enfant influencera leur offre de travail dans les périodes qui suivront.

6.5 La formation et la dissolution de couples

La formation et la dissolution de couples est un module important dans SimUL pour plusieurs raisons. Premièrement, former un couple est préliminaire à avoir des enfants (dans notre modèle). Deuxièmement, certaines formes de revenus peuvent être déclarées de façon conjointe par un couple pour des fins fiscales. Troisièmement, le fait d’être en couple influencera l’offre de travail des individus et l’estimation de leurs diverses formes de revenus dans SimUL. Finalement, il est souvent intéressant dans les simulations de comparer la situation des célibataires à celle des personnes en couple.

Tous les individus qui ne sont pas déjà en couple et qui ont entre 16 et 69 ans font face à chaque période à une probabilité de se trouver un conjoint. Comme pour le module précédent, les coefficients qui servent à calculer ces probabilités sont préalablement estimés à l’aide des données de l’ESG 2001 et 2006, par le biais de deux modèles Logit dans lesquels la variable dépendante est l’union avec un partenaire. Le premier modèle concerne les individus qui en sont à leur première union, alors que le deuxième concerne ceux qui ont déjà été en union et qui se sont séparés. Les variables explicatives utilisées pour chaque modèle sont présentées dans l’annexe C

Par la suite, on calcule à l’intérieur de SimUL les probabilités de tous les hommes et de toutes les femmes de former un couple. À l’aide de ces probabilités, on calibre les cibles d’unions séparément pour chaque sexe, qui seront donc différentes l’une de l’autre. Comme on doit forcément avoir le même nombre d’hommes que de femmes qui s’unissent (on ne modélise que les unions de même sexe), on utilise comme cible de nombre d’unions la moyenne des deux cibles. On sélectionne finalement les observations individuelles en fonction de leurs probabilités jusqu’à ce que la cible totale de formation de couples soit atteinte. Il

s'agit d'une présélection, car on doit encore associer chaque homme à une femme. Pour ce faire, on crée d'abord une mesure des différences d'âges ainsi que des niveaux de scolarité entre chaque homme et chaque femme. La mesure est la suivante :

$$D_{hf} = (A_h - A_f + 2)^2 + (S_h - S_f)^2, \quad (1)$$

où A_h et A_f sont l'âge de l'homme h et de la femme f , et S_h et S_f sont le nombre d'années de scolarité pour l'homme h et la femme f . On additionne 2 à la différence d'âge entre l'homme et la femme, car, dans l'échantillon initial, on observe que les hommes sont en moyenne deux ans plus vieux que leur conjointe. Pour chaque homme, nous créons ensuite une probabilité d'union avec chaque femme proportionnelle à l'inverse de cette mesure de distance. On sélectionne finalement la femme qui entre en union avec l'homme en tenant compte des probabilités de sélection, sous contrainte que le nombre total d'enfants dans la famille réunie ne dépasse pas le nombre maximal permis par notre modèle (8 enfants)¹⁷.

Un problème est que les hommes qui forment une union n'auront pas nécessairement le même poids statistique que celui de leur conjointe. On veut cependant attribuer un poids commun à tous les membres d'un même ménage. Pour contourner ce problème, nous utilisons le poids le plus faible des deux observations du couple pour déterminer le poids du nouveau ménage formé. Nous séparons ensuite l'observation avec un « excédent de poids » en deux. Une partie entre en relation avec le (la) conjoint(e) et l'autre non. On attribue à la première partie le poids du nouveau ménage (le poids plus faible du partenaire) et à la deuxième partie la différence entre les deux poids.

La deuxième partie de ce module est la dissolution des couples. Toutes les observations en couples font face à une probabilité de se séparer. À partir des données de l'ESG, on estime un modèle Logit où la variable dépendante est la dissolution et où les variables explicatives sont présentées dans l'annexe C.

On estime ensuite la probabilité de chaque personne de se séparer à l'aide des paramètres estimés, puis on prend la moyenne des deux probabilités dans le couple pour déterminer la probabilité de dissolution du couple. On calibre finalement le nombre de séparations en multipliant la probabilité de dissolution de chaque couple par son poids statistique, puis en en prenant la somme. Tous les enfants des ménages séparés suivent le même parent. Dans 90% des cas, la femme obtient la garde des enfants.

6.6 La natalité

La natalité est un module important dans SimUL, non seulement pour simuler la croissance démographique, mais aussi parce que cet événement influencera les choix et les caractéristiques futurs des parents. En effet, le nombre d'enfants engendrés et/ou l'âge des enfants servent entre autres de variables explicatives dans les modules de formation et de dissolution de couple, de scolarité et d'offre de travail.

17. Pour chaque homme, les probabilités générées sont normalisées pour que la somme des probabilités d'union avec chaque femme soit égale à un. On ordonne ensuite aléatoirement les femmes potentielles, puis on calcule la fonction cumulative des probabilités normalisées. On tire finalement une variable aléatoire d'une loi uniforme(0,1) puis on sélectionne la première femme pour laquelle la fonction cumulative est supérieure à la variable aléatoire.

Nous calibrons le nombre total de naissances à chaque période à partir des prévisions de l'ISQ. Le nombre total de naissances est donc exogène au modèle. La sélection des femmes qui donneront naissance à ces enfants est cependant déterminée à l'intérieur de SimUL selon les caractéristiques individuelles des femmes. Toutes les femmes entre 15 et 49 ans qui sont en couple et qui ont terminé leurs études font face à une probabilité de donner naissance à un enfant. Comme pour les modules précédents, on calcule la probabilité de chaque femme de tomber enceinte à partir d'un modèle Logit estimé sur les données de l'ESG de 2001 et de 2006. Les variables explicatives sont décrites dans l'annexe C. Nous imposons ensuite que le nombre de naissances soit égal à la prévision agrégée de l'ISQ pour l'année en question, puis nous sélectionnons les femmes qui tomberont enceintes avec des probabilités de sélection proportionnelles aux probabilités estimées plus haut.

6.7 L'éducation

Nous imposons d'abord que tous les individus étudient jusqu'à l'âge de 15 ans. De plus, nous ne modélisons pas les études à temps partiel pour le moment ; tous les étudiants étudient à temps plein et ne travaillent pas. Ce module considère les individus de 16 ans ou plus qui sont toujours aux études, évalue tout d'abord leurs probabilités d'obtenir un diplôme et, deuxièmement, leurs probabilités de quitter les études.

Les probabilités de diplomation sont calculées à partir des données de l'ESG de 2001 et de 2006. On utilise une fonction de survie, qui détermine la probabilité d'un individu d'obtenir un diplôme en tenant compte du fait qu'il n'est toujours pas diplômé. Pour le diplôme d'études secondaire (DES), on utilise l'âge de l'individu. La fonction de survie détermine donc la probabilité de diplomation à l'âge de 16 ans. Elle détermine ensuite cette probabilité pour les individus de 17 ans, conditionnellement à la population des individus de 17 ans n'ayant pas été diplômés à 16 ans. Le même exercice est effectué pour les diplômes collégiaux et universitaires, mais on utilise le nombre d'années d'études pour ces diplômes plutôt que l'âge. Pour l'instant, les autres caractéristiques individuelles n'affectent pas les probabilités de diplomation. Notons que nous ne calibrons pas le nombre de diplomations pour l'instant.

On identifie ensuite les personnes qui quittent définitivement les études. Les probabilités des individus sont calculées à partir d'un modèle Logit. Pour les individus qui habitent toujours chez leurs parents, les variables explicatives sont présentées dans l'annexe C.

6.8 Le marché du travail

Le module du marché du travail modélise tout d'abord la participation au marché du travail pour les individus entre 16 et 69 ans. Pour l'instant, nous ne permettons pas aux étudiants de travailler et nous supposons que personne ne travaille à partir de 70 ans. Aussi, nous ne tenons pas compte de la distinction entre travailleur salarié et travailleur autonome. Nous prévoyons tenir compte de ces éléments additionnels dans les versions futures de SimUL. Nous utilisons les coefficients d'un modèle Logit estimé à partir des données de l'EDTR, où la variable dépendante est une variable binaire (1 si l'individu travaille, 0 sinon) et où les variables explicatives sont présentées dans la section C.6 de l'annexe C. Le statut d'emploi influencera les estimations des revenus de retraite dans les prochains modules.

Ce module modélise ensuite le salaire des individus employés en fonction de leurs caractéristiques. Les salaires prédits sont d’abord estimés à l’aide des coefficients d’une régression des moindres carrés ordinaires effectuée sur les données de l’EDTR. Les variables utilisées sont décrites dans l’annexe C. Cette régression tient compte du problème de sélection potentiel. En effet, on observe seulement les salaires pour ceux qui travaillent. Il est possible que ces individus travaillent à cause de facteur non-observés, par exemple des opportunités d’emplois supérieures à la moyenne. Ceci fera que les revenus d’emplois observés dans les données auront tendance à être associés à des termes d’erreurs élevés. Une régression linéaire simple, qui suppose que les termes d’erreur sont centrés à zéro, estimera donc des coefficients biaisés. C’est pourquoi cette régression inclue une correction de Heckman (voir Heckman (1979)), qui utilise une variable indicatrice égale à 1 s’il y a présence d’au moins un enfant en âge préscolaire (moins de 5 ans) dans le ménage. Cette variable est utilisée comme restriction d’exclusion, c’est-à-dire qu’on suppose qu’elle affecte la décision de travailler sans affecter le salaire prédit. Cette variable entraîne donc des variations exogènes dans la participation au marché du travail. La correction de Heckman permet ainsi d’utiliser ces variations exogènes (qui ne sont pas dû à un terme d’erreur particulièrement élevé sur le revenu) pour corriger les coefficients. Une fois les coefficients estimés, on les utilise pour estimer les salaires prédits des individus à l’intérieur de SimUL. On leur attribue ensuite un terme d’erreur tiré des résidus du modèle afin de reproduire la variance observée de la distribution.

Le module du marché du travail modélise finalement le nombre d’heures travaillées par semaine. On crée trois catégories : moins de 30 heures, entre 30 et 45 heures, et plus de 45 heures. On estime ensuite un modèle Logit multinomial à l’aide des données de l’EDTR. Ce modèle estime les effets des variables explicatives décrites dans l’annexe C sur les probabilités d’être dans chacune des trois catégories. On note que ces variables incluent le salaire prédit de façon à obtenir une cohérence entre le nombre d’heures travaillées et le salaire annuel. Une fois les coefficients estimés, on utilise ceux-ci à l’intérieur de SimUL pour sélectionner la catégorie d’heures travaillées pour chaque individu qui travaille selon les probabilités qui découlent du modèle Logit. Afin de reproduire la distribution d’heures, nous tirons ensuite pour chaque individu une valeur de la distribution empirique des heures travaillées à l’intérieur de la catégorie de l’individu.

6.9 Les revenus de placements

Les revenus de placements sont une source de revenus non négligeable pour les personnes âgées. Dans SimUL, chaque individu de 16 ans et plus a une certaine probabilité de recevoir des revenus de placements. Cette probabilité est calculée à l’aide des paramètres d’un modèle Logit estimé sur les données de l’EDTR. Les variables explicative sont décrites dans l’annexe C.

On doit ensuite calculer le montant que l’individu reçoit. On utilise une estimation par moindres carrés ordinaires sur les données de l’EDTR avec correction de Heckman pour corriger pour le problème de sélection. Les variables explicatives sont décrites dans l’annexe C. Les coefficients estimés sont ensuite utilisés pour calculés les revenus de placements des individus qui en reçoivent à l’intérieur de SimUL.

6.10 Les revenus de retraite

Le système canadien et québécois de revenus de retraite est composé de trois paliers¹⁸. Le premier est le Programme fédéral de la sécurité de la vieillesse, constitué de la Pension de la sécurité de la vieillesse (SV), du Supplément de Revenu garanti (SRG), de l'Allocation et de l'Allocation au survivant. Ce programme est considéré comme une mesure d'assurance et de lutte contre la pauvreté visant à venir en aide aux personnes âgées dont les revenus sont insuffisants. Ce régime est financé par les impôts des contribuables.

Le deuxième palier est constitué de prestations de retraite de régimes publics, financées par les cotisations des travailleurs et des employeurs. Le Régime de pension du Canada (RPC) s'applique dans toutes les provinces canadiennes sauf au Québec, cette dernière ayant son propre régime de pension, soit le Régime des rentes du Québec (RRQ), qui est géré par la Régie des rentes du Québec. Ces deux régimes homologues, en plus d'offrir des pensions de retraite, offrent des prestations d'invalidité et des prestations de survivant.

Le dernier pilier est celui des régimes enregistrés d'épargne-retraite (REER), qui représentent un abri fiscal pour l'épargne en vue de la retraite, et des régimes complémentaires de retraite (RCR) qui sont mis en place par un employeur. Ces deux derniers paliers sont constitués de mesures d'assurance, soit des mesures visant à permettre une certaine continuité du revenu au moment de la retraite. Aussi, en plus des trois paliers canadiens, certaines provinces ont choisi de mettre en place différents programmes d'aide aux aînés ou aux retraités, qui peuvent être pris en compte lorsque vient le moment de la retraite. Ceux-ci peuvent consister tant en un allègement au niveau de l'imposition qu'en des prestations de vieillesse accrues. De plus, il existe, au delà de ces piliers, un ensemble relativement complexe de programmes gouvernementaux et de systèmes de gestion des épargnes pour la retraite.

SimUL modélise chacun de ces trois paliers de façon relativement simplifiée. Nous débutons par le modèle de prestations de RRQ qui, pour l'instant, considère seulement les prestations de retraite et ignore les prestations d'invalidité et de survivant. Le module sélectionne d'abord tous les individus d'au moins 60 ans qui ne reçoivent pas encore de prestation de la RRQ et calcule leurs probabilités de commencer à en recevoir. Ces probabilités sont calculées à l'aide des coefficients d'un modèle Logit estimé à l'aide des données de l'EDTR. Les variables explicatives utilisées sont énumérées dans la section C.7 de l'annexe C. Notons que nous incluons un effet de cohorte pour les femmes, car les femmes plus jeunes ont un taux d'emploi beaucoup plus élevé que celui qu'avaient les femmes dans le passé. Les jeunes femmes d'aujourd'hui cotisent donc en plus grand nombre au régime de la RRQ et recevront donc davantage de prestations dans le futur, ce qui est capté par l'effet de cohorte des femmes dans le modèle.

On doit ensuite calculer les montants des prestations de la RRQ pour les personnes qui en reçoivent pour la première fois. Nous estimons par les moindres carrés ordinaires (MCO) l'effet des variables explicatives (énumérées dans la section C.8 de l'annexe C) sur les prestations en incluant la correction de Heckman pour corriger pour un biais potentiel de sélection. Une fois qu'un individu commence à recevoir une prestation de la RRQ, il

18. Karim MOUSSALY, « Participation aux régimes privés d'épargne-retraite », Statistique Canada, [En ligne], 2010, <http://www.statcan.gc.ca/pub/13f0026m/13f0026m2010001-fra.htm>, (consulté le 16 juin 2011).

conserve le même niveau de prestation pour le reste de sa vie ¹⁹.

Nous modélisons ensuite les revenus de pensions privés provenant de régimes complémentaires de retraite. Les REER ne peuvent cependant pas être estimés séparément car ils sont considérés comme des revenus de pension après l'âge de 65 ans à la fois dans la BDMSPS et l'EDTR. Avant 65 ans, il est impossible de savoir si les retraits de REER observés ont été utilisés comme des revenus de retraite ou à d'autres fins.

Nous déterminons d'abord les probabilités de tous les individus de plus de 50 ans de recevoir des revenus de pensions. Les variables explicatives du modèle Logit utilisé sont présentées dans la section C.9 de l'annexe C. Nous calibrons ensuite le nombre de prestataires de pension de retraite sur le nombre prédit pour chacune des classes d'âges suivante : 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84 et 85 et plus. Si, pour une classe d'âge donnée, on a un nombre de prestataires prédit supérieur au nombre actuel, on sélectionne les nouveaux prestataires avec des probabilités de sélection proportionnelles aux probabilités calculées précédemment, jusqu'à ce qu'on atteigne le nombre prédit. Si le nombre prédit est inférieur au nombre de prestataires, on sélectionne ceux qui cesseront de retirer des revenus de pension avec une probabilité proportionnelle à l'inverse des probabilités des individus de recevoir des pensions. Les montants de pensions reçus sont ensuite estimés par moindres carrés ordinaires avec correction de Heckman ; voir la section C.10 pour la liste des variables.

À la différence des prestations de la RRQ, les revenus de pensions ne sont pas fixés pour toujours une fois que l'individu commence à les recevoir. Nous utilisons plutôt une fonction de différence, qui estime par MCO le revenu de pension pour une période en fonction de celui de la période précédente et d'autres variables ; voir la section C.11.

Le dernier palier à modéliser est le Programme fédéral de la sécurité de la vieillesse. Nous tenons compte ici des pensions de sécurité de la vieillesse et de supplément de revenu garanti, mais pas des allocations et allocations au survivant. Contrairement aux éléments précédents, nous calculons les montants selon les règlements en vigueur en 2004 (comme année de référence) plutôt que de les estimer à l'aide de modèles économétriques.

7 Prévisions de SimUL

Cette section présente les prévisions de SimUL pour diverses variables démographiques et économiques. La Figure 2 montre d'abord les variations de population prévues par SimUL jusqu'à 2030. Les variations positives proviennent de la natalité et de l'immigration (internationale) et les variations négatives proviennent de la mortalité et de l'émigration (internationale + interprovinciale nette). Tel qu'expliqué plus haut, le nombre total d'individus affectés par chacun de ces quatre événements est calibré sur des prévisions exogènes au modèle et provenant de l'ISQ. Il est donc naturel que les prévisions de SimUL correspondent de très près à celles de l'ISQ.

Le premier élément d'intérêt est la comparaison entre la natalité et la mortalité. Le nombre annuel prévu de nouveau-nés diminue légèrement de 2010 à 2030, alors que le nombre annuel prévu de décès augmente fortement sur la même période. On prévoit ainsi

19. SimUL utilise des dollars en valeur de 2004 pour toutes les estimations et pour toutes les simulations. Nous n'avons donc pas à indexer les prestations selon l'inflation.

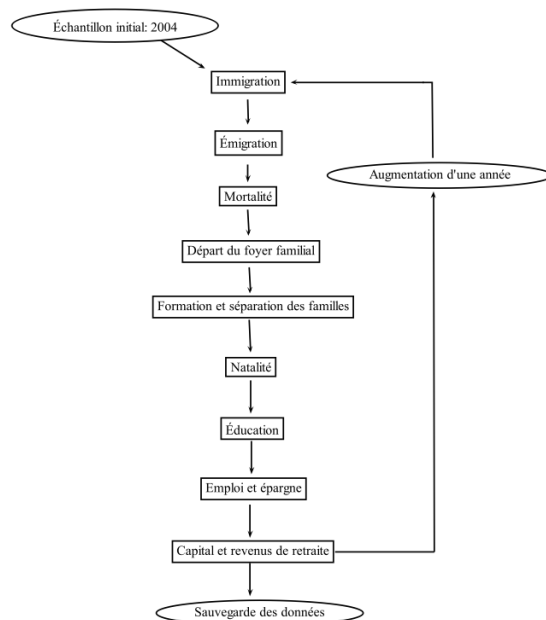


FIGURE 1 – Structure de SimUL

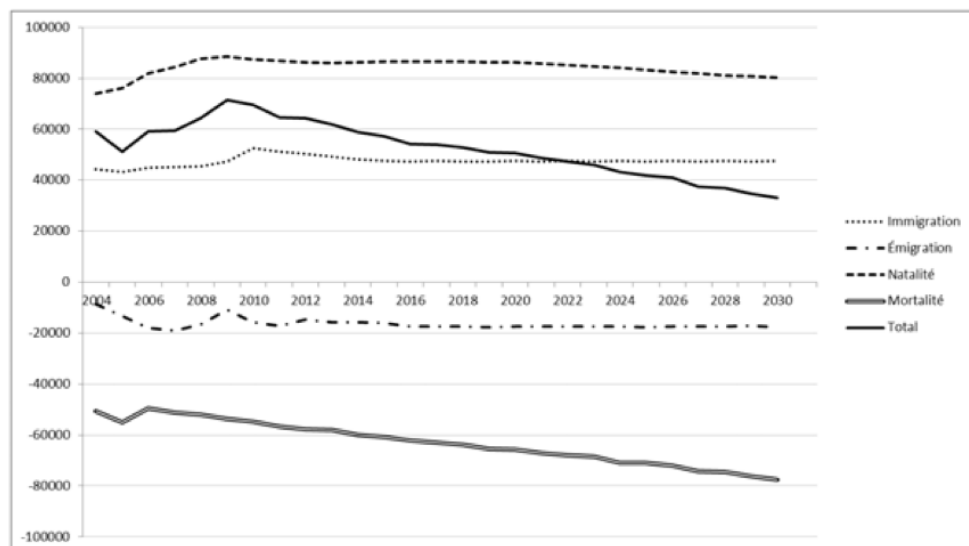


FIGURE 2 – Variations de population

que le nombre de décès rattrapera le nombre de naissances autour de 2030. L'immigration, qui continuera de surpasser largement l'émigration, viendra contrebalancer cette tendance. En additionnant toutes ces variations, on constate que la croissance totale de la population demeurera positive, mais diminuera continuellement avec le temps.

La Figure 3 illustre comment évoluera la population par tranche d'âge. Les classes d'âge sont construites de façon à séparer très sommairement les individus qui en sont à différentes étapes dans leur vie : les enfants de 0 à 15 ans, les individus de 16 à 25 ans en âge d'étudier ou de commencer à travailler, ceux de 26 à 55 ans principalement en âge de travailler, les individus de 56 à 70 ans qui doivent décider s'ils prennent leur retraite et ceux de 71 ans et plus qui sont presque tous retraités.

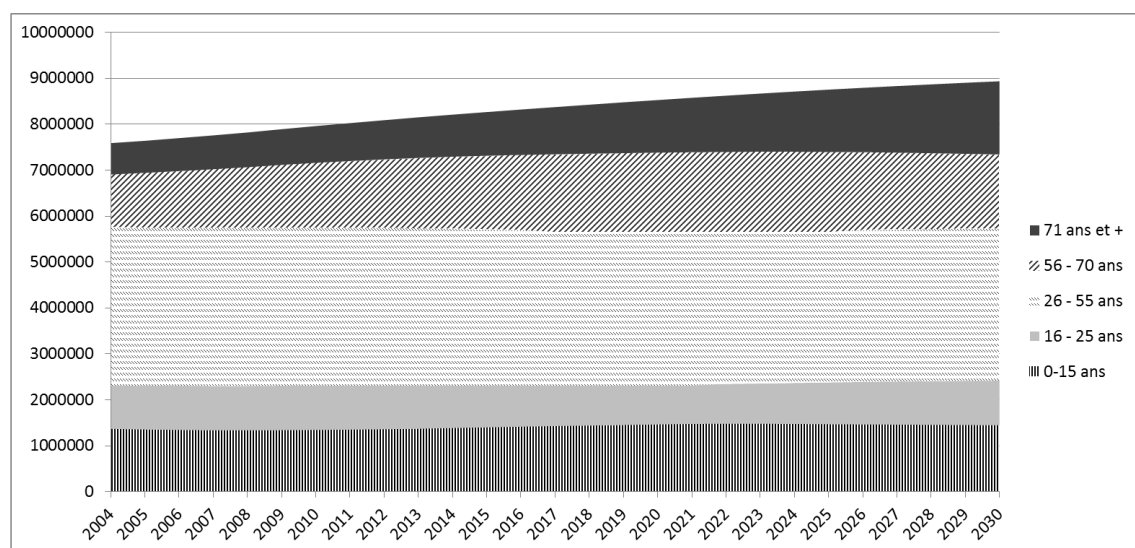


FIGURE 3 – Population par classes d'âges

La Figure 3 illustre clairement la situation. La croissance de la population de 2010 à 2030 proviendra presque exclusivement de la croissance de la population en âge de retraite ; la taille de la population dans les autres classes d'âge demeurera plutôt stable. En termes de distribution d'âge, la population non-retraîtée du Québec en 2030 sera très semblable à celle observée aujourd'hui, alors que les retraités seront considérablement plus nombreux.

Il est intéressant de s'attarder plus en détails aux effets de l'immigration, car ceux-ci peuvent contrebalancer les problèmes liés au vieillissement de la population en menant à une population en moyenne plus jeune. Rappelons que, dans notre modèle, un immigrant est défini comme étant un individu qui a immigré au Québec en provenance de l'extérieur du Canada ; les immigrants interprovinciaux ne sont pas considérés comme immigrants, car ils font partie de l'émigration interprovinciale nette (émigration-immigration). De plus, quand une immigrante donne naissance à un enfant au Québec, ce dernier n'est pas considéré comme un immigrant, ce qui peut avoir comme effet de sous-estimer l'apport de l'immigration à la croissance de la population.

La Figure 4 présente l'évolution de la population du Québec jusqu'à 2030 pour les immigrants et les non-immigrants. On voit qu'il y aura environ deux fois plus d'immigrants

en 2030 qu'en 2010. Aussi, la part des immigrants dans la population augmentera significativement par rapport à celle des non-immigrants. Sans immigration, la croissance de la population totale dans les années à venir serait pratiquement nulle.

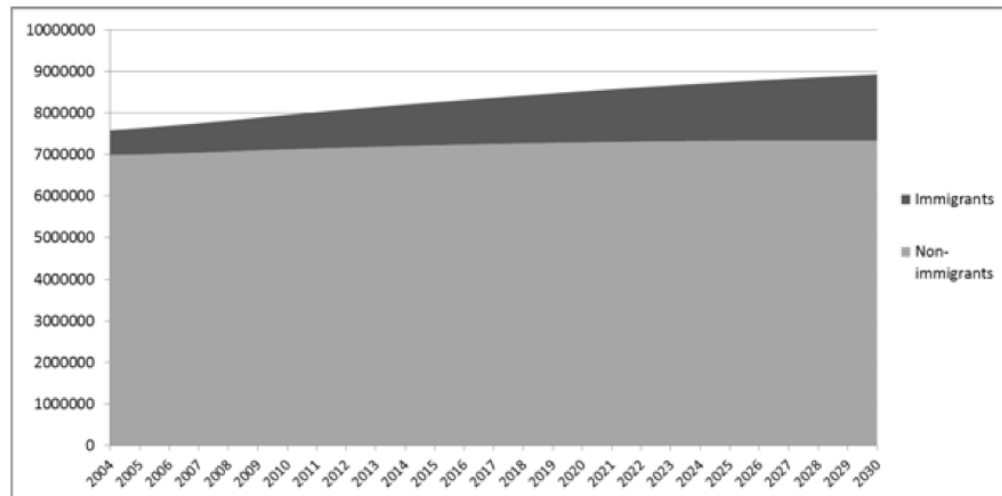


FIGURE 4 – Population immigrante et non-immigrante

Il est utile de voir comment l'immigration affecte la distribution de la population par âge. La Figure 5 présente la distribution de l'âge des immigrants dans la base d'immigrants de SimUL. Cette base contient un échantillon représentatif des individus qui ont immigré au Québec entre 1996 et 2001. Notre module d'immigration sélectionne aléatoirement (avec remise) dans cette base les ménages qui immigreront au Québec. En d'autres mots, un échantillon aléatoire de cette distribution s'ajoute à la population du Québec à chaque année de la simulation. On voit ainsi comment l'immigration fournit au Québec une population qui est davantage en âge de travailler, sans quoi les effets du vieillissement décrits plus haut seraient probablement amplifiés.

La Figure 6 illustre plus clairement l'effet de l'immigration en séparant les populations immigrante et non-immigrante par âge selon les prévisions du modèle. La distribution de 2004 montre un effet du vieillissement de la population : le nombre de jeunes en âge de commencer à travailler deviendra bientôt inférieur au nombre d'individus en âge de prendre leur retraite. Les prévisions pour les années suivantes montrent dans quelle mesure l'immigration contrebalancera en partie de phénomène en augmentant le rapport entre la population en âge de travailler et celle à la retraite.

Pour ce qui est de la mortalité, SimUL utilise les taux de mortalité prévus par l'ISQ pour chaque âge (de 0 à 120 ans) jusqu'à 2030. Un élément intéressant est la comparaison entre les prévisions des espérances de vie des hommes et des femmes : les femmes ont jusqu'à

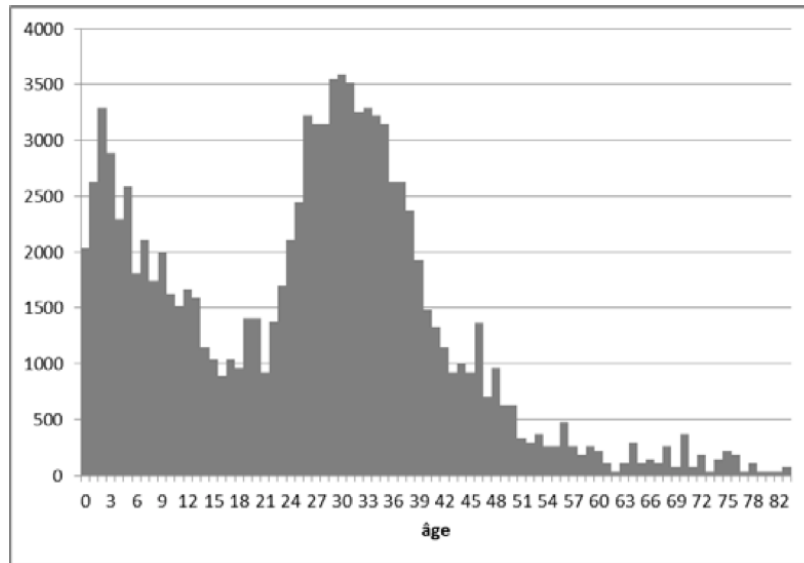


FIGURE 5 – Population d’immigrants potentiels

ce jour une espérance de vie plus élevée que les hommes, mais on prévoit que cet écart diminuera avec le temps. Cela se reflète dans les prévisions de SimUL, comme le montre la Figure 7, qui présente la proportion d’hommes dans la population pour les 0 à 69 ans et pour les 70 ans et plus et ce, jusqu’à 2030. La proportion d’hommes pour les 69 ans et moins est toujours proche de 0,5, ce qui est attendu. Pour les 70 ans et plus, la proportion d’hommes en 2004 est d’environ 0,4. À mesure que l’espérance de vie des hommes rattrape relativement celle des femmes, cette proportion se rapproche de 0,5.

Un autre module important de SimUL est la scolarité. La qualité de la modélisation effectuée par ce module est primordiale, car la scolarité influence fortement la participation au marché du travail et les revenus de travail futurs, qui à leur tour influencent l’épargne et les revenus de retraite. Aussi importe-t-il de vérifier la cohérence des résultats obtenus par ce module, entre autres car la diplômation totale et la décision de quitter les études ne sont pas calibrés sur des prévisions exogènes au modèle. Les résultats sont générés uniquement par les probabilités estimées par le modèle logit décrit plus haut.

La Figure 9 présente les prévisions des proportions des diplômés par âge. La Figure 8 présente les statistiques de la base initiale. On distingue entre le diplôme d’études universitaires (Université), le diplôme d’études collégiales (DEC), le diplôme d’études secondaires (DES) et l’absence de diplôme (Aucun).

La Figure 8 montre les proportions de diplômés observés pour chaque diplôme par âge dans notre base initiale. On constate la prévalence plus importante de niveaux d’éducation élevés pour les cohortes plus jeunes (du moins, après 20 ans). L’effet est important, et il y a lieu de se demander si cette tendance se poursuivra pour les nouvelles cohortes. De manière à être « conservateurs », nous avons choisi d’inclure des effets de cohortes seulement dans le modèle de choix de quitter les études. Nous ne prolongeons pas cette tendance très forte, car

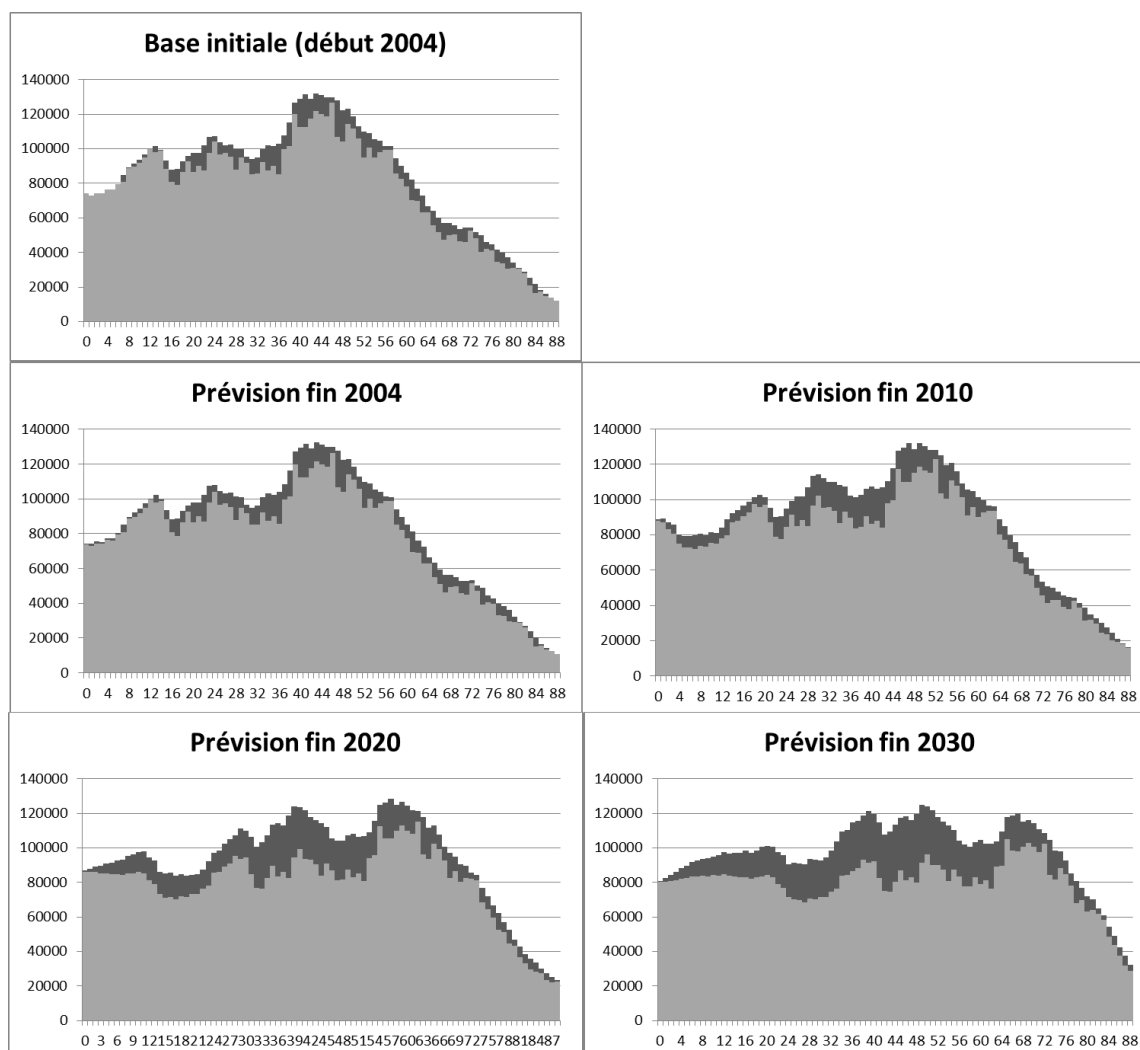


FIGURE 6 – Immigrants et non-immigrants par âge

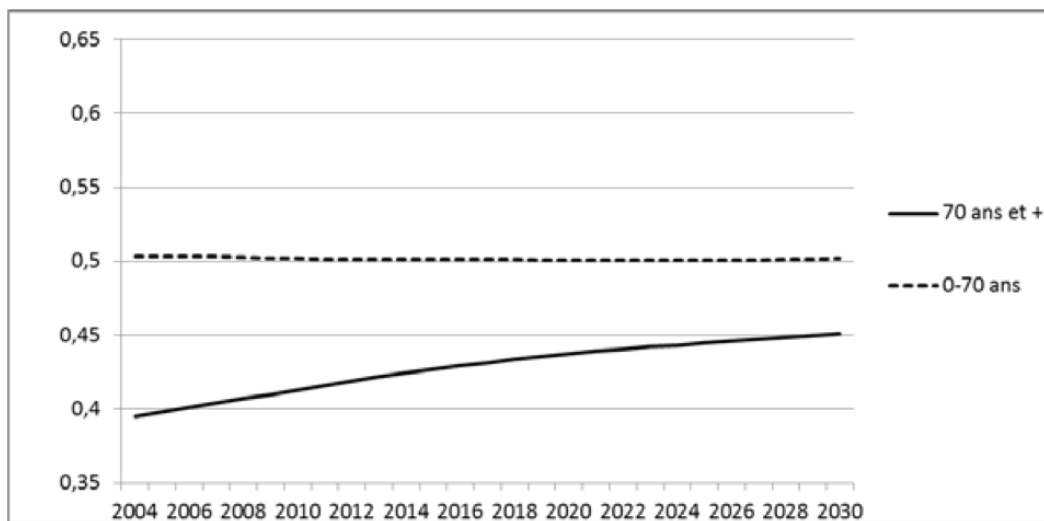


FIGURE 7 – Proportion d’hommes dans la population

nous utilisons seulement les observations qui ont terminé leurs études dans les 30 années précédant les enquêtes utilisés (ESG 2001). Nous prolongeons ainsi la légère tendance à la hausse observée dans le niveau de scolarité des dernières années, ce que montrent les graphiques pour les périodes futures.

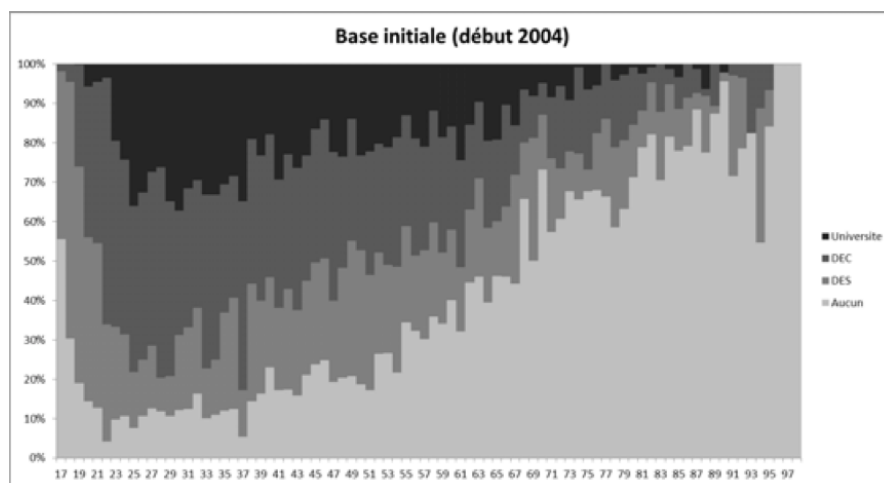


FIGURE 8 – Proportion des diplômes par âge

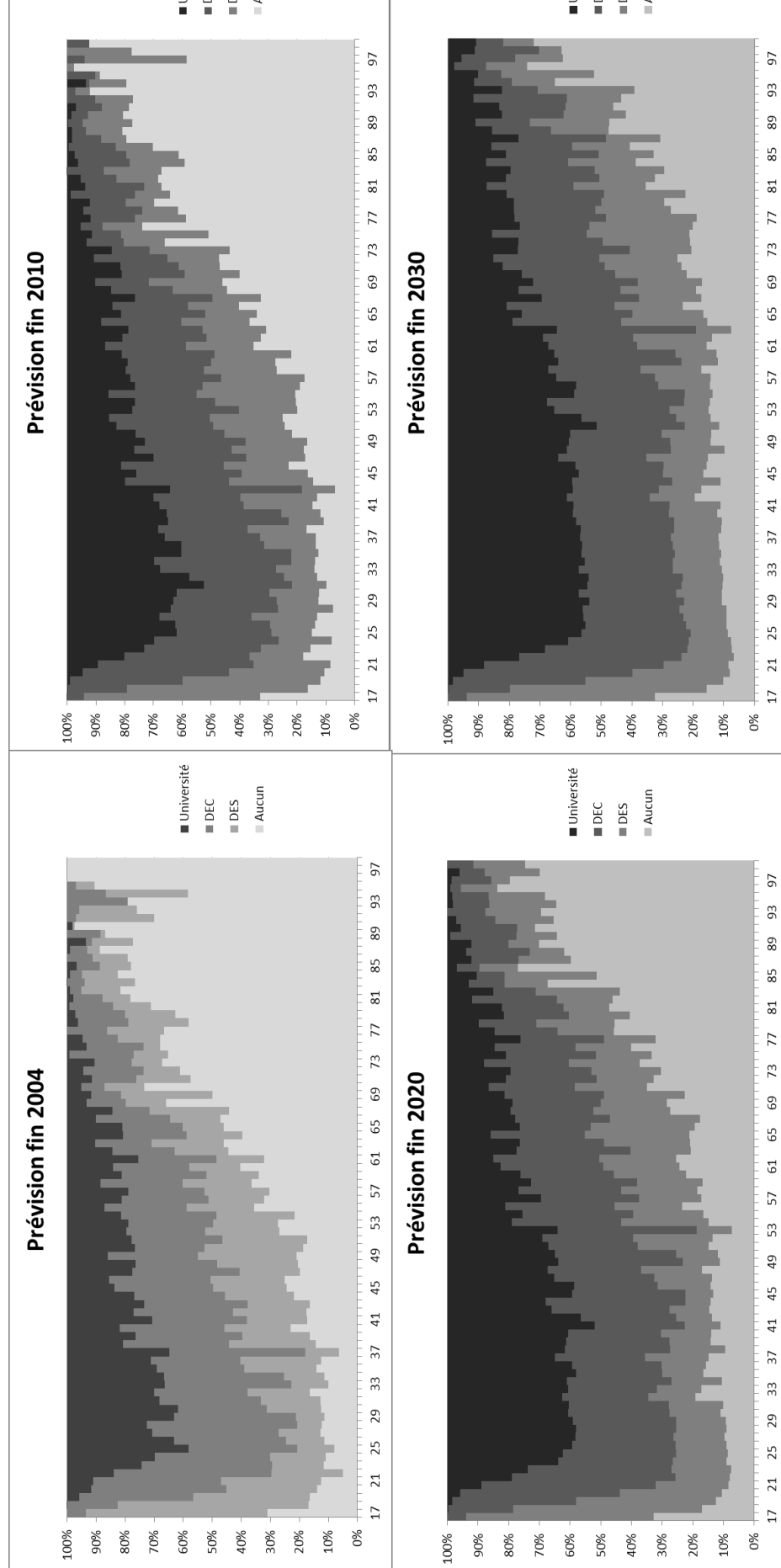


FIGURE 9 – Proportion des diplômes par âges

La Figure 10 porte sur le nombre de diplômés pour chaque diplôme dans la population en âge de travailler (de 16 à 65 ans). On voit que, bien qu'on ne puisse s'attendre à une croissance dans le nombre de travailleurs disponibles, il est possible d'anticiper une augmentation dans le niveau de scolarité de ces travailleurs. Ceci est observable entre autres par l'augmentation du nombre de diplômés universitaires. On observera également, toujours selon les prévisions de SimUL, moins d'individus entre 16 et 65 ans sans diplôme.

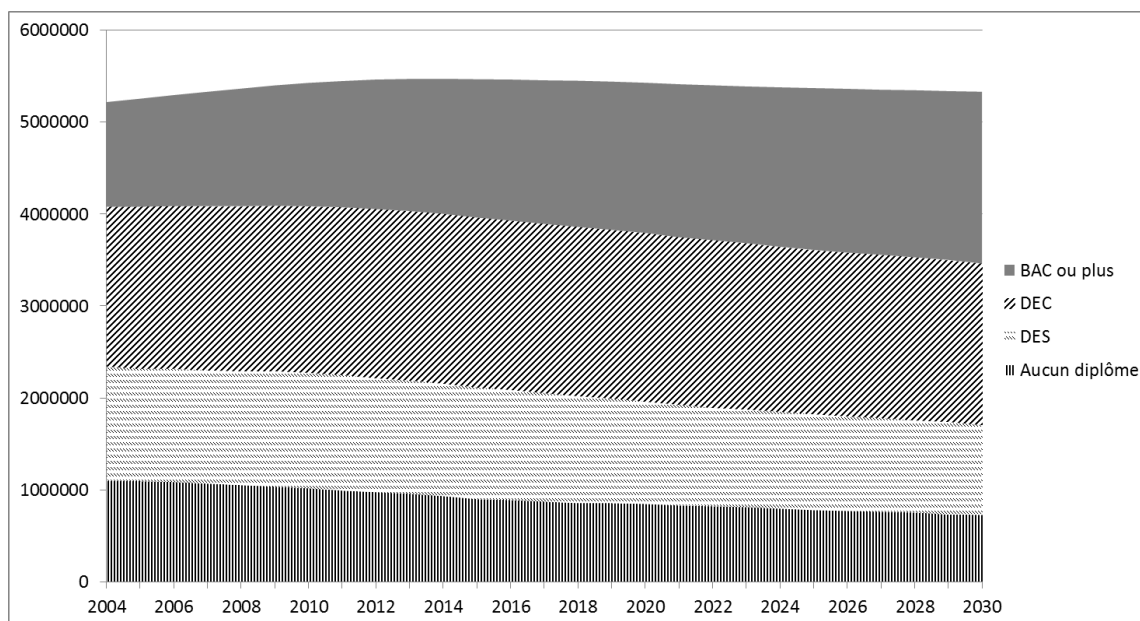


FIGURE 10 – Population entre 16 et 65 ans par diplôme obtenu

Un module primordial de SimUL est celui de l'emploi. La Figure 11 présente les proportions de travailleurs par âge. Les différences entre les années sont présentes, mais peu perceptibles ; pour plus de clarté, elles sont présentées dans la Figure 12 (par rapport à 2004). L'évolution des graphiques de 2004 à 2030 montre deux tendances.

Premièrement, les individus dans la vingtaine travaillent de moins en moins, ce qui vient en grande partie de la tendance des jeunes à étudier plus longtemps. Deuxièmement, les individus dans la quarantaine et la cinquantaine travaillent davantage. Ceci provient encore une fois de l'éducation : cette tranche d'âge sera passablement plus éduquée en 2030 qu'en 2004 (voir la Figure 9), ce qui influencera l'emploi. On voit cependant qu'une fois l'âge de la retraite atteint, ces individus n'auront pas tendance à prolonger leur carrière davantage.

La participation au marché du travail se mesure également par le nombre d'heures travaillées. La Figure 13 présente les prévisions des distributions d'heures de travail par semaine dans la population des hommes de 16 à 64 ans. Afin de bien cerner l'évolution des distributions, les prévisions pour 2010, 2020 et 2030 sont présentées en différences par rapport aux prévisions pour la fin de 2004. On observe dans la base initiale qu'environ 25% des hommes dans ces âges ne travaillent pas et que la grande majorité des autres travaillent 40 heures ou un peu moins. Les prévisions pour les années suivantes montrent peu de différences par rapport à 2004, sauf pour ce qui est de l'augmentation de la proportion

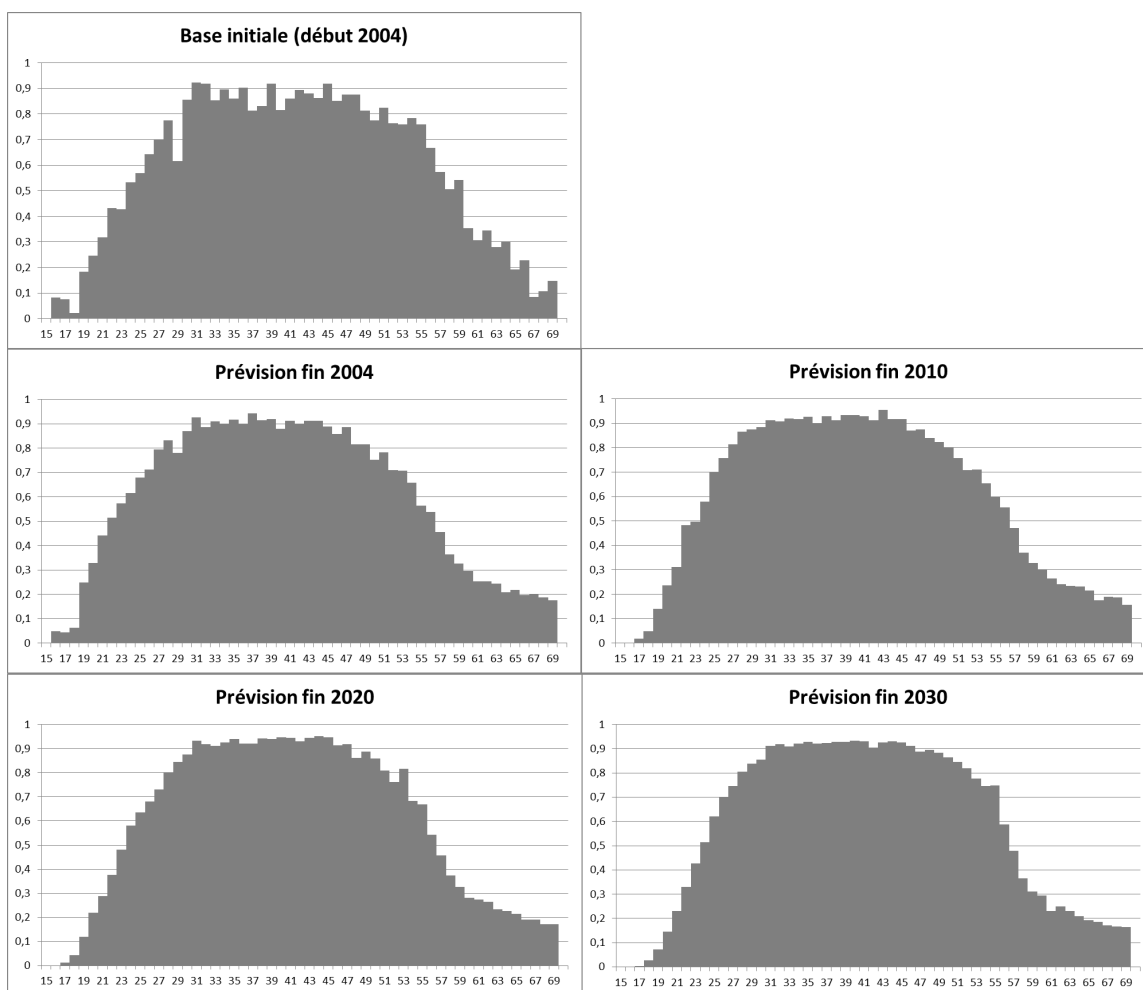


FIGURE 11 – Taux d'emploi par âge

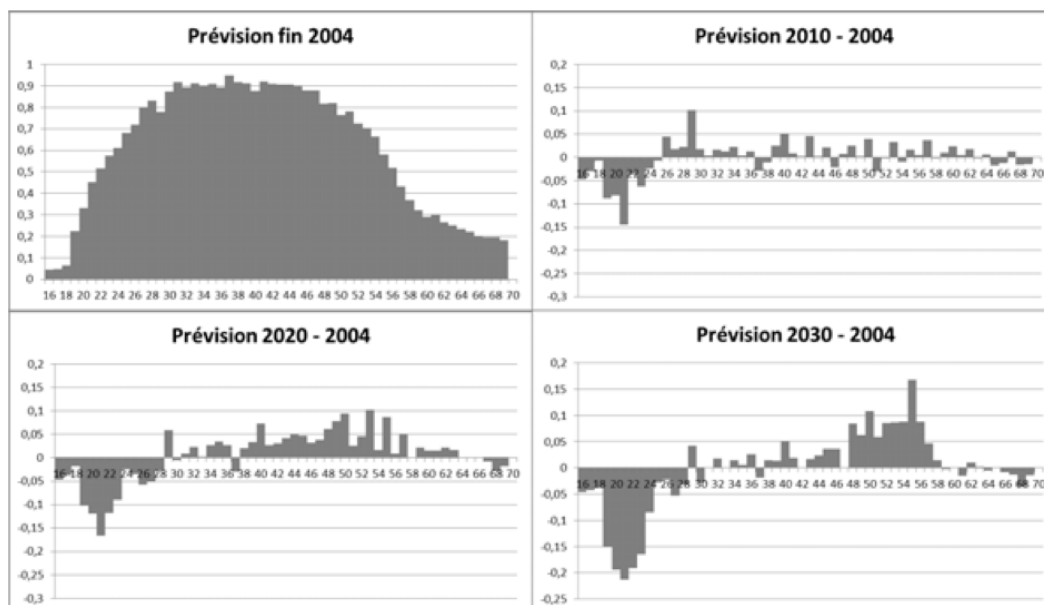


FIGURE 12 – Taux d’emploi par âge et différences par rapport à 2004

des hommes entre 16 et 64 ans qui ne travaillent pas. La cause probable est que plusieurs individus prennent leur retraite avant 65 ans, et que le vieillissement de la population augmentera la proportion des individus en âge de prendre leur retraite.

La Figure 14 présente les mêmes distributions, mais pour les femmes entre 16 et 64 ans. La base initiale montre deux différences par rapport à la distribution des heures de travail pour les hommes. Premièrement, la proportion des femmes qui ne travaillent pas est beaucoup plus grande (presque 40%, contre 25% pour les hommes). Deuxièmement, la proportion des femmes qui travaillent 40 heures n’est pas aussi élevée que pour les hommes. On observe plutôt deux sommets moins importants à 35 et à 40 heures. L’évolution des distributions montre également deux tendances : une légère diminution dans le nombre de femmes qui ne travaillent pas et une augmentation plus importante de la proportion des femmes qui travaillent entre 30 et 40 heures.

Un élément intéressant lié à l’emploi et aux changements démographiques est le ratio de dépendance. La méthode traditionnelle pour mesurer ce ratio est de diviser le nombre de personnes qui ne sont pas en âge de travailler (les 0 à 15 ans et les 65 ans et plus) par le nombre de personnes en âge de travailler (les 16 à 64 ans). Ainsi, un ratio de dépendance de 0,5 signifie qu’on a deux fois plus de personnes en âge de travailler que de personnes « dépendantes » (les enfants et les personnes en âge de retraite). Une augmentation dans la valeur de ce ratio suggère donc que chaque contribuable devra contribuer davantage aux recettes gouvernementales pour fournir les mêmes services à toute la population, incluant les enfants (*e.g.*, éducation) et les retraités (*e.g.*, soins de santé et revenus de retraite), toutes choses étant égales par ailleurs. Ce ratio de dépendance traditionnel est présenté dans

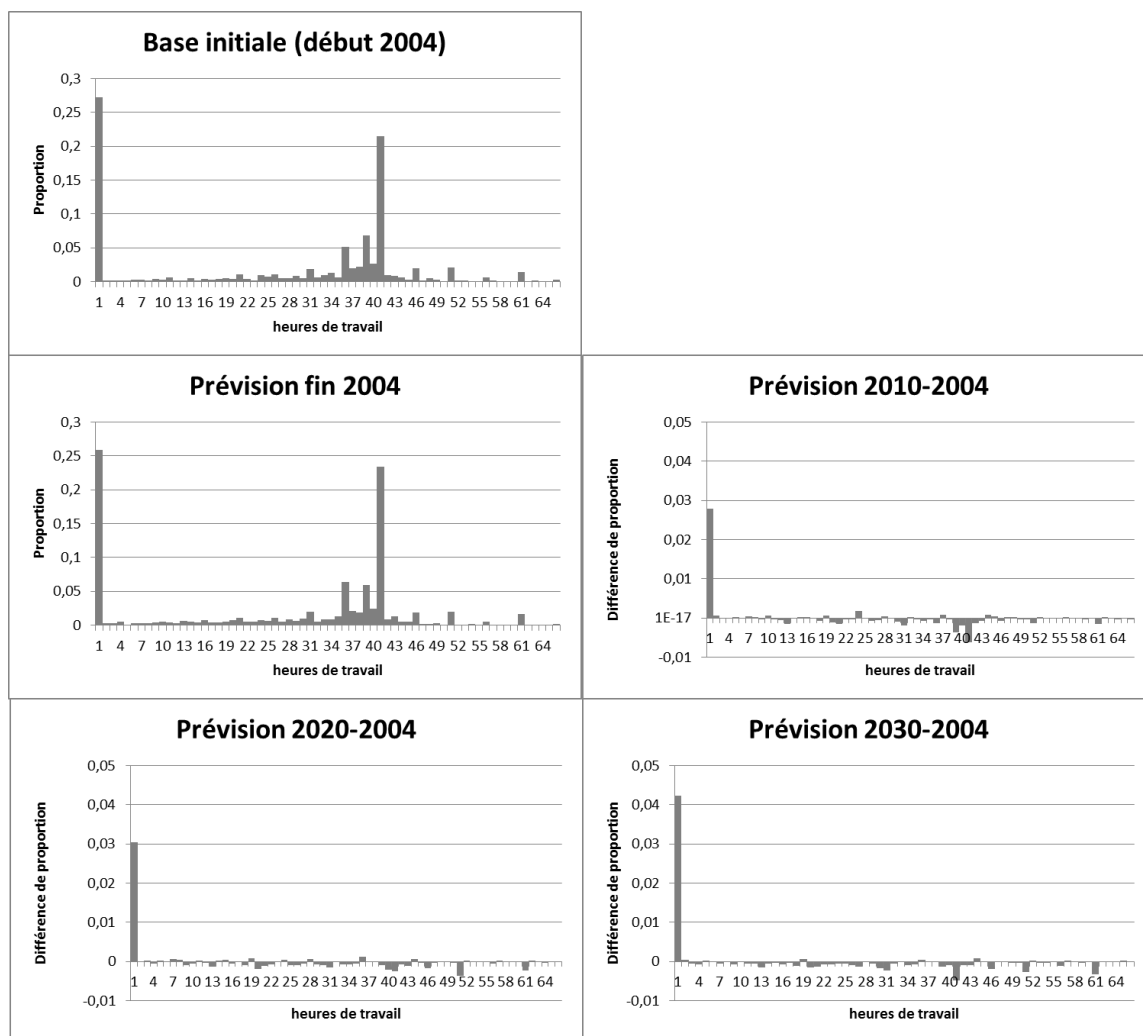


FIGURE 13 – Distribution des heures travaillées par semaine pour les hommes de 16 à 64 ans

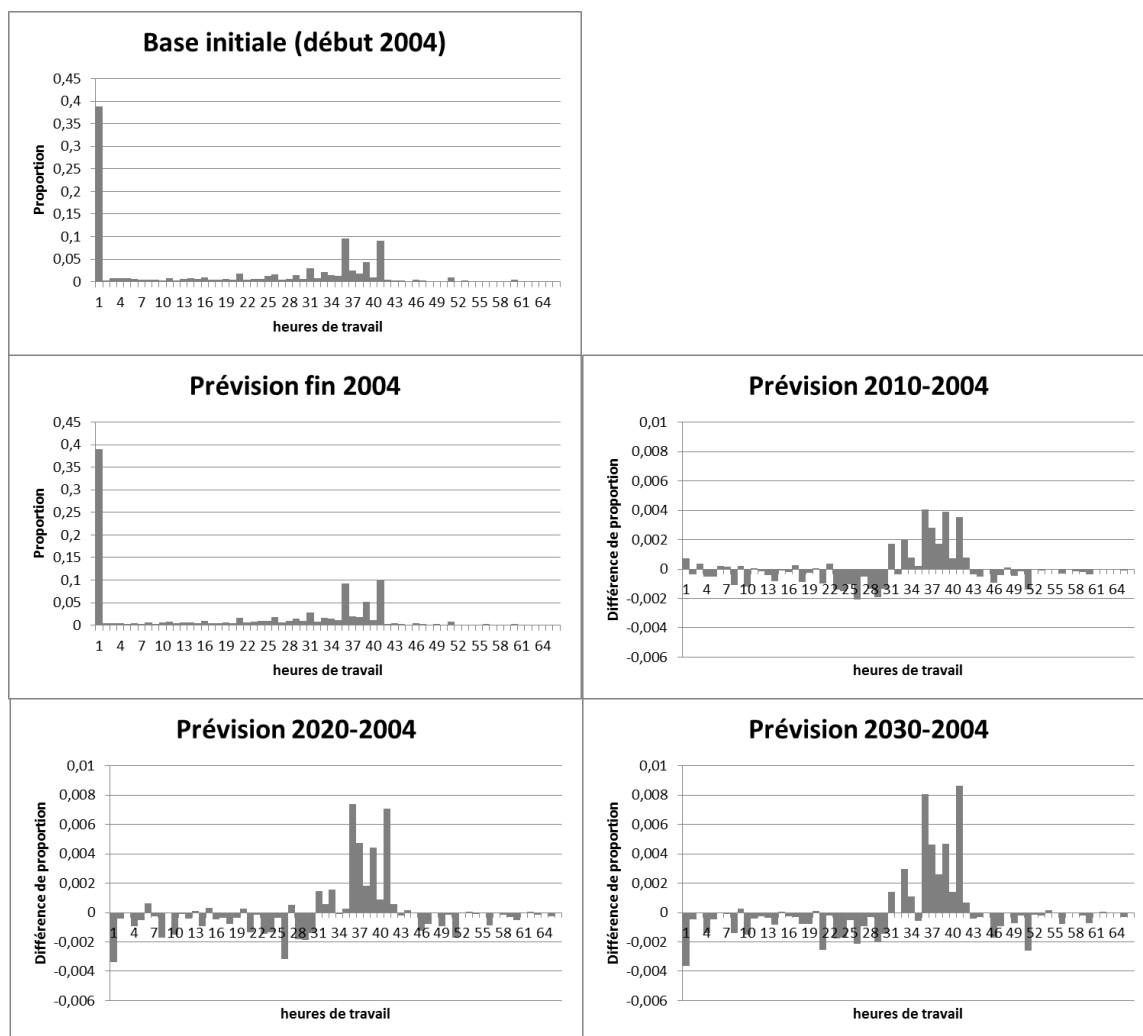


FIGURE 14 – Distribution des heures travaillées par semaine pour les femmes de 16 à 64 ans

la Figure 15. Sans surprise, le vieillissement de la population entraînera une importante augmentation du ratio de dépendance, qui passera de moins de 0,5 en 2010 à plus de 0,7 en 2030.

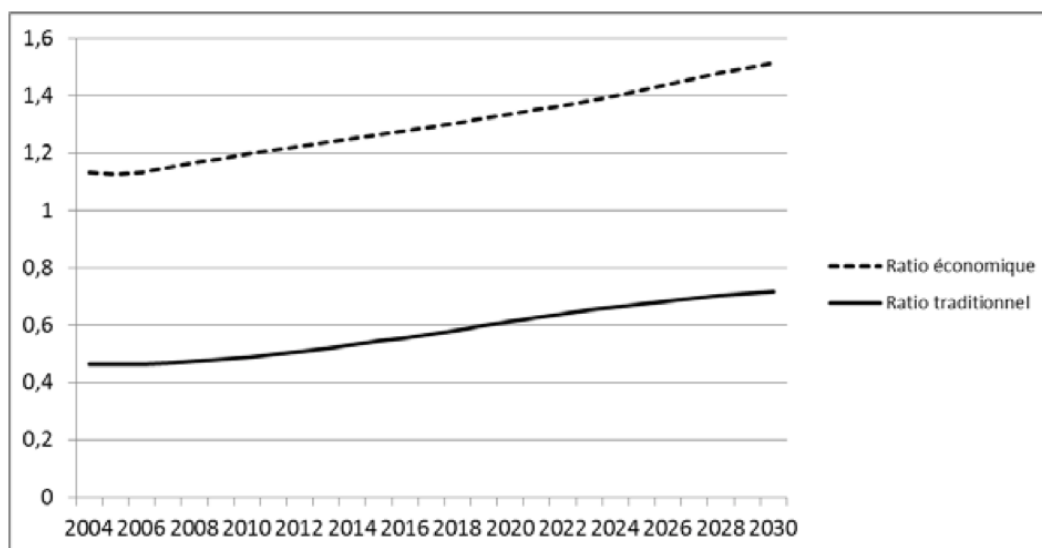


FIGURE 15 – Ratios de dépendance

Ce ratio a cependant une faiblesse importante. En approximant grossièrement le nombre de travailleurs et de non-travailleurs par l'âge, on ne peut voir l'effet de modifications de l'offre de travail sur les taux de dépendance économique. La Figure présente donc aussi un ratio alternatif : le « ratio de dépendance économique », qui divise le nombre de non-travailleurs (incluant les enfants, les retraités, les chômeurs et les inactifs) par le nombre de travailleurs. Notons que ce ratio est systématiquement plus élevé que le ratio traditionnel, car ce dernier suppose implicitement que tous les individus en âge de travailler travaillent. Le ratio suit la même tendance que le ratio précédant. L'interprétation est cependant différente : la valeur d'environ 1,2 en 2010 signifie que l'on compte 12 non-travailleurs pour chaque groupe de 10 travailleurs dans la population. En 2030, selon les prévisions de SimUL, on comptera plutôt 15 non-travailleurs pour chaque groupe de 10 travailleurs. Le changement est important, mais proportionnellement moins fort (1,5/1,2 relativement à 0,7/0,5) que pour le ratio traditionnel.

Le module du marché du travail modélise également les revenus de travail. La Figure 16 présente les revenus annuels de travail à l'aide de diagrammes de quartiles. La ligne du centre représente la médiane, les extrémités de la boîte montrent les 25^{ème} et 75^{ème} centiles, et les extrémités du diagramme indiquent les quasi minimum et maximum. Les valeurs externes, qui sont considérées statistiquement comme des valeurs aberrantes, ne sont pas incluses, car les valeurs extrêmes rendraient l'échelle du graphique inconmode. On voit dans la

base initiale que les distributions de revenus augmentent progressivement avec l'âge jusqu'à environ 30 ans, se stabilisent, puis diminuent à partir d'environ 52 ans. Cette dernière diminution peut avoir deux causes. Premièrement, bien que le début de la cinquantaine soit relativement jeune pour prendre sa retraite, il se peut que les individus dans cette tranche d'âge diminuent leurs activités en travaillant à temps partiel, ou qu'ils envisagent de travailler de manière moins intensive. Deuxièmement, en raison de la croissance de la productivité du travail, on s'attend à ce que les générations les plus vieilles soient moins riches que les plus jeunes.

Les prévisions pour les années futures montrent une tendance à la hausse dans l'ensemble des distributions, mais cette tendance est peu visible sur les graphiques. C'est pourquoi la Figure 17 présente également les prévisions des différences de revenus moyens par rapport aux prévisions de 2004. On voit que, de 2004 à 2030, les revenus moyens augmentent généralement de plusieurs milliers de dollars pour la population en âge de travailler²⁰.

Comme nous n'incluons pas d'effets de cohorte dans l'estimation des revenus de travaux (pour l'instant), l'augmentation observée des revenus provient forcément de l'évolution des variables démographiques. Il est entre autres attendu que l'augmentation continue dans le niveau d'éducation (voir la Figure 9) entraînera une augmentation continue des salaires.

La Figure 18 présente l'évolution des revenus moyens de travail pour les individus de 16 à 64 ans. La croissance des revenus de travail au début de la simulation est légèrement négative. Encore une fois, ceci est probablement dû à l'augmentation du nombre d'individus qui prennent leur retraite avant 65 ans. La croissance des revenus surpasse cet effet à partir de 2008, ce qui fait que le taux de croissance des revenus de travail varie entre 0 et 1% pour les années suivantes.

Cette dernière figure dissimule cependant plusieurs effets, car les revenus de travail moyens de différentes classes d'âges peuvent évoluer de manière différente. Le Tableau 3 présente les prévisions de revenus moyens par classe d'âges. La première classe d'âges, les 16 à 64 ans, comprend une proportion importante d'étudiants. On note que le revenu moyen de cette classe (4785\$) est significativement différent de celui présenté pour cette même classe dans les statistiques descriptives (9680\$). La différence est que, comme nous ne permettons pas aux étudiants de travailler dans SimUL, nous fixons ici les revenus des étudiants à zéro pour mieux comparer les prévisions avec la base initiale. On remarque aussi, toujours pour cette classe d'âge, la différence majeure entre la moyenne de la base initiale et de la prévision pour la fin de 2004. Ceci est probablement causé par les différences entre la BDMSPS, qui sert de base initiale, et l'EDTR, qui sert à modéliser les revenus futures (voir la section 5). Les prévisions futures pour cette classe d'âges, qui montrent une diminution continue du revenu moyen, sont probablement causées par le fait que les jeunes étudient de plus en plus longtemps dans le modèle.

Les individus de 25 à 55 ans sont principalement en âge de travailler. SimUL prévoit que les revenus d'emploi de ces individus augmenteront considérablement, passant d'environ 30 000\$ au début de 2004 à plus de 39 000\$ en 2030. Le saut entre la base initiale et la première prévision est encore une fois important. Les personnes âgées entre 55 et 64 ans

20. Le graphique montre une exception flagrante pour les individus de 53 ans. Le revenu moyen trop élevé pour cet âge en 2004 est probablement dû au nombre trop faible d'individus échantillonnés dans l'EDTR pour cette tranche d'âge. Nous comptons remédier éventuellement à ce genre de problème en fusionnant davantage de panels de l'EDTR aux données que nous utilisons présentement.

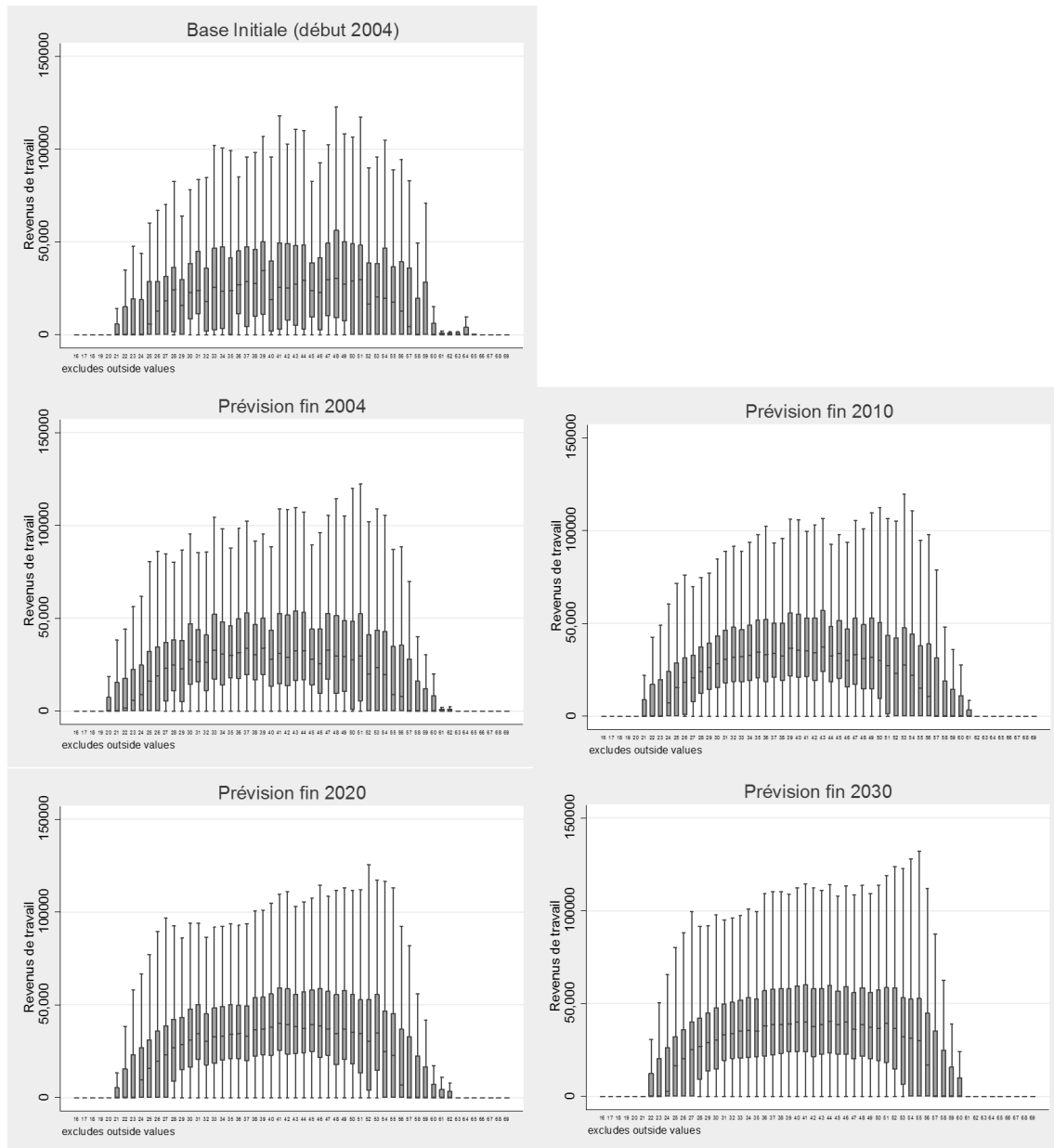


FIGURE 16 – Revenus de travail par âge

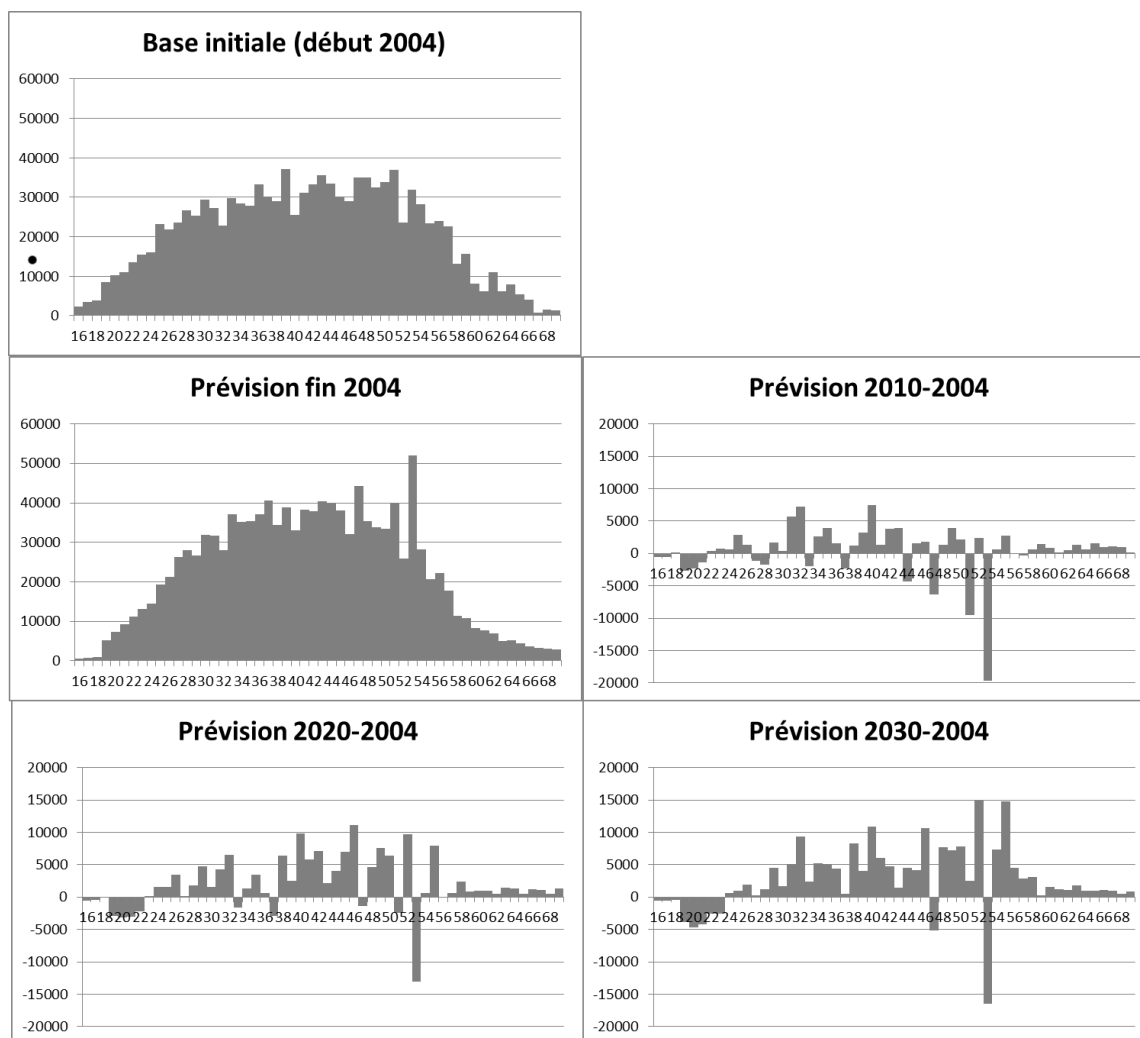


FIGURE 17 – Revenus de travail par âge

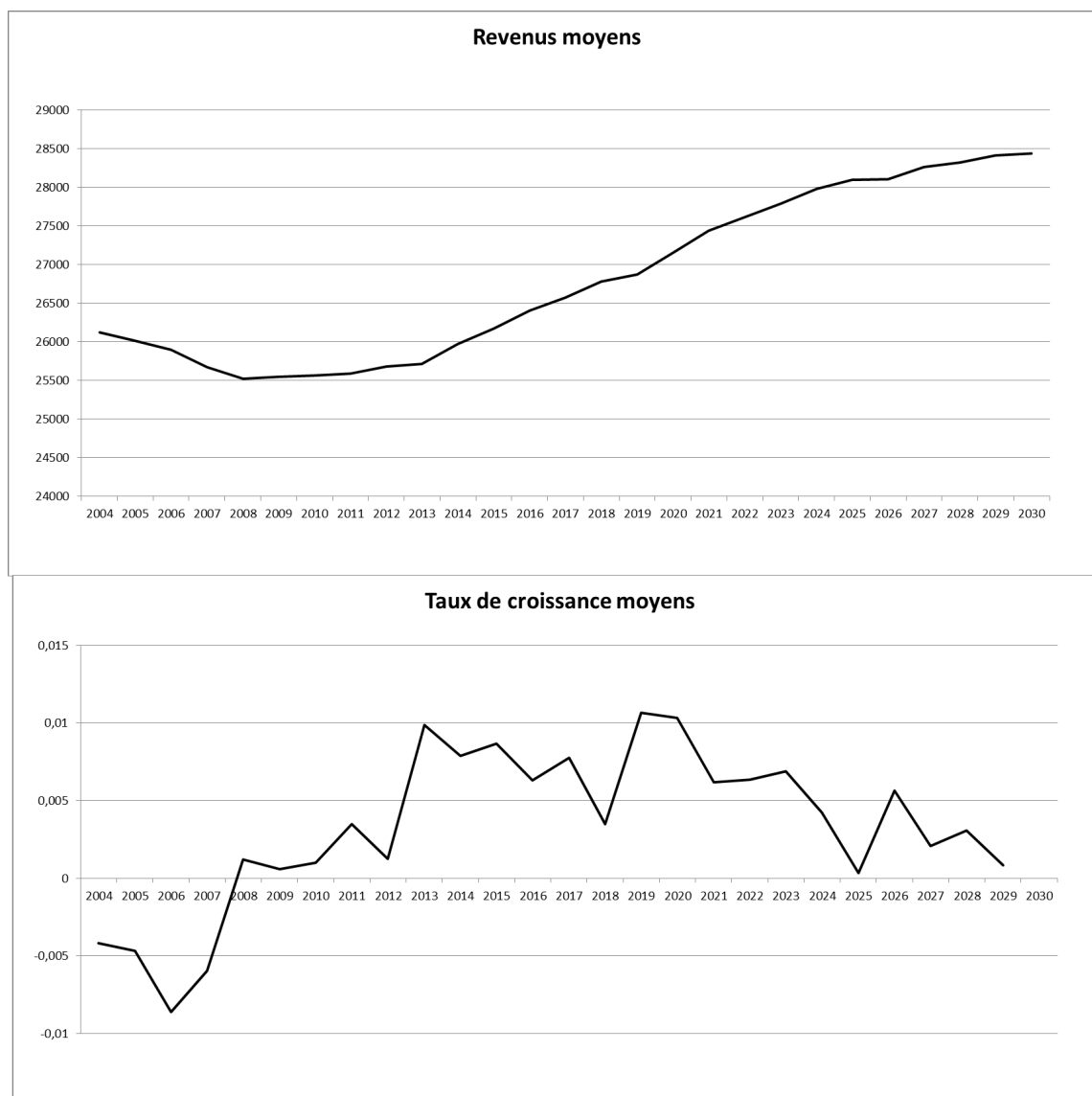


FIGURE 18 – Évolution des revenus de travail moyens pour les 16 à 64 ans

TABLE 3 – Prévisions des revenus de travail moyens par classes d'âges

	16 - 24 ans	25 - 54 ans	55 - 64 ans	65 ans et plus
Base initiale	4785	29961	14618	1326
Fin 2004	7015	34554	12479	1111
Fin 2010	5468	34492	12956	1309
Fin 2020	5960	37269	13390	1435
Fin 2030	4807	39110	14687	1228

sont partagées entre travailleurs et retraités. Sauf pour ce qui concerne le saut entre la base initiale et la première prévision, on constate une augmentation progressive du revenu moyen pour cette classe d'âges. Les prévisions des revenus de travail moyens pour les 65 ans et plus augmentent jusqu'en 2020, puis diminuent en 2030. Il est cependant embêtant de tirer des conclusions pour la population en âge de prendre sa retraite, car les variations de revenus moyens peuvent être causées par deux effets. L'augmentation de la productivité ou de l'éducation peut influencer ces revenus à la hausse, mais la décision des individus de prendre leur retraite plus tôt influencera les revenus moyens à la baisse.

Nous pouvons ensuite considérer les revenus de placements pour les 50 ans et plus. SimUL permet aux individus de recevoir des revenus de placements à partir de 16 ans, mais ces revenus, observés ou prédits, sont négligeables pour les moins de 50 ans. La Figure 19 présente les résultats. Les valeurs hors diagramme sont encore une fois exclues pour les mêmes raisons que pour la Figure précédente. Les médianes de tous les graphiques sont toujours très près de zéro, mais augmentent légèrement avec l'âge. Le modèle prédit peu de changements majeurs au fil du temps.

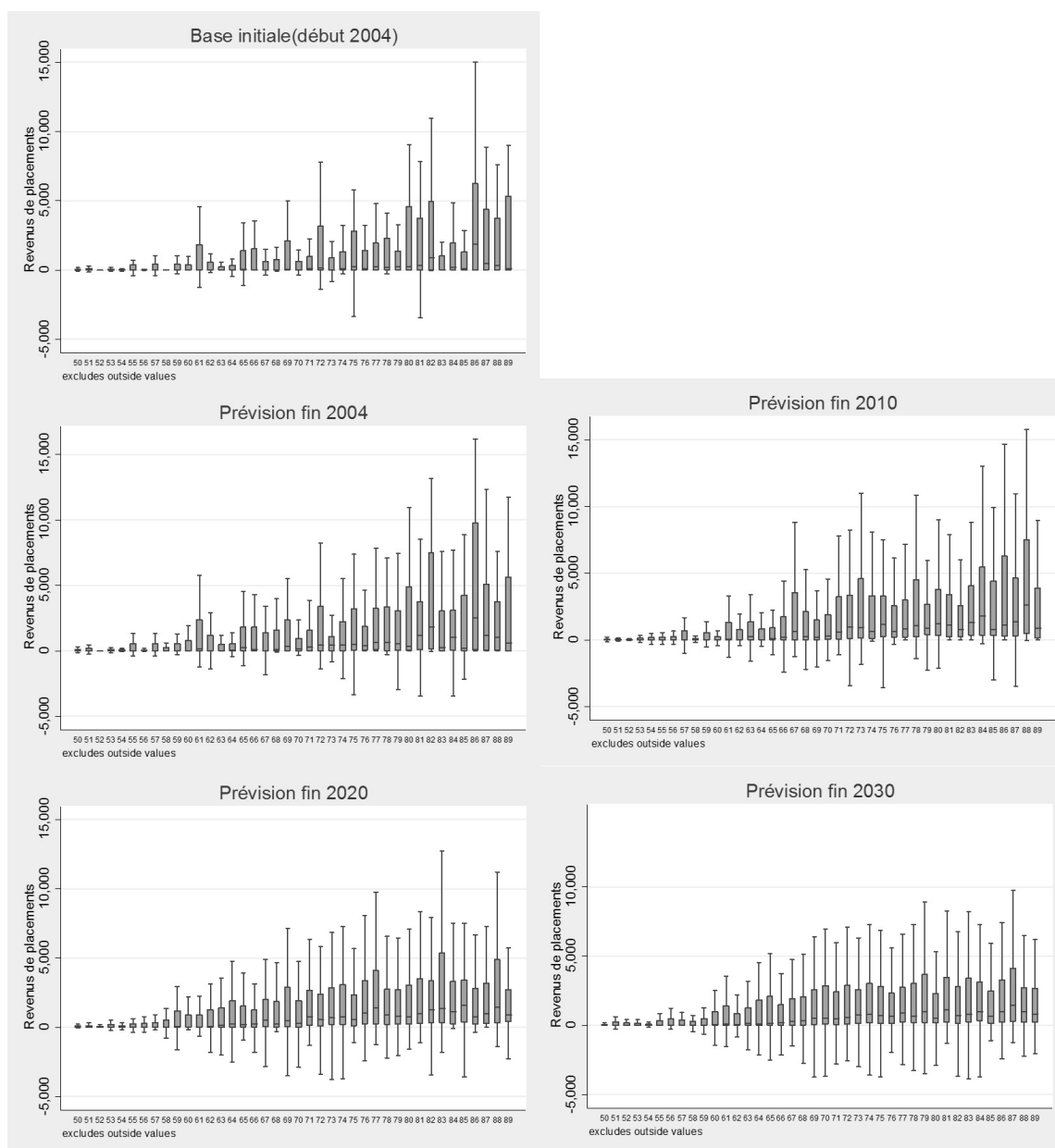


FIGURE 19 – Revenus de placements par âge

8 Simulations supplémentaires

8.1 Effet de l'augmentation du niveau de scolarité sur le revenu moyen

Comme SimUL n'inclut pas pour l'instant d'augmentation exogène de la productivité, ni d'effets de cohortes sur les revenus de travail, il semble raisonnable d'attribuer une partie de l'augmentation du revenu de travail moyen à l'augmentation continue du niveau d'éducation. Il est possible de tester cette hypothèse à l'aide d'une technique relativement simple : en modifiant les poids des individus en 2030 de façon à reproduire la distribution observée dans les niveaux de diplômes en 2004. On peut ensuite comparer les revenus de travail en 2030 avant et après pondération, la différence entre les deux étant attribuable à l'effet des différences dans les niveaux d'éducation.

Comme la distribution observée des diplômes selon l'âge est volatile (voir la Figure 8), nous choisissons de lisser les niveaux de diplômes selon l'âge à l'aide d'une estimation par noyau (kernel density estimation). On utilise le noyau d'Epanechnikov et la taille de fenêtre (bandwidth) choisie est de 2. L'estimation de la proportion des diplômés d'un certain niveau pour un âge donné prend donc en compte la proportion observée pour cet âge, mais aussi celle observée pour les individus un an plus vieux et un an plus jeunes.

TABLE 4 – Revenus de travail moyens en 2030 (incluant les revenus nuls) selon la distribution des diplômes

	16 - 24 ans	25 - 54 ans	55 - 64 ans	65 ans et plus
2030 avant pondération	4807	39110	14687	1228
2030 avec éducation 2004	5146	33208	11485	1100
Différence	-339	5902	3202	128

Le Tableau 4 présente les résultats. Les revenus de travail moyens avant pondération sont les même que ceux présentés plus haut. Ceux de 2030 après pondération représentent les revenus moyens que l'on observerait en 2030 dans un Québec hypothétique où le niveau d'éducation par âge n'aurait pas augmenté depuis 2004. La différence représente donc l'augmentation du revenu moyen attribuable à l'augmentation du niveau d'éducation depuis 2004. Cette différence est négative pour les individus de 16 à 24 ans (effet négatif de l'éducation sur le revenu moyen), car le prolongement de la scolarité fait que les jeunes entrent sur le marché du travail plus tard. Les individus de 25 à 54 ans ont majoritairement terminé leurs études et très peu sont déjà à la retraite. Cette classe d'âge permet donc de voir plus clairement l'effet de l'éducation sur le revenu moyen pour les individus qui travaillent. L'effet est fort, soit une augmentation de 5902\$ due à l'éducation. L'augmentation est plus modeste pour les 55 à 64 ans, mais quand même importante. Pour les 65 ans et plus, l'effet est très peu important, car ils sont pour la plupart à la retraite dans les deux situations.

8.2 Effet de l'augmentation de l'âge sur le revenu moyen

Une procédure semblable peut être utilisée pour analyser l'effet de l'évolution d'autres variables. La distribution de l'âge au Québec variera considérablement d'ici 2030 (voir la Figure 6). Nous vérifions comment l'évolution de la distribution d'âge affectera le revenu moyen des québécois selon SimUL. Nous modifions les poids des individus en 2030 de façon à reproduire les distributions d'âge observées en 2004, conditionnellement à un niveau de diplômentation donné. (La distribution des diplômes de 2030 est donc maintenue.) Le Tableau 5 présente les résultats de la simulation sur tous les revenus moyens (incluant les revenus nuls) pour chaque niveau de diplôme. On voit que la différence totale de revenu moyen est négative, ce qui indique que l'évolution de la distribution de l'âge a comme effet de réduire le revenu de travail moyen. Comme les moyennes incluent les revenus nuls, cet effet s'explique possiblement par la prise de retraite plutôt que par la réduction des salaires des travailleurs avec l'âge. Le Tableau 6 présente la même simulation, mais seulement pour les travailleurs de 2030. On voit ici que l'évolution de la distribution de l'âge affecte positivement le revenu moyen des travailleurs pour un diplôme donné.

TABLE 5 – Revenus de travail moyens en 2030 (incluant les revenus nuls) selon la distribution de l'âge

	Aucun	DES	DEC	BAC	Total
2030 avant pondération	11911	12739	22079	41874	25964
2030 âge 2004	11605	16027	26661	43413	27685
Différence	306	-3288	-4582	-1539	-1721

TABLE 6 – Revenus de travail moyens en 2030 (pour les travailleurs seulement) selon la distribution de l'âge

	Aucun	DES	DEC	BAC	Total
2030 avant pondération	29220	28978	35483	54919	42657
2030 âge 2004	27326	27858	34089	53137	41171
Différence	1894	1120	1394	1782	1486

8.3 Effets de l'immigration

L'immigration affectera significativement la distribution de l'âge dans la population de 2004 à 2030 (voir la Figure 6). SimUL permet d'observer l'effet de l'immigration sur le rapport de dépendance en simulant l'évolution du Québec en absence d'immigration. Comme les cibles exogènes de fécondité et d'immigration sont ajustées en fonction de l'immigration prévue, on ne peut simplement désactiver le module d'immigration de SimUL. On relance plutôt la simulation en supposant que les enfants des immigrants sont également des immigrants, puis on supprime les immigrants des données à la fin de la simulation. La distribution de l'âge de la population se trouvant dès lors modifiée, on observera forcément un effet sur le ratio de dépendance démographique.

La Figure 20 présente cet effet. Rappelons que le ratio de dépendance est calculé en divisant les « dépendants » (les 0 à 15 ans et les 65 ans et plus) par les individus en âge de travailler (les 16 à 64 ans). On voit que les immigrants ont eu un effet relativement faible sur le rapport de dépendance à ce jour. Cependant, cet effet augmente avec le temps de sorte que le rapport de dépendance en 2030 serait plus élevé d'environ 10 points de pourcentage dans un Québec sans immigration.

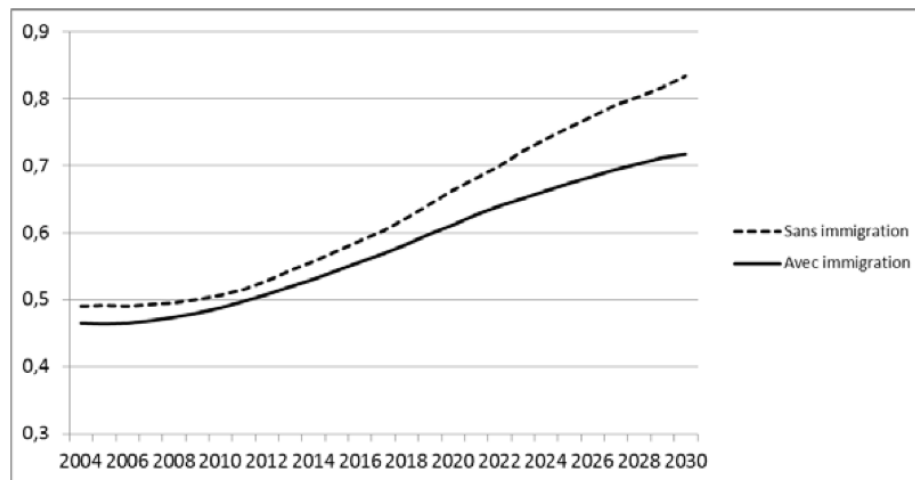


FIGURE 20 – Ratios de dépendance démographique avec et sans immigration

9 Conclusion

Le modèle SimUL permet d'étudier l'effet des changements socio-économiques (dont les changements démographiques) en procédant à la prévision et à la projection de distributions d'indicateurs socio-économiques sur la base d'échantillons représentatifs de la population et d'une riche analyse de micro-simulation dynamique. SimUL peut ainsi modéliser l'évolution de plusieurs variables affectées par le vieillissement de la population.

En tenant compte de l'évolution prévisible de la natalité, la mortalité, l'immigration et l'émigration, le modèle prévoit que la croissance de la population deviendra de plus en plus faible avec le temps, et que le ratio de dépendance passera de moins de 0,5 en 2010 à plus de 0,7 en 2030. Une évolution importante dans le temps du niveau de scolarité de la population québécoise est aussi à prévoir ; elle proviendra à la fois du vieillissement naturel des jeunes diplômés actuels (qui remplaceront les cohortes de travailleurs plus âgés moins scolarisés) et de la hausse anticipée de l'obtention de diplômes par les nouvelles cohortes.

On prévoit que les revenus de travail augmenteront significativement à travers les temps. Pour les individus de 25 à 54 ans, on prévoit que, de 2010 à 2030, le revenu de travail moyen (incluant les revenus nuls) passera de 34 492\$ à 39 110\$ en dollars constants de 2004, soit une augmentation d'environ 13%. Comme le modèle n'inclut pas (pour l'instant) d'augmentation exogène de la productivité ni d'effets de cohortes sur les revenus de travail, l'augmentation prédite de ces revenus pour la population en âge de travailler provient entièrement de l'évolution des caractéristiques socio-économiques de la population. Une de ces caractéristiques importantes est le niveau de scolarité. Le modèle attribue à l'augmentation du niveau d'éducation un effet d'environ 5 900\$ sur les revenus de travail des individus de 25 à 54 ans.

SimUL prévoit également une augmentation du taux d'emploi des individus dans la cinquantaine, une diminution du taux d'emploi des jeunes, car ceux-ci étudieront plus longtemps, ainsi qu'une augmentation dans le nombre d'heures travaillées des femmes.

SimUL n'est encore qu'à un stade assez préliminaire et un certain nombre de résultats doivent être validés et précisés. Il faudra en outre en accroître l'utilité en y ajoutant des modules permettant par exemple de procéder à des prédictions quant aux impôts prélevés, à l'accumulation des actifs (immobiliers, financiers, *etc*), aux dépenses en santé et en éducation et aux taxes à la consommation.

L'utilité de SimUL à des fins de politique publique pourra aussi être accrue en mettant en place des modèles structurels de comportements pour certains modules. SimUL n'utilise pour l'instant que des modèles de forme réduite qui évaluent la probabilité d'un individu de vivre un événement en fonction des valeurs de variables explicatives, sans modélisation structurelle des comportements. La validité de ces modèles de forme réduite risque d'être dépendante des politiques publiques et des circonstances en vigueur au moment où les données comportementales ont été observées. Si les politiques ou les conditions économiques changent, rien ne dit que les paramètres des modèles de forme réduite estimés sur des données antérieures seront toujours valides.²¹ Enfin, dans le cas de certains programmes

21. Il s'agit de la Critique de Lucas (1976) pour laquelle celui-ci a obtenu le Prix Nobel d'économie. Notons que la presque totalité des modèles de micro-simulation dynamique utilisent des modèles de forme réduite, une exception étant le modèle Suédois SESIM, qui utilise un modèle structurel pour la décision de prendre sa retraite (Flood et al., 2005).

sociaux (*e.g.*, RRQ), il importerait de générer les montants de cotisations et de prestations non pas à partir de modèles de forme réduite mais à partir des paramètres mêmes permettant de les calculer.

Références

- Agence de Revenu du Canada (2011). <http://www.cra-arc.gc.ca>.
- Bouffard, N., Easther, R., Johnson, T., Morrison, R. J., and Vink, J. (2001). Matchmaker, matchmaker, make me a match. *Brazilian Electronic Journal of Economics*, 4(2).
- Cassells, R., Harding, A., and Kelly, S. (2006). Problems and Prospects for Dynamic Microsimulation : A Review and Lessons for APPSIM. *NATSEM Woriking Paper Series*, no.63. 50 p.
- Chen, J., Sitter, R. R., and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89 :230 – 237.
- Favreault, M. and Smith, K. (2004). A Primer on the Dynamic Simulation of Income Model (DYNASIM3). Discussion Paper 02-04, The Urban Institute. 22 p.
- Flood, L., Jansson, F., Pettersson, T., Pettersson, T., Sundberg, O., and Westerberg, A. (2005). SESIM III : A Swedish dynamic micro simulation model. Handbook of sesim, Ministry of Finance, Stocholm.
- Fredriksen, D. (1998a). Projections of Population, Education, Labour Supply and Public Pension Benefits : Analyses with the Dynamic Microsimulation Model MOSART. *Social and Economic Studies* 101, 123 p.
- Fredriksen, D. (1998b). Projections of Population, Education, Labour Supply and Public Pension Benefits : Analyses with the Dynamic Microsimulation Model MOSART. Technical report, Statistics Norway.
- Gupta, A. and Kapur, V. (2000). *Microsimulation in Government Policy and Forecasting*. North-Holland Elsevier Science, Amsterdam.
- Harding, A. (1993). *Lifetime Income Distribution and Redistribution : Application of a Microsimulation Model*. Contributions to Economic Analysis 221, Amsterdam, North Holland. 436 p.
- Harding, A., Kelly, S., Percival, R., and Keegan, M. (2009). Population Ageing and Government Age Pension Outlays : Using Microsimulation Models to Inform Policy Making. Technical report, NATSEM, Canberra.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1) :153–61.
- Imhoff, E. and Post, W. (1997). Méthodes de micro-simulation pour des projections de population. *Population*, 52 :889–932.
- ISQ (2009). Perspectives démographiques du Québec et des régions, 2006-2056. Document de travail, ISQ, Québec. 132 p.
- Kelly, S. (2003). Autralias’s Microsimulation Model - DYNAMOD. Discussion paper, NATSEM, Canberra.

- Kelly, S. and Percival, R. (2009). Longitudinal benchmarking and alignment of a dynamic microsimulation model. Discussion paper, NATSEM, Canberra. 120 p.
- Lucas, R. (1976). Econometric policy evaluation : A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1 :19–46.
- Meyerson, N., Pineles-Mark, C., Schwabish, J. A., Simpson, M., and Topoleski, J. (2009). CBO’s Long-Term Model : An Overview. Background paper, Congressional Budget Office, Washington D.C.. 16 p.
- Milligan, K. (2010). Canadian tax and credit simulator. Database, software and documentation, Version 2010-1.
- Morrison, R. (2007). DYNACAN (Longitudinal Dynamic Microsimulation Model). In *Harding, A., Modelling Our Future : Population Ageing, Health and Aged Care*, chapter Modèle 6. Emerald. p. 461-465.
- Morrison, R. and Dussault, B. (2000a). Overview of DYNACAN : A full-fledged Canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes. Discussion paper, Actuarial International Association. 29 p.
- Morrison, R. and Dussault, B. (2000b). Overview of dynacan : A full-fledged Canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes. Technical report, Association actuarielle internationale.
- Moussaly, K. (2011). Participation aux régimes privés d’épargne-retraite. Technical report, Statistique Canada.
- Ressources humaines et développement des compétences Canada (2011). Modifications touchant le régime de pensions du canada. [http ://www.rhdcc.gc.ca/fra/sv-rpc/changements.shtml](http://www.rhdcc.gc.ca/fra/sv-rpc/changements.shtml).
- Revenu Québec (2011). [http ://www.revenuquebec.ca](http://www.revenuquebec.ca).
- Robert-Bobée, L. (2001). Modelling demographic behaviours in the French microsimulation model Destinie : An analysis of future change in completed fertility. Série des documents de travail de la Direction des Études et Synthèses Économiques 14, INSEE. 36 p.
- Service Canada (2011). [http ://www.servicecanada.gc.ca](http://www.servicecanada.gc.ca).
- Smith, K., Favreault, M., Ratcliffe, C., Butrica, B., Toder, E., and Bakija, J. (2007). FINAL REPORT : Modeling Income in the Near Term 5. Discussion paper, The Urban Institute, Washington D.C.. 135 p.
- Spielauer, M. (2007). Dynamic Microsimulation of Health Care Demand, Health Care Finance and the Economic Impact of Health Behaviours : Survey and Review. *International Journal of Microsimulation*, 1 (1) :35–53.
- Statistique Canada (2009). Le modèle de microsimulation LifePaths : Une vue d’ensemble. 37 p. www.statcan.gc.ca/microsimulation/lifepaths/lifepaths-fra.htm.

- Wu, C. (2004). Some Algorithmic Aspects of the Empirical Likelihood Method in Survey Sampling. *Statistica Sinica*, 14 :1057 – 1067.
- Zaidi, A. and Rake, R. (2001). Dynamic Microsimulation Models : A Review and Some Lessons for SAGE. SAGE Discussion Paper 2, ESRC-Sage Research Group. 36 p.

A Construction des bases de données

A.1 Données initiales

Comme il a été mentionné dans la section 5, les données brutes servant de point de départ à SimUL proviennent de la BDMSPS. Il est cependant utile d'effectuer un traitement préliminaire des données avant de les importer dans le modèle. Tout d'abord, de nouveaux poids échantillonnaires ont été estimés et appliqués aux observations de la BDMSPS. Ceci a été fait pour mieux refléter le nombre total de personnes par genre et par tranche d'âge (1 an) dans la population. En effet, certaines différences entre les totaux de la BDMSPS et ceux estimés par l'ISQ ont été remarquées. Les Tableaux 7 à 10 en font la présentation.

TABLE 7 – Description de la population par tranche d'âge (0-25) en 2004

Âge	Homme			Femme		
	Nb. pers. ISQ	Différence initiale	Différence finale	Nb. pers. ISQ	Différence initiale	Différence finale
0	38 060	2 995	-2,2	36 074	7 015	-2,9
1	37 394	557	-2,8	35 677	567	-3,9
2	38 001	1 777	-3,7	36 351	5 402	-2,5
3	38 293	4 781	-1,5	35 904	-623	-1,3
4	39 267	7 497	-0,2	37 095	-9 574	1,0
5	39 265	-3 173	-2,0	37 157	-20 399	-0,3
6	40 773	2 668	-1,6	38 853	3 275	-4,8
7	43 764	3 907	-1,6	41 046	687	-1,0
8	45 334	1 013	-0,6	43 695	1 815	-1,5
9	46 646	2 180	9,3	44 864	-10 477	-1,4
10	48 057	-2 727	-1,3	45 423	6 072	0,3
11	49 471	12 141	-2,0	45 423	9 417	-0,4
12	50 982	-11 668	-2,9	48 743	-7 981	-2,1
13	52 048	-3 919	-1,9	49 659	2 378	-1,8
14	50 921	6 036	1,0	48 448	-1 946	-1,5
15	47 782	-3 393	-1,9	45 274	10 000	-0,7
16	44 848	-4 964	-1,1	43 055	7 914	-2,3
17	44 860	-6 474	-2,1	43 405	-6 576	-2,3
18	47 595	5 321	-0,8	45 190	3 328	-0,6
19	49 353	5 956	-3,1	46 415	-9 393	0,2
20	50 232	7 191	0,9	47 327	-80	0,3
21	50 079	-14 525	-0,1	47 746	13 464	0,4
22	52 326	-1 382	1,2	49 554	2 203	0,4
23	54 707	7 220	1,0	52 180	4 809	1,2
24	55 254	8 746	1,7	51 971	-3 712	2,5
25	52 931	-13 542	1,5	50 905	-7 768	1,6

TABLE 8 – Description de la population par tranche d'âge (26-50) en 2004

Âge	Homme			Femme		
	Nb. pers. ISQ	Différence initiale	Différence finale	Nb. pers. ISQ	Différence initiale	Différence finale
26	52 314	-1 728	1,6	49 903	-6 952	2,9
27	52 241	1 703	3,0	50 321	3 316	2,5
28	51 478	5 774	2,4	48 394	-4 585	1,8
29	51 348	5 635	4,5	48 554	2 005	3,7
30	48 788	-1 275	5,4	46 523	2 295	3,6
31	48 297	6 781	2,1	45 582	5 670	4,6
32	48 582	-1 521	2,4	46 420	6 371	4,0
33	51 213	-2 561	1,4	48 553	-8 045	3,2
34	52 448	-3 418	5,9	49 427	37	5,4
35	51 960	-7 767	3,0	49 498	-15 785	3,8
36	52 301	6 811	2,1	50 492	-6 543	3,0
37	54 980	8 397	5,0	52 608	7 705	3,9
38	59 057	-11 986	1,5	56 272	-2 520	3,9
39	64 465	5 071	1,6	62 116	6 267	4,0
40	65 981	-1 260	3,1	63 110	-99	7,3
41	66 672	-1 199	1,5	64 785	1 947	6,9
42	65 549	-436	-1,5	63 419	687	4,5
43	66 494	-1 677	0,4	65 706	1 026	2,9
44	65 533	13 862	0,8	65 489	20 198	5,3
45	64 951	5 557	2,7	65 017	-2 805	-0,4
46	64 965	5 534	2,9	64 651	964	6,3
47	63 858	-482	0,8	63 969	-10 694	3,5
48	60 650	-9 389	2,3	61 449	1 874	2,2
49	60 938	460	-1,1	62 123	-205	0,7
50	58 827	-4 741	3,6	60 129	15 661	0,7

TABLE 9 – Description de la population par tranche d'âge (51-75) en 2004

Âge	Homme			Femme		
	Nb. pers. ISQ	Différence initiale	Différence finale	Nb. pers. ISQ	Différence initiale	Différence finale
51	55 899	-1 209	1,5	57 142	-10 351	4,2
52	54 390	-6 223	7,6	55 461	13 806	0,9
53	53 694	1 158	1,8	55 293	-5 197	1,7
54	52 083	4 769	2,0	53 544	-9 578	1,7
55	51 457	13 991	0,9	53 161	-906	1,3
56	50 002	-4 175	1,3	51 664	-1 161	1,0
57	49 947	10 359	1,4	51 488	4 830	-0,9
58	46 288	-5 863	-1,3	48 249	-3 449	1,1
59	44 111	1 340	2,8	46 152	-6 359	0,2
60	41 999	-1 966	2,5	43 987	1 645	1,6
61	40 196	-11 361	2,6	41 900	9 201	-1,5
62	37 433	9 191	2,5	39 268	1 793	4,0
63	35 765	-9 154	-4,8	37 369	-3 874	5,9
64	32 343	-981	0,6	34 570	-5 491	4,1
65	31 015	-7 895	6,1	33 104	-2 541	8,5
66	28 931	-4 514	0,7	31 275	-3 672	7,7
67	27 029	-30	0,2	29 966	4 299	7,8
68	26 980	1 185	-2,0	30 267	-9 863	10,4
69	25 977	5 865	-2,1	29 767	-11 319	12,1
70	24 896	3 602	-3,9	28 849	-4 295	20,6
71	25 135	-2 858	-10,6	29 178	677	7,8
72	24 791	3 110	-6,5	29 466	8 495	11,3
73	22 906	-4 772	-16,1	28 700	-5 268	12,6
74	21 892	1 842	-16,8	27 976	-2 202	21,6
75	19 995	-4 386	-8,5	26 223	-3 125	17,3

TABLE 10 – Description de la population par tranche d'âge (76-99) en 2004

Âge	Homme			Femme		
	Nb. pers. ISQ	Différence initiale	Différence finale	Nb. pers. ISQ	Différence initiale	Différence finale
76	19 051	-3 440	-15,3	25 677	-1 852	27,0
77	17 242	-3 368	-0,3	24 255	-3 808	15,4
78	16 251	-1 703	-20,7	23 632	-2 881	25,1
79	14 922	2 478	-39,3	22 448	-5 497	25,2
80	13 156	1 349	-57,1	21 169	7500	-12,2
81	11 637	1 167	-5,9	19 585	2 974	41,2
82	10 544	3 894	-39,1	18 311	6 512	40,6
83	8 889	898	-127,1	16 484	1 451	6,2
84	7 324	-1 152	48,3	14 419	-3 790	99,3
85	5 964	1 743	-132,7	12 406	-557	25,4
86	4 987	-49	-18,7	11 183	3 216	-6,7
87	3 939	-651	203,9	9 712	-1 880	103,6
88	3 334	-43	-43,5	8 688	1 104	-34,9
89	2 820	-292	48,6	4 129	3 390	-102,0
90	2 200	-1 379	590,4	4 607	1 704	-159,8
91	1 741	964	-387,1	8 472	-3 328	588,0
92	1 253	1 253	1 253,0	4 174	59	34,4
93	872	-7,1	300,4	3 161	1 798	-661,4
94	694	-299	203,7	2 480	-328	84,5
95	436	-198	89,9	1 872	369	-179,1
96	327	327	327,0	1 331	936	-751,5
97	209	209	209,0	950	-943	580,9
98	118	118	118,0	649	-1 149	736,6
99	82	82	82,0	441	441	441,0

Ensuite, certaines variables nominales ont été fractionnées en variables indicatrices pour refléter des états précis :

TABLE 11 – Description du fractionnement des variables

Variable initiale	Variable fractionnée
Niveau de scolarité	Indicatrice de l'atteinte d'un diplôme secondaire. Indicatrice de l'atteinte d'un diplôme de niveau collégial. Indicatrice de l'atteinte d'un diplôme de niveau universitaire.
Statut marital	Indicatrice de personne mariée. Indicatrice de personne en union de fait.
Statut d'études	Indicatrice de fréquentation scolaire.

Puis, l'imputation de variables temporelles a été effectuée, en prévision du calcul des probabilités de transition des phénomènes démographiques. Celles-ci sont présentées dans le Tableau 12. Ce type de variables n'est pas disponible dans la BDMSPPS, puisque cette base de données est en coupe transversale. Pour effectuer l'imputation, la méthode de l'appariement de données²² est utilisée en se basant sur les données publiques de l'ESG de 2001 et 2006.

TABLE 12 – Description des variables temporelles importées

Nom	Description
Temps union	Temps depuis le début de l'union.
Temps célibataire	Temps depuis le début de la période de célibat.
Épisode de conception	Temps depuis le début de l'union sans avoir donné naissance à un enfant ou temps depuis la naissance du dernier enfant durant l'union actuelle.
Temps hors études	Temps depuis la fin des études.
Ind. relation antérieure	Variable indicatrice de l'existence d'une relation antérieure.
Temps départ maison	Temps depuis le départ du foyer familial.

Lors de la comparaison des proportions de réception de prestations de la RRQ entre la BDMSPPS et les données panel de l'EDTR, d'importantes divergences ont été découvertes pour les femmes. Comme l'EDTR est la source de données sur lesquelles les prévisions sont extrapolées, il a été décidé de calibrer les proportions de réception de prestation de la RRQ des femmes sur celles de l'EDTR. Le Tableau 13 illustre les différences de proportions par groupe d'âge.

22. Voir B.2 pour davantage de détails sur la technique d'appariement.

TABLE 13 – Description de la proportion de femmes recevant des prestations de RRQ par groupe d'âge

Âge	Proportions avant ajus.	Proportions après ajus.
60-64	0,81	0,63
65-69	0,90	0,64
70-74	0,87	0,67
75-79	0,95	0,77
80-84	0,94	0,84
85-89	0,92	0,85
90-94	0,85	0,83
95-99	1,00	1,00

Par la suite, on procède au clonage de certaines observations de manière à obtenir un poids échantillonnal maximal égal à 50. Cette procédure est utilisée dans le but d'accroître les possibilités de parcours de vie d'une même observation et de diminuer la variance des poids. Au terme du traitement des données de la BDMSPS pour le fichier des particuliers, celui-ci est composé des variables suivantes :

TABLE 14 – Description des variables du fichier famille

Nom	Description
Ids	Identifiant statique.
Poids	Poids échantillonal.
Age	Âge.
Sexe	Sexe.
Idfam	Identifiant statique de la famille.
TempsUnion	Temps depuis le début de l’union.
TempsFinEtudes	Temps depuis la fin des etudes.
IndEtudes	Variable indicatrice de la frequentation scolaire.
IndRel	Variable indicatrice de relation conjugale.
IndMar	Variable indicatrice de personne mariée.
IndRelPrec	Variable indicatrice de l’existence d’une relation antérieure.
IndMarPrec	Variable indicatrice de l’existence d’un mariage précédent.
NbEnfants	Nombre d’enfants engendrés dans le passé.
TempsNaissance	Temps depuis le début de l’union sans avoir donné naissance à un enfant ou temps depuis la naissance du dernier enfant durant l’union actuelle.
IndImmigrant	Variable indicatrice du Statut d’immigrant.
LienChef	Lien avec chef de famille.
NbAnneesScol	Nombre d’année de scolarité.
TempsCelibat	Temps depuis début du célibat.

TABLE 15 – Description des variables du fichier famille (suite)

Nom	Description
TempsDepMais	Temps depuis le départ du foyer familial.
TempsNiveauScol	Temps passé dans le niveau de scolarité actuel.
IndEmploi	Variable indicatrice de la participation au marché du travail.
NbHeuresTrav	Nombre d’heures travaillées.
Salaire	Salaire horaire.
RevTravail	Valeur du revenu de travail annuel.
AideSociale	Valeur de l’aide-sociale.
GainCap	Valeur du gain de capital.
RevPlac	Valeur du revenu de placement.
RevAutre	Valeur des autres revenus.
RevPen	Valeur du revenu de pension privée.
RevReer	Valeur des retraits de RÉER.
RevRrq	Valeur des prestations de la RRQ.
RevSvSrg	Valeur de la sécurité de la vieillesse et du supplément de revenu garanti

Pour terminer, le fichier des particuliers est réorganisé en fichier des familles. Les fichiers des particuliers et des familles servent d’intrant principal au modèle SimUL. Dans ce dernier type de fichier, une observation équivaut à une famille nucléaire et des variables propres à la famille sont créées :

TABLE 16 – Description des variables du fichier famille

Nom	Description
NbPersonnes	Nombre de personnes dans la famille.
IndRelChef	Statut marital du chef de la famille.
SexeChef	Sexe du chef de la famille.
NbEnfantsMen	Nombre d’enfants vivant avec la famille.
NbAdultes	Nombre d’adultes dans la famille.
AgeEnf1-AgeEnf8	Âge de l’enfant 1 à 8. L’ordre des enfants est du plus jeune au plus vieux.

A.2 Données des immigrants internationaux

La simulation de l’immigration internationale a nécessité la création d’une banque de données d’immigrants potentiels. Pour ce faire, il a été choisi de sélectionner les immigrants étant arrivés au Canada²³ durant les cinq années précédant le Recensement de 2001. Ce critère de sélection a été retenu, car pour une telle durée, il est plus probable que la structure familiale des immigrants soit restée identique à celle au moment de leur arrivée. Deux types

23. Même si on réfère à l’arrivée des immigrants au Canada, seuls les immigrants vivant au Québec ont été inclus dans la banque de données.

de traitement ont dû être employés selon la structure familiale des immigrants (les familles²⁴ et les immigrants hors famille de recensement). Pour les familles immigrantes, le traitement des données a été plus complexe. En effet, dans le fichier des particuliers du Recensement, on retrouve des informations précises sur l'âge et l'année d'immigration. Cependant, il est impossible de relier les personnes entre elles et de connaître leur structure familiale. Pour le fichier des familles, les informations sur la structure familiale sont présentes, mais c'est au niveau de l'âge des adultes et des enfants, en plus de l'année d'immigration, que les données ont été agrégées. Comme il est important de disposer de l'âge précis des individus dans SimUL, il est primordial d'effectuer une désagrégation des catégories d'âge et de durée depuis l'immigration. Pour ce faire, la technique de l'appariement²⁵ est employée en se basant sur le fichier des particuliers du Recensement de 2001. Pour le cas des immigrants hors famille de recensement, cette complication est superflue, puisque par définition cette « famille » ne contient qu'une seule personne. Il est donc possible d'utiliser directement le fichier des particuliers pour ce type de ménage.

Une fois que l'âge et le moment précis de l'immigration sont obtenus, l'extrapolation des caractéristiques des immigrants à leur arrivée au Canada est faite, en soustrayant à leur âge, le temps depuis leur immigration. Les enfants qui se retrouvent avec un âge négatif sont éliminés de la famille. Par la suite, un traitement relativement semblable²⁶ à celui de la base de données initiale est effectué, de manière à obtenir une structure de fichiers des particuliers et des familles identique à la base initiale. Finalement, pour ce qui est des variables économiques des immigrants, elles sont mises égales à zéro et elles sont entièrement déterminées par les équations de SimUL.

B Annexe technique

B.1 Ajustement des poids échantillonnaux

$N_{ISQ} \rightarrow$ Taille de la population totale selon l'ISQ

$N \rightarrow$ Taille de la population totale selon la BDMSPS

$n \rightarrow$ Taille de l'échantillon de la BDMSPS

$w_i \rightarrow$ Poids échantillonnal de la famille i dans la BDMSPS

$s_i \rightarrow$ Taille de la famille i

Premièrement, les w_i sont ajustés de manière à ce que $N = \sum_{i=1}^n s_i w_i = N_{ISQ}$.

Puis, l'échantillon est séparé en 200 groupes (représentés par $g = 1, \dots, 200$) comprenant 100 groupes d'âge (de 0 à 99 ans) par sexe. Pour chaque famille i , un vecteur colonne (200×1) x_i est dérivé, dans lequel chaque élément représente le nombre d'individus dans la famille appartenant à un groupe correspondant. Supposons que q est un vecteur colonne (200×1) de la proportion de la population dans chaque groupe selon l'ISQ.

La fonction de vraisemblance peut être exprimée comme suit :

24. Pour être incluse dans la base de données des immigrants, tous les adultes d'une famille doivent être immigrants.

25. Voir B.2 pour davantage de détails sur la technique d'appariement.

26. Les traitements appliqués sont le fractionnement des variables nominales, l'imputation des variables temporelles et la réorganisation des données en fichier de familles.

$$L(p) = \sum_{i=1}^n w_i \log p_i \quad (2)$$

Pour calibrer les poids échantillonnaires de manière à ce que la taille de la population des 200 groupes soit la même pour l'ISQ et la BSPS, il est nécessaire de maximiser l'équation (2) sous les contraintes suivantes :

$$\sum_{i=1}^n p_i = 1 \quad (3)$$

$$\sum_{i=1}^n p_i x_i - q = 0 \quad (4)$$

Il est possible de considérer q_i pour chaque famille i plutôt que q , lorsque $q_i = q$ et que $\sum_{i=1}^n p_i q_i = q$.

La contrainte (4) peut être alors réécrite comme suit :

$$\sum_{i=1}^n p_i u_i = 0, \quad (5)$$

où $u_i = (x_i - q_i)$.

Le Lagrangien est le suivant :

$$\mathcal{L} = \sum_{i=1}^n w_i \log p_i - \mu \left(\sum_{i=1}^n p_i - 1 \right) - \lambda' \left(\sum_{i=1}^n p_i u_i \right) \quad (6)$$

La condition de premier ordre donne :

$$p_i = \frac{w_i}{\sum_{i=1}^n w_i + \lambda' u_i} \quad (7)$$

Le remplacement de p_i par son expression dans la deuxième contrainte donne :

$$f(\lambda) = \sum_{i=1}^n \frac{w_i u_i}{\sum_{i=1}^n w_i + \lambda' u_i} = 0 \quad (8)$$

Puisque $p_i > 0$, λ est tel que $\sum_{i=1}^n w_i + \lambda' u_i > 0$. La solution peut être trouvée en maximisant la nouvelle fonction de vraisemblance :

$$\tilde{L}(\lambda) = \sum_{i=1}^n w_i \log(\sum_{i=1}^n w_i + \lambda' u_i) \quad (9)$$

selon λ et sous la contrainte que :

$$\sum_{i=1}^n w_i + \lambda' u_i > 0. \quad (10)$$

Tel que mentionné par Wu 2004, lorsqu'une solution unique existe, il est possible de la dériver en utilisant l'algorithme Newton-Rapson, suggéré par Chen et al. 2002. Les étapes à suivre sont les suivantes :

Étape 0 :

Poser $j = 0$, $\lambda_0 = 0$, $\gamma_0 = 1$ and $\epsilon = 10^{-8}$.

Étape 1 :

Calculer $\Delta(\lambda_j)$ comme suit :

$$\Delta(\lambda) = \left\{ \frac{\partial}{\partial \lambda} f(\lambda) \right\}^{-1} f(\lambda) = \left\{ - \sum_{i=1}^n \frac{w_i u_i u_i'}{(\sum_{i=1}^n w_i + \lambda' u_i)^2} \right\}^{-1} \sum_{i=1}^n \frac{w_i u_i}{\sum_{i=1}^n w_i + \lambda' u_i}$$

Si $\|\Delta(\lambda_j)\| < \epsilon$, arrêter l'algorithme et afficher λ_j ; sinon passer à l'étape 2.

Étape 2 :

Calculer $\delta_j = \gamma_j \Delta(\lambda_j)$. Si $\sum_{i=1}^n w_i + (\lambda_j - \delta_j)' u_i \leq 0$ ou $\tilde{L}(\lambda_j - \delta_j) < \tilde{L}(\lambda_j)$, alors poser $\delta_j = 0.5 \delta_j$ et répéter l'étape 2 ; sinon passer à l'étape 3.

Étape 3 :

Poser $\lambda_{j+1} = \lambda_j - \delta_j$, $j = j + 1$ et $\gamma_{j+1} = (j + 1)^{-0.5}$. Retourner à l'étape 1.

B.2 Méthode d'appariement

La méthode d'appariement a pour but de transférer la valeur d'un ensemble de variables ($Y = y_1, \dots, y_p$) d'une base de données donneuse (BD_d) à une receveuse (BD_r). Pour ce faire, des variables communes à la BD_d et BD_r sont identifiées ($X = x_1, \dots, x_n$). Puis, différents identifiants ($Id_1, \dots, Id_s$) sont créés. Ceux-ci se différencient par le nombre de x qui les composent. Ainsi, l'identifiant le moins précis (Id_1) est celui qui contient les variables de critère minimal de sélection ($X_m = x_1, \dots, x_m$). De nouvelles variables sont incluses

dans les identifiants suivants (Id_j), jusqu'à ce que Id_s contienne l'ensemble des X . Par la suite, la procédure d'appariement consiste à transférer Y d'une observation donneuse à une observation receveuse à l'aide des identifiants des deux bases de données (Id_{jd} et Id_{jr}) :

Étape 0 :

On pose $j = s$.

Étape 1 :

On recherche une correspondance entre Id_{jd} et Id_{jr} . Si cette opération est fructueuse on passe à l'étape 2. Sinon on passe à l'étape 3.

Étape 2 :

Si une seule observation donneuse est trouvée pour une observation receveuse, Y_i est directement transféré à l'observation receveuse. Cependant, lorsque plusieurs observations donneuses sont identifiées, on détermine aléatoirement pour laquelle le transfert de Y_i aura lieu.

Étape 3 : On diminue le niveau de précision de Id_{jd} et Id_{jr} en fixant $j = \max(s - l, 1)$, où l est égal au nombre de répétitions de la procédure d'appariement. Puis, on retourne à l'étape 1.

Les identifiants Id_{1d} et Id_{1r} sont construits de manière à assurer une correspondance entre les observations donneuses et receveuses.

B.3 Dérivation du risque de sortie pour le logit

Il est possible de partitionner le temps en une séquence infinie de périodes de temps disjointes, $l = (\tau_{l-1}, \tau_l]$, où $l = 1, 2, \dots$ et $\tau_0 = 0$. Pour chaque période, le temps est observé en un seul point (τ_l). Ainsi, le risque de sortie à chaque période est défini par :

$$h(l, x) = \Pr(T \in l | T > \tau_{l-1}, \mathbf{x}). \quad (11)$$

De plus, il est possible de définir la fonction de survie conditionnelle par

$$S(l, \mathbf{x}) = \Pr(T > \tau_l | \mathbf{x}) = \prod_{j=1}^l [1 - h(j, \mathbf{x})] \quad (12)$$

et la fonction de densité conditionnelle par

$$f(l, \mathbf{x}) = h(l, \mathbf{x}) \prod_{k=1}^{l-1} (1 - h(k, \mathbf{x})). \quad (13)$$

En considérant la censure aléatoire non-informative, la contribution d'une observation non-censurée est $\Pr(T_i \in l) = f(l, \mathbf{x}_i)$, alors que la contribution des observations censurées est $\Pr(T_i > l) = S(l, \mathbf{x}_i)$. Ainsi, la vraisemblance peut être écrite comme suit :

$$\begin{aligned} L &= \prod_{i=1}^N \Pr(T_i \in l_i)^{d_i} \Pr(T_i > l_i)^{1-d_i} \\ &= \prod_{i=1}^N \left[h(l_i, \mathbf{x}_i) \prod_{k=1}^{l_i-1} (1 - h(k, \mathbf{x}_i)) \right]^{d_i} \left[\prod_{k=1}^{l_i} (1 - h(k, \mathbf{x}_i)) \right]^{1-d_i} \end{aligned} \quad (14)$$

Le logarithme de la fonction de vraisemblance est

$$\begin{aligned} \ell &= \sum_{i=1}^N \left[d_i \log(h(l_i, \mathbf{x}_i)) + d_i \sum_{k=1}^{l_i-1} \log(1 - h(k, \mathbf{x}_i)) + (1 - d_i) \sum_{k=1}^{l_i} \log(1 - h(k, \mathbf{x}_i)) \right] \\ &= \sum_{i=1}^N \left[d_i \log \frac{(h(l_i, \mathbf{x}_i))}{(1 - h(l_i, \mathbf{x}_i))} + \sum_{k=1}^{l_i} \log(1 - h(k, \mathbf{x}_i)) \right] \end{aligned} \quad (15)$$

En définissant l'indicateur $d_{il} = I(T_i \in l)$, ayant une valeur égale à 1 lorsque la période l est celle où l'événement se produit et ayant une valeur égale à 0 autrement ; Il est possible de réécrire la log-vraisemblance comme suit :

$$\begin{aligned} &= \sum_{i=1}^N \left[\sum_{k=1}^{l_i} d_{ik} \log \frac{(h(k, \mathbf{x}_i))}{(1 - h(k, \mathbf{x}_i))} + \sum_{k=1}^{l_i} \log(1 - h(k, \mathbf{x}_i)) \right] \\ &= \sum_{i=1}^N \sum_{k=1}^{l_i} d_{ik} \log(h(k, \mathbf{x}_i)) + (1 - d_{ik}) \log(1 - h(k, \mathbf{x}_i)). \end{aligned} \quad (16)$$

Enfin, un simple logit peut être effectué en créant de nouvelles observations pour chaque période de temps jusqu'à ce qu'un individu subissent l'événement ou qu'il soit censuré.

C Les variables explicatives des modèles

C.1 Le modèle de départ de ménage

- log de l'âge
- log de l'âge au carré
- log de l'âge au cube
- log du temps depuis la fin des études
- log du temps depuis la fin des études au carré
- log du temps depuis la fin des études au cube
- log du temps depuis la fin des études à la puissance quatre
- log de l'âge $\times$ log du temps depuis la fin des études
- indicateur d'études en cours
- indicateur d'individu sur le marché du travail
- indicateur du statut d'immigrant

- indicateur du statut d’immigrant pour le père
- indicateur du statut d’immigrant pour la mère
- indicateur d’un ménage monoparental
- indicateur d’un seul enfant engendré
- indicateur de deux enfants engendrés
- indicateur de plus de deux enfants engendrés
- indicateur d’un diplôme d’études secondaires (DES) pour le père
- indicateur d’un diplôme d’études collégiales (DEC) pour le père
- indicateur d’un diplôme d’études universitaires (BAC ou plus) pour le père
- indicateur d’un DES pour la mère
- indicateur d’un DEC pour la mère
- indicateur d’un BAC ou plus pour la mère
- indicateur d’un DES
- indicateur d’un DEC
- indicateur d’un BAC ou plus
- indicateur de l’âge 23 ans
- indicateur de l’âge 24 ans
- indicateur de l’âge 25 ans
- nombre de frères et soeurs plus jeunes
- constante

Toutes ces variables sont aussi incluses en interaction avec une variable indicatrice pour le sexe féminin, à l’exception des variables qui concernent le nombre d’enfants engendrés et de la constante.

C.2 Le modèle de formation de couple

Dans le cas d’une première union, les variables explicatives sont les suivantes :

- log de l’âge
- log de l’âge au carré
- log de l’âge au cube
- log du temps depuis le départ du ménage
- log du temps depuis le départ du ménage au carré
- log du temps depuis le départ du ménage au cube
- le temps depuis la fin des études
- le temps depuis la fin des études au carré
- le temps depuis la fin des études au cube
- indicateur d’un seul enfant engendré
- indicateur de deux enfants engendrés
- indicateur de plus de deux enfants engendrés
- indicateur d’un enfant de moins de 6 ans
- indicateur d’un DES
- indicateur d’un DEC
- indicateur d’un BAC ou plus
- indicateur d’études en cours
- constante

Quand il ne s'agit pas de la première union de l'individu, les variables sont plutôt les suivantes :

- log de l'âge
- log de l'âge au carré
- log de l'âge au cube
- log du temps depuis le départ du ménage
- log du temps depuis le départ du ménage au carré
- log du temps depuis le départ du ménage au cube
- le temps depuis la fin de la dernière relation
- le temps depuis la fin de la dernière relation au carré
- le temps depuis la fin de la dernière relation au cube
- le temps depuis la fin de la dernière relation multiplié par l'âge
- le temps depuis la fin de la dernière relation au carré multiplié par l'âge
- le temps depuis la fin de la dernière relation au cube multiplié par l'âge
- indicateur d'un seul enfant engendré
- indicateur de deux enfants engendrés
- indicateur de plus de deux enfants engendrés
- indicateur d'un enfant de moins de 6 ans
- indicateur d'un DES
- indicateur d'un DEC
- indicateur d'un BAC ou plus
- indicateur d'études en cours
- constante

C.3 Le modèle de dissolution de couple

- log de l'âge
- log de l'âge au carré
- log de l'âge au cube
- log du temps depuis le début de l'union
- log du temps depuis le début de l'union au carré
- log du temps depuis le début de l'union au cube
- log du temps depuis le début de l'union multiplié par l'âge
- log du temps depuis le début de l'union au carré multiplié par l'âge
- log du temps depuis le début de l'union au cube multiplié par l'âge
- indicateur d'un seul enfant engendré
- indicateur de deux enfants engendrés
- indicateur de plus de deux enfants engendrés
- indicateur d'un enfant de moins de 6 ans
- indicateur d'une union antérieure
- indicateur d'un DES
- indicateur d'un DEC
- indicateur d'un BAC ou plus
- indicateur d'études en cours
- constante

C.4 Le modèle de naissance

- log de l'âge
- log de l'âge au carré
- log de l'âge au cube
- log du temps depuis le début de l'union
- log du temps depuis le début de l'union au carré
- log du temps depuis le début de l'union au cube
- log du temps depuis le dernier enfant engendré
- log du temps depuis le dernier enfant engendré au carré
- log du temps depuis le dernier enfant engendré au cube
- log du temps depuis la fin des études
- log du temps depuis la fin des études au carré
- log du temps depuis la fin des études au cube
- indicateur d'un union précédente
- indicateur d'un DES
- indicateur d'un DEC
- indicateur d'un BAC ou plus
- constante

Pour les femmes qui n'ont jamais eu d'enfant, on utilise les mêmes variables, à l'exception de celles qui concernent le temps écoulé depuis le dernier enfant.

C.5 Le modèle de scolarité

Si l'individu est toujours chez ses parents, les variables explicatives sont les suivantes :

- log de l'âge
- log de l'âge au carré
- log de l'âge au cube
- log du temps depuis le début du niveau d'études actuel
- log du temps depuis le début du niveau d'études actuel au carré
- log du temps depuis le début du niveau d'études actuel au cube
- log du temps depuis le début du niveau d'études actuel à la puissance quatre
- log du temps depuis le début du niveau d'études actuel x l'âge
- indicateur du statut d'immigrant
- indicateur du statut d'immigrant pour le père
- indicateur du statut d'immigrant pour la mère
- indicateur d'un ménage monoparental
- indicateur de sexe féminin
- indicateur d'un seul enfant engendré
- indicateur de deux enfants engendrés
- indicateur d'un diplôme d'études secondaires (DES) pour le père
- indicateur d'un diplôme d'études collégiales (DEC) pour le père
- indicateur d'un diplôme d'études universitaires (BAC ou plus) pour le père
- indicateur d'un DES pour la mère
- indicateur d'un DEC pour la mère
- indicateur d'un BAC ou plus pour la mère

- indicateur d'un DES
- indicateur d'un DEC
- indicateur d'un BAC ou plus
- Indicateur de relation conjugale
- indicateur d'un diplôme d'études secondaires (DES) pour le père x indicateur de sexe féminin
- indicateur d'un diplôme d'études collégiales (DEC) pour le père x indicateur de sexe féminin
- indicateur d'un diplôme d'études universitaires (BAC ou plus) pour le père x indicateur de sexe féminin
- indicateur d'un DES pour la mère x indicateur de sexe féminin
- indicateur d'un DEC pour la mère x indicateur de sexe féminin
- indicateur d'un BAC ou plus pour la mère x indicateur de sexe féminin
- indicateur d'un DES x indicateur de sexe féminin
- indicateur d'un DEC x indicateur de sexe féminin
- indicateur d'un BAC ou plus x indicateur de sexe féminin
- Indicateur de relation conjugale x indicateur de sexe féminin
- cohorte (Année de naissance)
- cohorte x log de l'âge
- cohorte x log de l'âge au carré
- constante

Si l'individu a fondé son propre ménage, on utilise les mêmes variables, sauf celles qui concernent les caractéristiques des parents.

C.6 Le modèle de marché du travail

- splines d'âges²⁷
 - plus de 16 ans
 - plus de 20 ans
 - plus de 25 ans
 - plus de 35 ans
 - plus de 45 ans
 - plus de 55 ans
 - plus de 60 ans
 - plus de 65 ans
 - plus de 55 ans x indicateur de diplôme collégial
 - plus de 60 ans x indicateur de diplôme collégial
 - plus de 65 ans x indicateur de diplôme collégial
 - plus de 55 ans x indicateur de diplôme universitaire
 - plus de 60 ans x indicateur de diplôme universitaire
 - plus de 65 ans x indicateur de diplôme universitaire

27. L'estimation par splines est une estimation linéaire, similaire aux moindres carrés ordinaires, où on permet au coefficient de varier selon la valeur de la variable. Dans ce cas-ci, on estime le coefficient associé à l'âge pour les 16 ans et plus, puis on permet au coefficient de s'ajuster après 20 ans, 25 ans, 35 ans, 45 ans, 55 ans, 60 ans et 65 ans.

- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur de relation conjugale
- indicateur d’un enfant de moins de 6 ans
- le nombre d’enfants de plus de 5 ans
- indicateur de sexe féminin
- constante

Toutes ces variables sont aussi incluses en interaction avec une variable indicatrice pour le sexe féminin, à l’exception des variables qui concernent le nombre d’enfants engendrés et de la constante.

C.7 Le modèle de RRQ (1 si l’individu en reçoit, 0 sinon)

- splines d’âges
 - plus de 60 ans
 - plus de 65 ans
 - plus de 70 ans
- splines d’âges ci-dessus x indicateur de sexe féminin
- cohorte (année de naissance) x indicateur de sexe féminin
- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur de relation conjugale
- indicateur du statut d’immigrant
- indicateur de sexe féminin
- indicateur de diplôme secondaire x indicateur de sexe féminin
- indicateur de diplôme collégial x indicateur de sexe féminin
- indicateur de diplôme universitaire x indicateur de sexe féminin
- constante

C.8 Le modèle de de prestation de RRQ)

- cohorte
- cohorte x indicateur de sexe féminin
- splines d’âges
 - plus de 60 ans
 - plus de 65 ans
 - plus de 70 ans
 - plus de 80 ans
- splines d’âges ci-dessus x indicateur de sexe féminin
- cohorte (année de naissance) x indicateur de sexe féminin
- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur de relation conjugale

- indicateur du statut d’immigrant
- indicateur de sexe féminin
- indicateur de diplôme secondaire x indicateur de sexe féminin
- indicateur de diplôme collégial x indicateur de sexe féminin
- indicateur de diplôme universitaire x indicateur de sexe féminin
- constante

C.9 Le modèle de pension privée (1 si l’individu en reçoit, 0 sinon)

Pour les moins de 65 ans, les variables sont :

- splines d’âges
 - plus de 50 ans
 - plus de 55 ans
 - plus de 60 ans
- splines d’âges ci-dessus x indicateur de sexe féminin
- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur de relation conjugale
- indicateur du statut d’immigrant
- indicateur de sexe féminin
- indicateur de prestations de RRQ
- indicateur d’emploi
- indicateur d’emploi x indicateur de sexe féminin
- indicateur de diplôme secondaire x indicateur de sexe féminin
- indicateur de diplôme collégial x indicateur de sexe féminin
- indicateur de diplôme universitaire x indicateur de sexe féminin
- indicateur de relation conjugale x indicateur de sexe féminin
- constante

Pour les plus de 65 ans, les variables sont :

- splines d’âges
 - plus de 65 ans
 - plus de 70 ans
 - plus de 75 ans
 - plus de 80 ans
- splines d’âges ci-dessus x indicateur de sexe féminin
- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur de relation conjugale
- indicateur du statut d’immigrant
- indicateur du statut d’immigrant x indicateur de sexe féminin
- indicateur de sexe féminin
- indicateur d’emploi
- indicateur d’emploi x indicateur de sexe féminin

- indicateur de diplôme secondaire x indicateur de sexe féminin
- indicateur de diplôme collégial x indicateur de sexe féminin
- indicateur de diplôme universitaire x indicateur de sexe féminin
- indicateur de relation conjugale x indicateur de sexe féminin
- constante

C.10 Le modèle de niveau de pension privée

Pour les moins de 65 ans, les variables sont :

- splines d'âges
 - plus de 50 ans
 - plus de 55 ans
 - plus de 60 ans
- splines d'âges ci-dessus x indicateur de sexe féminin
- cohorte
- cohorte (année de naissance) x indicateur de sexe féminin
- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur de relation conjugale
- indicateur du statut d'immigrant
- indicateur de sexe féminin
- indicateur de prestations de RRQ
- indicateur de diplôme secondaire x indicateur de sexe féminin
- indicateur de diplôme collégial x indicateur de sexe féminin
- indicateur de diplôme universitaire x indicateur de sexe féminin
- indicateur de relation conjugale x indicateur de sexe féminin
- constante

Pour les plus de 65 ans, les variables sont :

- splines d'âges
 - plus de 65 ans
 - plus de 70 ans
 - plus de 75 ans
 - plus de 80 ans
- splines d'âges ci-dessus x indicateur de sexe féminin
- cohorte
- cohorte (année de naissance) x indicateur de sexe féminin
- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur de relation conjugale
- indicateur du statut d'immigrant
- indicateur du statut d'immigrant x indicateur de sexe féminin
- indicateur de sexe féminin
- indicateur de diplôme secondaire x indicateur de sexe féminin

- indicateur de diplôme collégial x indicateur de sexe féminin
- indicateur de diplôme universitaire x indicateur de sexe féminin
- indicateur de relation conjugale x indicateur de sexe féminin
- constante

C.11 Le modèle de différence de pension privée

- revenu de pension de la période précédente
- splines d'âges
 - plus de 50 ans
 - plus de 55 ans
 - plus de 60 ans
- splines d'âges ci-dessus x indicateur de sexe féminin
- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur de relation conjugale
- indicateur de sexe féminin
- indicateur de prestations de RRQ
- indicateur de diplôme secondaire x indicateur de sexe féminin
- indicateur de diplôme collégial x indicateur de sexe féminin
- indicateur de diplôme universitaire x indicateur de sexe féminin
- indicateur de relation conjugale x indicateur de sexe féminin
- constante

Pour les plus de 65 ans, les variables sont :

- splines d'âges
 - plus de 65 ans
 - plus de 70 ans
 - plus de 75 ans
 - plus de 80 ans
- splines d'âges ci-dessus x indicateur de sexe féminin
- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur de relation conjugale
- indicateur du statut d'immigrant
- indicateur de sexe féminin
- indicateur de diplôme secondaire x indicateur de sexe féminin
- indicateur de diplôme collégial x indicateur de sexe féminin
- indicateur de diplôme universitaire x indicateur de sexe féminin
- indicateur de relation conjugale x indicateur de sexe féminin
- constante

C.12 Le modèle de revenus de placements

- splines d'âges

- plus de 16 ans
- plus de 25 ans
- plus de 35 ans
- plus de 45 ans
- plus de 55 ans
- plus de 60 ans
- plus de 65 ans
- plus de 70 ans
- plus de 80 ans
- splines d'âges ci-dessus x indicateur de sexe féminin
- indicateur de diplôme secondaire
- indicateur de diplôme collégial
- indicateur de diplôme universitaire
- indicateur d'emploi
- indicateur de relation conjugale
- indicateur du statut d'immigrant
- indicateur de diplôme secondaire x indicateur de sexe féminin
- indicateur de diplôme collégial x indicateur de sexe féminin
- indicateur de diplôme universitaire x indicateur de sexe féminin
- indicateur d'emploi x indicateur de sexe féminin
- indicateur de relation conjugale x indicateur de sexe féminin
- indicateur du statut d'immigrant x indicateur de sexe féminin
- indicateur d'un seul enfant dans le ménage
- indicateur de deux enfants dans le ménage
- indicateur de plus de deux enfants dans le ménage
- constante