Micro-simulation sur le cycle de vie : le modèle TaxIPP-Life

Béatrice Boutchenik et Alexis Eidelman
IPP

29 avril 2013

Introduction

L'objet du projet TaxIPP-Life, débuté en septembre 2012, est la réalisation d'études sur la population française portant sur l'ensemble du cycle de vie. Le regard que l'on porte sur les inégalités peut se voir modifié dans cette optique. Sur des données administratives suédoises, [1] montre que l'état de pauvreté sur le cycle de vie est temporaire alors que les personnes en haut de la distribution des revenus y restent. Les aller-retours de part et d'autres du seuil de pauvreté est aussi un phénomène connu en France. Se pose alors la question de la redistribution sur le cycle de vie. Cette question est en général étudiée sur une base annuelle. Si cette démarche est intéressante pour considérer l'état de la population à un instant donné, elle est souffre de plusieurs limitations. D'abord l'estimation de la pauvreté et de la richesse est établie à partir de la situation d'une année seulement et uniquement en fonction des revenus. Ainsi seront pratiquement toujours pauvres les étudiants et souvent le seront les retraités. Si l'on pense que la consommation peut être lissée au cours de la vie et que le revenu permettant d'apréhender cette consommation est le revenu permanent. Une étude annuel ne peut être satisfaisante. Par exemple, un retraité aura en général moins de revenu que lorsqu'il est actif, cependant son niveau de vie a probablement été surestimé lorsqu'il était actif car il a épargné en pensant à sa retraite et est sous-estimé car il peut consommer plus que ses revenus une fois inactif. La question du patrimoine est elle aussi primordiale dans l'étude des inégalités. Il est en effet raisonnable de croire qu'un individu ayant un capital (éventuellement un capital attendu à travers un héritage) n'aura pas la même attitude de consommation qu'une personne qui n'en a pas. Si on pense à deux étudiants, l'un aidé par ses parents, l'autre non, dans l'imaginaire collectif celui aidé par ses parents est plus aisé que l'autre. Pourtant, si pour subvenir à ses besoin le second travaille en même temps que ses études, il sera paradoxalement considéré comme plus riche.

Le premier point pour justifier une approche sur cycle de vie est donc que l'on peut mieux appréhender la situation des individus. Un deuxième est que si l'on veut étudier la progressivité des transferts ou du système socio-fiscal dans son ensemble, une étude en coupe souffre de limitation importante. En effet, les études en coupe doivent contourner la difficulté causée par les transferts dits assuranciels (assurance chômage, retraite et maladie). Ces assurances ne peuvent être vues uniquement comme des transferts. Celui qui cotise ouvre un droit pour plus tard, c'est le principe de l'assurance. Par sa cotisation, il s'assure un revenu futur ou du moins l'espérance

d'un revenu. Une étude de la redistribution sur cycle de vie permet d'intégrer ces transferts assuranciels dans l'étude alors qu'ils doivent exclus d'une étude en coupe. Pour ces transferts, comme pour les autres, on est capable d'identifier une redistribution individuel temporelle et une redistribution entre agent. On sépare ainsi la partie d'un prélèvement qu'un individu va « récupérer » ou qui lui a été avancée de la partie qui bénéficiera vraiment à d'autres. Pour les prestations, on peut aussi isoler ce qui relève purement de la solidarité de la partie de la prestation que l'individu a lui-même financée

L'étandue des questions auxquelles TaxIPP-Life est très grande. On ne les détaillera pas ici car l'état actuel du projet ne permet pas encore d'aborder ces points. Le présent document détaillera essentiellement la méthode utilisée et donnera de premiers résultats provisoires.

L'étude sur le cycle de vie exige, et c'est une lapalissade, des données sur l'ensemble de ce cycle de vie. Cela n'est pas sans poser de problème. Le plus évident est que lorsque l'on interroge un individu, il ne connait pas son avenir. Le modèle TaxIPP-Life contient un module réalisant une projection à un niveau individuel du futur. Ceci sera développé dans la deuxième partie de ce document. La première s'attardera sur un point moins évident qui est que le passé des individus n'est pas non plus si bien connus. En dehors des effets de mémoire qui font qu'une partie du passé est oubliée, aucune enquête ne demande à un individu l'ensemble de ses revenus passés. Des données administratives semblent être la seule solution pour connaitre par exemple la trajectoire professionnelle des individus. Mais les données administratives ne sont pas non plus une panacée, elles ne donnent en général pas d'information sur la vie matrimoniale des personnes, or, savoir qui vit avec qui est déterminant pour étudier la consommation et les niveaux de vie. Enfin, une troisième voie serait l'utilisation d'un panel qui suivrait les individus tout au long de leur vie, toutefois une telle base de données, avec des informations assez riches pour pouvoir appliquer la législation socio-fiscal n'existe pas. Nous avons donc procédé à un matching statistique de données d'enquête et de données administratives. La première partie de document présentera ce matching avec entre les données de l'enquête patrimoine et les données de l'échantillon EIR-EIC.

1 Construction de la base de données

1.1 Appariement statistique des données EIC-DADS avec l'Enquête Patrimoine

Un des intérêts majeurs de TAXIPP Life réside dans l'utilisation de trajectoires salariales effectivement observées, plutôt que simulées. Les données administratives des EIC et DADS ont été utilisées à cette fin. Celles-ci ont été couplées à l'enquête Patrimoine afin de disposer d'un certain nombre d'informations, notamment au niveau du ménage plutôt que de l'individu. Un appariement statistique a donc été effectué entre les deux sources de données : appariement statistique exact sur un certain nombre de variables considérées au moment de l'enquête (sexe, âge, PCS à 1 chiffre, tranche de revenus salariaux et de remplacement), puis appariement avec

le plus proche voisin, en particulier en termes de trajectoire vis-à-vis de l'emploi.

Il a été dans un premier temps nécessaire de manipuler les fichiers EIC-DADS (et UNEDIC) d'une part et l'enquête Patrimoine d'autre part afin de les rendre plus exactement comparables quant aux variables utilisées pour l'appariement statistique : il faut en effet que ces variables soient définies de la façon la plus proche possible dans une base et dans l'autre. En particulier, il a fallu utiliser l'information des fichiers EIC-DADS-UNEDIC afin de retracer les trajectoires vis-à-vis de l'emploi, selon les mêmes modalités que celles renseignées dans le calendrier rétrospectif de l'Enquête Patrimoine, calendrier renseignant les changements entre des situations ayant duré au moins un an.

1.1.1 Mise en forme des fichiers EIC-DADS-UNEDIC

On conserve uniquement les individus nés à partir de 1942, les données concernant les générations 1934 et 1938 n'étant pas complètes.

Reconstitution des trajectoires vis-à-vis de l'emploi Les observations du fichier DADS correspondent au croisement individu x année x entreprise. A partir de ce fichier et pour les années 1976 à 2001, on détermine le statut de l'individu au 1er janvier de chaque année, en termes d'emploi à temps plein ou temps partiel. Les données du fichier EIC caisse x individu x année permettent par ailleurs de croiser cette information avec le fait que le salarié était dans le public, le privé ou encore était indépendant pour l'année donnée. Ces sources permettent de retracer les modalités suivantes de la variable de statut vis-à-vis de l'emploi de l'Enquête Patrimoine : salarié du public à temps complet, salarié du public à temps partiel, salarié du privé à temps complet, salarié du privé à temps partiel et à son compte . Lorsqu'il y a cumul de plusieurs emplois au premier janvier de l'année, que ceux-ci soient à temps plein ou temps partiel, on considère que l'individu travaille à temps plein. Enfin, l'appartenance au public ou à la catégorie des indépendants étant obtenue dans l'EIC grâce à l'appartenance à différentes caisses - appartenance dont seule l'année est connue et non les dates exactes au cours de cette année -, on a 2116 individus qui pour au moins une année sont affiliés à la fois à une caisse du public et à une caisse d'indépendants. On attribue aléatoirement un des deux statuts à ces individus pour l'année donnée.

Le fichier UNEDIC permet quant à lui de retracer pour les années 1984 à 2001 et grâce au type d'allocation perçue le fait qu'au 1er janvier de chaque année l'individu soit en situation de chômage indemnisé ou non (chômage indemnisé seulement jusqu'en 1992), de formation et de préretraite. S'il y a cumul de plusieurs allocations dont une de préretraite ou de formation au 1er janvier de l'année considérée, on classe l'individu comme étant plutôt en préretraite ou formation. Lorsqu'on a à la fois un statut provenant du fichier DADS et un statut provenant du fichier UNEDIC pour le même individu et la même année, on conserve le statut DADS. L'information provenant du fichier EIC caisse x individu x année , complétée par les dates de sortie des fichiers EIC-DADS-UNEDIC, nous permet de retracer les départs à la retraite.

Enfin, le calendrier rétrospectif de l'enquête Patrimoine comprend une modalité "succession de courtes périodes (inférieures à un an) d'emploi et de chômage", qu'il serait dommage d'assimiler aléatoirement à de l'emploi ou du chômage, celle-ci nous renseignant sur un type de situation particulière vis-à-vis de l'emploi. De telles périodes ont ainsi été repérées dans le fichier DADS.

Revenus salariaux et de remplacement pour l'année 2001 On additionne les salaires nets renseignés dans les DADS et les salaires nets des fonctionnaires de l'Etat donnés dans la base EIC caisse x individu x année. On prend enfin en compte les revenus de remplacement, que l'on divise par 0,7 afin de comparer des salaires de référence et non des revenus effectifs, qui nous intéresseraient moins ici : il serait par exemple peu souhaitable d'apparier un individu gagnant 2000 euros mensuels avec un autre gagnant généralement 3000 euros environ, mais qui l'année considérée se trouve être en situation de chômage et donc le revenu de remplacement est donc plus proche de 2000 euros.

1.1.2 Mise en forme de l'Enquête Patrimoine

L'appariement se fera naturellement lorsque l'on disposera du fichier EIC-DADS 2010, mais on matche pour l'instant l'Enquête Patrimoine 2009-2010 et l'EIC 2001 en alignant la première sur l'EIC, et en considérant donc la situation d'un individu pour l'année 2010 dans l'Enquête Patrimoine comme correspondant à celle pour l'année 2001 de l'EIC. Ainsi, on utilise pour l'appariement selon les trajectoires des dates retardées de 9 années pour l'Enquête Patrimoine, ce qui revient à comparer des situations vis-à-vis de l'emploi pour un âge donné. On déflate par ailleurs les revenus salariaux et de remplacement de l'Enquête Patrimoine 2010 par la croissance nominale observée des salaires afin qu'ils soient comparables à ceux provenant des fichiers EIC-DADS-UNEDIC en 2001. De même que pour les données issues du fichier EIC, on multiplie les revenus de remplacement par un facteur de 0.7 avant de les additionner aux revenus salariaux.

Les données EIC étant disponibles pour une génération sur quatre, on regroupe également les individus de l'Enquête Patrimoine en catégories correspondant à quatre années de naissance : les individus de l'Enquête Patrimoine nés entre 1941 et 1944 seront par exemple appariés avec la génération née en 1942 du fichier EIC, ceux nés entre 1945 et 1948 avec la génération 1946, et ainsi de suite. L'échantillon est donc composé d'individus nés entre 1941 et 1972.

On conserve pour l'appariement statistique uniquement les individus qui au moment de l'enquête ne sont pas agriculteurs ou indépendants, et qui dans leur trajectoire rétrospective de l'Enquête Patrimoine ont déclaré au moins une période comme salarié (à temps plein ou temps partiel, dans le public ou le privé), comme au chômage ou en alternance de courtes périodes d'emploi et de chômage, entre 1984 et 2001. Pour ceux qui n'ont pas connu de tels épisodes sur cette période, on matche séparément ceux qui en ont connu au moins une sur la période 1976 à 1983. Les autres ne sont donc pas appariés et se verront attribuer des revenus salariaux nuls entre 1976 et 2001, ce qui est cohérent leur déclaration dans l'enquête Patrimoine. On dissocie

en effet 1976-1983 et 1984-2001 car la variable de statut vis-à-vis de l'emploi ne peut être définie de la même manière sur ces deux périodes, les données UNEDIC n'étant pas renseignées avant 1984.

Enfin, on n'apparie pas pour l'instant les individus à la retraite [1].

1.1.3 Appariement des deux bases

L'appariement est effectué de façon exacte sur les variables de sexe, d'âge par tranches de quatre ans, de PCS à un chiffre et de revenus salariaux et de remplacement par tranches au nombre de 12, entre les données de l'année 2001 dans le fichier EIC-DADS et celles de l'année 2010 pour l'Enquête Patrimoine (ces dernières étant ré-évaluées pour les revenus salariaux et de remplacement). Au sein d'une cellule d'individus possédant les mêmes caractéristiques, on recherche alors un plus proche voisin quant à la trajectoire en termes de statut entre 1984 et 2001, sauf pour les individus n'ayant pas eu de période salariée ou de chômage entre 1984 et 2001, et aux revenus salariaux et de remplacement.

La distance entre trajectoires est calculée en spécifiant des coûts de suppression d'un statut - il est ainsi coûteux de passer d'une trajectoire à la trajectoire identique, avec simplement une période en moins ou en plus - et de substitution d'un statut à un autre. Le coût de suppression est le même quel que soit le statut pendant la période supprimée, mais les coûts de substitution varient selon le statut que l'on substitue à un autre statut donné : on doit ainsi définir une matrice de coûts (cf. infra). On n'autorise pas les statuts manquants au cours ou à la fin d'une trajectoire : une période manquante est catégorisée comme "Inactif et autres". On autorise par contre les statuts manquants au début des trajectoires, c'est-à-dire avant l'entrée dans une des bases de données, si celle-ci a lieu après 1976. Une trajectoire considérée est ainsi plus courte si elle comporte des statuts manquants au début, mais cela n'est pas vrai si elle comporte des statuts manquants au milieu ou à la fin.

Définition de la matrice de coûts pour la distance entre trajectoires La distance entre deux statuts est définie comme étant la différence en valeur absolue entre les revenus (d'activité et de remplacement) médians parmi les sous-groupes se trouvant dans chacun des deux statuts, en 2010. On choisit de considérer des distances en termes de salaire puisque c'est finalement l'adéquation de la trajectoire salariale de l'EIC avec la trajectoire salariale réelle connue par l'individu de l'Enquête Patrimoine qui nous importe. Il est évident que les distances entre statuts en termes de revenus salariaux et de remplacement ont pu varier entre 1984 et 2010, mais il nous faut définir une unique matrice de coûts de substitution.

Poids relatifs de la distance entre les salaires en 2001 et de la distance entre trajectoires Différents essais ont été effectués pour l'appariement afin de déterminer l'ordre de grandeur du poids qu'il faut attribuer à la distance entre les trajectoires relativement à celle

1. On ne dispose pas de suffisamment d'individus à la retraite dans l'échantillon EIC 2001 (puisqu'on ne prend en compte que les générations nées à partir de 1942).

entre les revenus salariaux et de remplacement en 2001. Lorsque l'on augmente le poids relatif attribué à la distance entre trajectoires, même considérablement, le pourcentage de statuts qui ne correspondent pas pour une période donnée entre ce qui est déclaré dans l'Enquête Patrimoine et ce qui est connu à travers l'EIC ne diminue pas beaucoup. Au contraire, la distance moyenne entre les revenus dans l'Enquête Patrimoine et dans l'EIC augmente fortement. On choisit donc d'accorder un poids relativement élevé à la distance entre les revenus en 2001, puisque cela ne détériore pas grandement la qualité de l'appariement en termes de trajectoires de statuts, tout en améliorant assez nettement celle en termes de revenus salariaux et de remplacement en 2001.

1.2 Trajectoires salariales

1.2.1 Imputation des années manquantes

Imputation des années manquantes pour la Fonction Publique d'Etat Les années 1979, 1981 et 1987 sont manquantes en totalité pour la Fonction Publique d'Etat. Dans le fichier EIC, les salaires de ces années ont été imputées en étudiant la probabilité d'être entré dans la fonction publique l'année N plutôt que l'année $N+1$ (si l'individu n'était pas déjà dans la base l'année $N-1$). Cette imputation de la présence dans la fonction publique est conservée, mais plutôt que de reprendre simplement le salaire de l'année précédente, on fait en sorte qu'il y ait une augmentation progressive entre les salaires des années $N-1$, N et $N+1$. Plus précisément, on veut qu'en moyenne :

$$W_{N+1} = W_N \,.\, g_{N+1} \,.\, t$$
$$W_N = W_{N-1} \,.\, g_N \,.\, t$$

avec g la croissance nominale des salaires dans le privé et semi-public (données INSEE) et t un facteur de croissance spécifique à la fonction publique et constant entre $N-1$ et $N+1$, et que l'on calcule donc comme $t = \sqrt{\frac{W_{N+1}}{g_{N+1} \,.\, W_{N-1} \,.\, g_N}}$.

On choisit ensuite d'appliquer ce facteur (multiplié par la croissance nominale) au salaire de l'année $N-1$ lorsque celui-ci existe, et au salaire de l'année $N+1$ dans le cas contraire, et ce si l'individu est présent dans la fonction publique l'année N. Les salaires des années 1976 et 1977 sont connus dans les DADS mais pas dans les EIC pour la Fonction Publique d'Etat (l'appartenance au fichier de paie des agents de l'Etat est par contre connue). On impute pour l'année 1977 le salaire de 1978 déflaté par la croissance réelle des salaires dans le privé et le semi-public si l'individu appartient à la Fonction Publique cette année-là, et de même pour 1976 en utilisant l'année 1977.

Imputation des années manquantes pour le secteur privé Les années 1981, 1983 et 1990 sont entièrement manquantes dans les DADS. On estime les revenus salariaux pour chacune de ces années grâce à des modèles tobit. Pour l'année 1981, on estime tout d'abord un modèle tobit en régressant les salaires nets totaux (issus de l'emploi privé) pour l'année 1980 sur les salaires (issus du privé) de 1979 et 1978, l'âge et l'âge au carré en 1980, le sexe, et la présence dans

une caisse du secteur privé en 1980 (connue grâce au fichier EIC), ce grâce à la procédure de Heckman. La sélection se fait sur les mêmes variables, avec en supplément le statut vis-à-vis de l'emploi en 1978 et 1979. Pour cette estimation, on utilise uniquement les individus pour lesquels il existe au moins un salaire strictement positif pour les années 1978, 1979 et 1982. Le salaire comme variable explicative est renseigné de façon catégorielle, par l'appartenance à l'un des neuf premiers déciles ou aux catégories P90-95, P95-99, P99-99.5, P99.5-99.9, P99.9-99.95, P99.95-99.99, et P99.99-100. Le fait de prendre en compte de façon fine l'appartenance au haut de la distribution les années passées permet par la suite de simuler des salaires suffisamment élevés pour l'année en cours. La variable de statut vis-à-vis de l'emploi est celle qui a été construite pour l'appariement, et qui reprend une partie des catégories de la variable CYACT de l'enquête Patrimoine. Une fois cette estimation effectuée, on effectue la prédiction des salaires nets de l'année 1981 ainsi que de la probabilité d'avoir un salaire positif, à partir des salaires des années 1980 et 1979, de l'âge et de l'âge au carré en 1981, etc. On n'applique cela qu'aux individus pour lesquels au moins l'un des salaires de 1979, 1980 ou 1982 est strictement positif. On impute un salaire nul aux individus dont les salaires de ces années sont tous nuls.

Pour l'année 1990, on transpose la même procédure en estimant tout d'abord un modèle tobit régressant les salaires de 1989 sur les données de 1987, 1988 et 1991 (et 1989 pour les variables d'âge et d'appartenance à une caisse du secteur privé), puis en effectuant la prédiction des salaires de 1990 à partir des données concernant les années 1988, 1989 et 1992. On applique la même procédure pour l'année 1983. Il n'est pas possible d'utiliser l'année 1981 pour l'estimation et/ou la prédiction, cette année ayant un statut particulier puisqu'elle a elle-même été imputée. On estime ainsi l'effet des salaires etc. des années 1980, 1985 et 1986 sur l'année 1984, puis on obtient la prédiction des salaires de 1983 en utilisant les données de 1979, 1984 et 1985.

1.3 Résultats préliminaires : description des trajectoires salariales

1.3.1 Volatilité des revenus sur le cycle de vie

On considère ici les trajectoires salariales sur la période allant de 1976 à 2001, dans le but d'examiner les inégalités intra-individuelles de revenu sur une partie du cycle de vie. Pour cela, on s'inspire du travail effectué par Björklund et Palme (1997) dans le cas de la Suède. Les auteurs utilisent des mesures d'entropie généralisées permettant de décomposer les inégalités en deux fractions, inégalités interindividuelles de revenu total calculé sur le cycle de vie d'une part, inégalités intra-individuelles au cours du cycle de vie d'autre part. Cette mesure des inégalités dépend d'un paramètre d'aversion à la pauvreté : lorsque celui-ci vaut 0 et 1 respectivement, on retrouve l'indice de Theil-L I_0 et l'indice de Theil I_1 respectivement. Le premier correspond à un degré plus grand d'aversion à la pauvreté, accordant une importance relativement plus grande aux revenus très faibles pour une période donnée. Les formules correspondant à ces deux indices sont les suivantes :

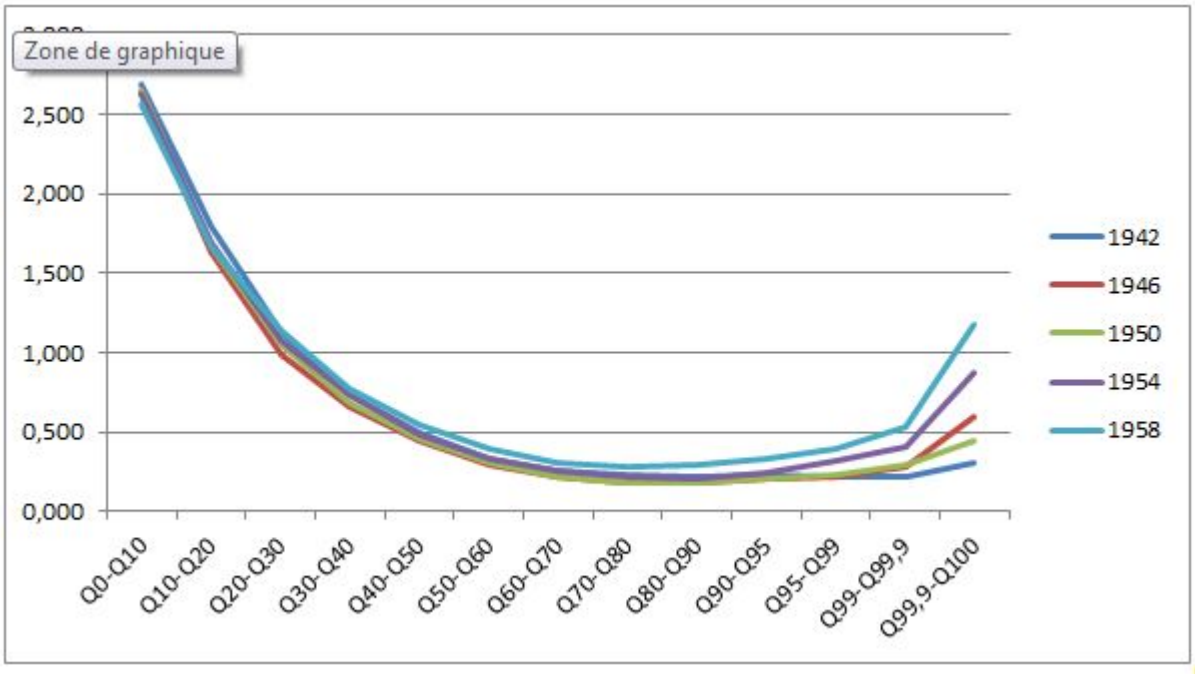
$$I_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\overline{y}} \log(\frac{y_i}{\overline{y}})$$

$I_0 = \frac{1}{n} \sum_{i=1}^{n} \log(\frac{\overline{y}}{y_i})$

avec y_i le revenu de l'individu à la période i et $\overline{y}$ son revenu moyen sur le cycle de vie. I_0 ne peut être calculé que dans le cas où les revenus sont non-nuls pour tous les individus à toutes les périodes. Cela n'est pas le cas pour nos trajectoires salariales. L'indice de Theil I_1 au contraire peut être calculé même lorsque le revenu des individus est nul à certaines périodes, en ne prenant pas en compte ces périodes dans la somme (on a $\lim_{y \to 0} y \log(y) = 0$). On calcule alors un indice de Theil intra-individuel pour chaque individu, sur la période 1976 à 2001 et pour les générations 1942 à 1958 (la génération 1958 a entre 18 et 43 ans et la génération 1942 entre 34 et 59 ans sur la période considérée). Cet indice individuel renseigne sur la volatilité des revenus sur le cycle de vie. Björklund et Palme considèrent quant à eux les revenus des individus sur une période de 18 années allant de 1974 à 1991, pour une catégorie d'individus " âgés " (ayant entre 33 et 47 ans en 1974) et une catégorie d'individus " jeunes " (ayant entre 18 et 32 ans en 1974). Pour chacun de ces deux groupes, ils examinent la corrélation entre l'indice de Theil intra-individuel (sur le cycle de vie) et le revenu total sur la période considérée. Celle-ci est significativement négative, signifiant que les carrières salariales sont plus instables dans le bas de la distribution. Lorsqu'ils effectuent la même analyse quartile par quartile, les auteurs trouvent une corrélation significativement négative pour le premier quartile, mais non significativement différente de zéro pour les trois quartiles de revenus (sur le cycle de vie) les plus élevés. Toutes générations confondues (1942 à 1958), on trouve pour les données EIC-DADS une corrélation entre indice de Theil intra-individuel et revenus salariaux sur le cycle de vie significativement négative (au seuil de 1 %), d'une valeur de -0.48. Lorsqu'on divise la population en quartiles, les résultats sont les suivants :

	P0-25	P25-50	P50-75	P75-100
Indice de Theil moyen	1.95	0.67	0.28	0.23
Corrélation avec le revenu total	-0.74 (***)	-0.47 (***)	-0.18 (***)	0.18 (***)

Et lorsqu'on examine les individus génération par génération, et en découpant la distribution à un niveau plus fin (particulièrement pour le haut de celle-ci), on peut figurer graphiquement les indices de Theil moyens par quantile :



On observe ainsi une baisse de l'instabilité des revenus salariaux jusqu'à un certain point, et une remontée assez prononcée dans l'extrémité haute de la distribution des revenus. Ceci pourrait être dû en particulier à la prise en compte des bonus et primes dans les salaires, bonus et primes qui sont variables au cours du temps et pouvant représenter des montants considérables pour les salariés les plus rémunérés.

2 Simulation sur le cycle de vie (Alexis)

Avant de commencer cette partie, certaines personnes doivent être remerciée. En effet, le projet TaxIPP-Life a bénéficié de l'aide de plusieurs autres projets. Le souci de partage des personnes concernées, leur réceptivité aux idées d'amélioration ainsi que leurs conseils et leurs temps ont été d'une aide extrêmement précieuse. TaxIPP-Lifen'est pas encore un sujet abouti mais il serait très éloigné de ce qu'il est en matière et auraient probablement des objectifs beaucoup plus restreint sans elles. Je remercie donc l'équipe d'OpenFisca [2] et en particulier M. Ben Jelloul, l'équipe de Liam2 [3] en en particulier G. Dekkers et G. de Menten, l'équipe Destinie [4] avec Aude Leduc et Anthony Marino et l'équipe de PensIPP et en particulier D. Blanchet et S.Rabaté.

2.1 Quelques mots de techniques

Dans cette courte section, seront décrits les choix techniques qui ont été fait pour réaliser la simulation. Nous verrons que la question de la performance ne doit pas être négligée. Appliquer

2.
3.
4.

la législation socio-fiscale d'une année prend plus de 15 minutes lorsqu'elle est codée en SAS par exemple. Dans un calcul sur l'ensemble du cycle de vie, cette législation doit être simulée disons cent fois, l'optimisation du temps de calcul peut donc avoir du bon pour ne pas prendre 25 heures de calculs. Cela d'autant plus que la volonté de travailler sur un gros échantillon peut augmenter considérablement le temps de calcul, et que l'on a l'idée un jour d'introduire de la décision comportementale des agents en fonction de leur espérance de gain actualisée, et donc de simuler la législation bien plus de cent fois.

2.1.1 La base de données initiales

Comme on peut l'imaginer à la lecture de la première partie de ce document c'est l'enquête patrimoine qui va servir de base d'information à notre modèle. La raison en est la connaissance des trajectoires professionnelle passée. La connaissance du patrimoine est aussi , bien sûr, une des forces de l'enquête. On aurait pu utiliser d'autres sources. Par exemple l'enquête Santé et Itinéraire Professionnel, qui comme son nom là aussi l'indique, est bien moins précise sur le patrimoine mais bien plus sur les conditions de santé. Pour la simulation du futur uniquement, on peut aussi envisager l'utilisation d'enquête plus générale et n'ayant pas de volet trajectoire. Ainsi, l'enquête budget des familles est par exemple attirante car la consommation est alors connue, au moins pour une année. On peut aussi penser à l'enquête Logement. L'idée est à étudier mais on peut essayer d'étendre le matching statistique réalisé avec deux enquêtes à plusieurs en constituant se faisant une sur-enquête couvrant la majorité des champs micro-économiques.

Dans la suite, nous ne parlerons que de l'enquête patrimoine 2009-2010. L'échantillon contient 29 951 individus répartis en 12788 ménages [5]. De plus, 14 954 déclarations fiscales sont imputées (pour l'instant sommairement).

2.1.2 La fermeture de l'échantillon

Une première limite de l'enquête patrimoine pour notre sujet est qu'elle ne donne pas de lien entre les ménages. Si les relations sont connues au sein d'un ménage, on veut aussi connaitre les relations et particulièrement les liens de filiation entre les individus qui ne vivent pas dans le même ménage. Cela peut permettre d'imputer ensuite les successions, les pensions alimentaires l'aide pendant les études mais aussi les différentes formes d'aide vers les ascendants, en particulier en cas de dépendance. On peut aussi imaginer d'imputer un recours aux grands parents pour les gardes d'enfants.

Dans l'enquête patrimoine, les parents ayant des enfants vivant à l'extérieur du domicile se voient poser quelques questions à leur sujet (date de naissance, nombre d'enfant, statut sur le marché de l'emploi, diplôme, etc). Nous créons ainsi des individus factices représentant ces enfants hors domicile. Puis par un procédé de matching, nous cherchons des individus correspondant, déclarant un ou deux parents en vie, qui correspondent le plus possible à ces statistiques.

5. Les ménages interrogés aux Antilles, ont un questionnaire légèrement différent. Par souci de simplicité et pour travailler avec un échantillon uniforme, ils sont pour l'instant exclu du champ.

On leur attribue alors les parents enquêté de l'individu fictif comme étant ses parents

Nous n'utilisons pas les informations ascendantes, celle que les enfants déclare à propos de leur parents. Elles sont assez pauvre et concerne pratiquement exclusivement l'enfance de l'enfant or il se peut que la situation des parents ait changé entre l'enfance de l'enfant et le moment de l'enquête. On pourrait aussi utiliser les informations sur le décès ou non des grands parents. Enfin, notons que nous n'utilisons que les données de l'enquête, on n'utilise pas d'information annexe par exemple sur la corrélation entre les revenus des parents et des enfants. Pour l'instant, on fait donc l'hypothèse que les variables utilisées sont suffisantes pour que les distributions croisées soient respectées.

Nous donnons ici une petite précision sur la méthode de calcul pour souligner au lecteur une petite limite actuelle de cette étape. Lorsque les parents sont interrogés sur leurs enfants vivant hors du domicile, on leur demande de préciser s'il s'agit de l'enfant du couple ou de l'un d'eux seulement. S'il s'agit de l'enfant du couple, alors on cherche des enfants dont les deux parents sont vivants. Il reste ensuite de tels enfants avec deux parents vivants à apparier ce qui est logique car les parents peuvent être vivant et ne plus vivre ensemble. On cherche pour eux, comme pour les enfants n'ayant qu'un parent à trouver, un père puis une mère. Si on ne cherche qu'un parent, il n'y a pas de problème, en revanche, un point à noter est que lorsque l'on cherche deux parents, on les cherche indépendamment. Par exemple, si on a deux enfants identiques, deux pères potentiels, l'un de 40 et l'autre de 60 ans et deux mères potentielles, l'une de 40 ans et l'autre de 60 ans, rien ne contrôle que l'on a plus de chance d'avoir un couple dont les deux membres ont 40 ans et un dont les deux couples ont 60 ans que deux couples avec 20 ans d'écart d'âge. On peut donc ne pas reproduire la réalité statistique en créant une distribution jointe des parents non réaliste. Du moins c'est le cas, si on pense que deux couples de parents peuvent avoir des enfants similaires même s'ils sont différents au départ. Une solution serait, une fois la mère trouver de chercher un père avec un age assez proche, une autre solution serait de réaliser un mariage fictif des parents ayant des enfants similaire puis de tirer un couple de parents aux enfants.

Enfin, précisons que le matching est fait à partir d'une méthode de score (voir Annexe pour plus d'information sur les méthodes de matching)

2.1.3 Étendre l'échantillon

Travailler avec des pondérations n'est en général pas un problème dans le cadre de statistiques statiques. Cependant, dans le contexte particulier de la microsimulation dynamique où des liens sont établis entre ménage cela pose problème. On ne peut en effet pas associer 50 enfants à 3 000 mères ou unir un homme en représentant 1000 avec une femme représentant 2000 femmes de la population française. On se demande par exemple, comment on pondérerait leurs enfants. Pour solutionner le problème, le choix a été fait de dupliquer les individus autant de fois que nécessaire pour ainsi avoir une pondération uniforme. Dans l'enquête patrimoine 2010,

qui contient un sur-échantillonnage des hauts-patrimoines, la plus petite pondération est de 6 et cela devrait être la pondération uniforme. Cela nous mène en théorie à une base de plus de 10 millions de lignes dans la base pour représenter la population française.

Le modèle DESTINIE tire ensuite un sous-échantillon à partir de cette population étendue. TaxIPP-Life, quant à lui, essaie de tourner sur l'ensemble de cet échantillon. Cela a aussi l'avantage qu'un même individu au départ, se verra associé plusieurs carrière, on conserve ainsi une certaine variabilité dans le futur de chaque individu de la base initial ce qui ne peut être que bon pour la précision statistique. Comme nous le verrons ci-dessous, travailler sur une base étendue est un vrai challenge. Pour l'instant, nous travaillons en général sur l'échantillon non étendu, sans tenir compte des pondérations dans les différentes étapes. Une alternative acceptable en terme de temps de calcul et qui permet de tester le modèle est d'utiliser comme pondération uniforme 200, ce qui fait une base de 300 000 individus. Les ménages avec une pondération initiale inférieure à 200 se voit attribuer un poids de 200 ce qui biaise donc en théorie pour l'instant un peu les résultats

2.1.4 Choix de langage de programmation

Toutes les étapes initiales, y compris l'extension de l'échantillon décrit ci-dessus, sont réalisées en R. Les étapes de simulation, vieillissement de la population et calcul de la législation sont réalisées en Python. On utilise ainsi les forces de chaque programme [6]. L'utilisation de Python permet une écriture adaptée à la microsimulation et est relativement incomparable avec les logiciels statistiques en terme de performances dont on a montré qu'elles étaient importantes pour TaxIPP-Life. L'étape de calcul de la législation qui prend 15 minutes en SAS est effectué dans un temps de l'ordre de grandeur de dix secondes. La possibilité de travailler avec un gros échantillon ou avec des équations comportementales devient envisageable.

2.2 Les étapes de la simulation dynamique

Avant d'entamer cette section, il faut préciser que TaxIPP-Life est encore un projet jeune en développement. Les étapes présentées ici peuvent toutes être améliorer, certaines sont tellement rudimentaires qu'elles n'ont d'autre mérite que d'exister. La description de ces étapes permet toute fois de donnée une première idée des fonctionnalité et possibilité de TaxIPP-Life. L'écriture sous forme de module indépendant permet à tout moment d'améliorer chacune de ces étapes.

2.2.1 Démographie

L'age est bien sûr incrémenté à chaque période.

6. Parce qu'il s'agit d'un matching, l'étape de créations des liens entre parents et enfants vivant hors domicile est, elle aussi, réalisée en Python.

Naissance . On donner à toutes les femmes en couple âgées de 16 à 50 ans se voir imputer un enfant [7]. Un certain nombre de ces femmes sont sélectionnées en s'assurant de vérifier les projections de naissance en fonction de l'âge de la mère. Le sexe de l'enfant est choisi aléatoirement.

Le fait de ne pas avoir de probabilité différente pour les femmes selon leur caractéristiques, autre que leur age est problématique. Le nombre d'enfant ainsi que le niveau d'étude devrait intervenir [8].

Décès . La probabilité de décès ne dépend que de l'age et du sexe. Le nombre de décès prédit par les projections démographiques de l'Insee est ainsi reproduit. Si une personne décédée vit seule avec des enfants, ces enfants sont attribués à l'autre parent si celui-ci est encore en vie. Si ce n'est pas le cas, ils sont affectés à un ménage spécial sans adulte avec tous les autres enfants dans ce cas.

Union . Pour l'instant, dans le modèle, l'union entre deux personnes correspond à la fois à un emménagement et à une union légale. Sans qu'il n'y ait de contrindication technique à le faire, il n'y a donc pas de création de concubinage. Il en existe toutefois dans la base initiale. Reprenant les équations utilisée par DESTINIE à partir du travail de Duée(? ?) on impute une probabilité d'avoir se mettre en couple. Cela est fait d'une part pour les premières unions et d'autre part pour les personnes ayant déjà été en couple séparément selon le sexe, en fonction de l'age depuis la fin des études, du fait d'avoir un enfant et, de la durée depuis la précédente séparation le cas échéant. En utilsant, une méthode d'alignement de Liam2 (voir ...), un tiers de ces personnes est sélectionné [9]. Ensuite, un matching est effectué pour associer les individus en fonction de leur âge, de leur différence d'âge et de leur niveau de diplôme comparé. Encore une fois, ce matching, crucial pour étudier bon nombre de question de redistribution, par exemple, l'impact de l'endogamie, mériterait d'être affiné. En cas d'union, les personnes rattachée aux nouveaux conjoints (enfants le plus souvent) emménage avec eux. Idem pour les déclarations fiscales.

Séparation . C'est plus crédible cette fois : les séparations se traduisent à la fois par un changement de logement d'un des deux conjoints et par une rupture du contrat (et donc le passage à des déclarations fiscales séparées). La probabilité de rupture dépend du nombre d'enfant du couple, de son ancienneté et de la différence d'âge entre ses membres. Ceci pour être régi par des probabilité mais pour l'instant c'est l'homme qui déménage (sauf lorsque le couple habitait chez ses parents) et la femme qui a une nouvelle déclaration [10]. Les enfants et personnes rattachées ne change de logement que s'ils sont liés à la personne qui part sans être liés à la personne qui reste, que ce soit pour le logement ou pour la déclaration fiscale.

7. On élimine les femmes qui ont 8 enfants ou plus ou dont le conjoint est dans cette situation. On pourrait toutefois supprimer sans difficulté cette condition.

8. On ne peut pas pour autant dire que le tirage obtenu et complétement indépendant du niveau d'étude et du nombre d'enfant dans la mesure où ceux-ci interviennent dans les équations de mise en couple et de divorce, pré-requis pour se voir imputer des enfants dans TaxIPP-Life.

9. Un calage plus pertinent ne serait pas une mauvaise chose.

10. Dans le cas de couple homosexuel, le choix est fait aléatoirement

2.2.2 Logement

Le travail sur le logement est extrêmement rudimentaire. Les déménagement ne se font que lors des unions et séparations et lorsque qu'une personne de plus de 24 ans n'est ni personne de référence ni en couple (i.e. vit avec ses parents). Le travail sur les loyers n'a pas encore été fait mais ne devrait pas poser de difficulté majeure.

On pourrait à terme utiliser avoir réellement un parc de logement dont on simulerait la variation chaque année. Lors des déménagement, les ménages choisirait l'un des logements en fonction de la taille du ménage et du logement ainsi qu'en fonction des revenus et du loyer. On simulerait ainsi des déménagements en fonction d'un surpeuplement ou sous-peuplement ainsi qu'en cas d'évolution dans les revenus. L'idée serait de reproduire des tensions sur le marché du logement.

2.2.3 Activité et Revenus

2.2.4 Épargne et consommation

2.3 Projets futurs

2.3.1 Améliorations des imputations

Consommation, importantes...

2.3.2 Simulation par mois

2.3.3 Transferts intergénérationnels

2.3.4 Comportements

2.3.5 Autres enquêtes

3 Résultats préliminaires : simulation des allocations familiales (...et de ? ?)

Annexe

Méthode de matching

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Salarié du public à temps complet (1)	0	8600	2900	12400	5300	15120	11900	19260	7100	23300
Salarié du public à temps partiel (2)	8600	0	5700	3800	3300	6520	3300	10660	1500	14700
Salarié du privé à temps complet (3)	2900	5700	0	9500	2400	12220	9000	16360	4200	20400
Salarié du privé à temps partiel (4)	12400	3800	9500	0	7100	2720	500	6860	5300	10900
A son compte (5)	5300	3300	2400	7100	0	9820	6600	13960	1800	18000
Chômage (6)	15120	6520	12220	2720	9820	0	3220	4140	8020	8180
Succession de périodes d'emploi et de chômage (7)	11900	3300	9000	500	6600	3220	0	7360	4800	11400
Reprise d'études ou formation (8)	19260	10660	16360	6860	13960	4140	7360	0	12160	4040
Préretraite (9)	7100	1500	4200	5300	1800	8020	4800	12160	0	16200
Inactif et autres (10)	23300	14700	20400	10900	18000	8180	11400	4040	16200	0

Tableau résumant la qualité du matching selon les poids relatifs utilisés pour les distances ?

Tableau décrivant les salaires imputés par rapport aux années adjacentes ?

Références

[1] Anders Bjorklund, Markus Jantti, and John Roemer. Equality of opportunity and the distribution of long-run income in sweden. IZA Discussion Papers 5466, Institute for the Study of Labor (IZA), 2011.