Statistiques socio-économiques et politique publique: Nouveau röle pour les modèles de microsimulation

par Michael C. Wolfson[1]

No. 81

11F0019MPF No. 81
ISSN: 1200-5231
ISBN: 0-662-99185-0

24 Immeuble R.H. Coats, Ottawa, K1A 0T6
(613) 951-8216 Télécopieur: (613) 951-5403
internet: wolfson@statcan.ca

Juillet 1995

Document rédigé pour les 50e assises de l'Institut international de statistique, Beijing, 21-29 août 1995 Communication sollicitée sur l'évolution des attentes des utilisateurs concernant la qualité des statistiques dans les secteurs public et privé

Also available in English

---

[1] Statistique Canada et Institut canadien des recherches avancées. Le présent document propose le travail d' une équipe principalement constituée des membres du Groupe de la modélisation socio-économique de la Direction des études analytiques de Statistique Canada. Geoff Rowe et John Armstrong ont mené l'analyse empirique tandis que Steve Gribble a adapté le logiciel ModGen qui sert d'environnement au modèle de microsimulation LifePathss et a assuré en permanence la direction de l'équipe. Les erreurs ou les incohérences qu'on pourrait relever dans le présent document doivent m'être entièrement imputées.

Resumé

Les utilisateurs de statistiques socio-économiques veulent habituellement plus de renseignements, et de meilleure qualité. Souvent, on peut répondre à ces besoins simplement par des collectes de données plus poussées, qui sont soumises aux contraintes habituelles relatives aux coûts et au fardeau de réponse des répondants. Les utilisateurs, particulièrement aux fins de politiques publiques, continuent de réclamer, et cette demande n'est toujours pas comblée, un système intégré et cohérent de statistiques socio-économiques. Dans ce cas, des données supplémentaires ne seront pas suffisantes; la contrainte la plus importante demeure l'absence d'une approche conceptuelle convenue.

Nous examinons brièvement ici l'état des cadres d'utilisation des statistiques sociales et économiques, y compris les genres d'indicateurs socio-économiques que les utilisateurs pourraient désirer. Ces indicateurs sont premièrement justifiés, en termes généraux, par des principes de base et des concepts intuitifs, ce laisse de côté les détails de leur construction, pour le moment. Ensuite, nous montrons comment une structure cohérente de tels indicateurs peut être assemblée.

Une conséquence fondamentale est que cette structure exige un réseau coordonné d'enquêtes et de processus de collecte de données, ainsi que des normes supérieures de qualité de données. Ceci, à son tour, implique une décomposition des systèmes sur mesure qui caractérisent la majeure partie du travail d'enquête des agences statistiques nationales (ex. des "chaînes de production" de données parallèles, mais généralement non reliées). De plus, les données émanant du réseau d'enquêtes doivent être intégrées. Puisque les données en question sont dynamiques, la méthode proposée dépasse la correspondance statistique et s'étend aux modèles de microsimulation. Enfin, ces idées sont illustrées avec les résultats préliminaires du modèle LifePathss actuellement en cours d'élaboration à Statistique Canada.

Mots clés: microsimulation, statistiques sociales, cadres statistiques

1. Introduction

Il est tout à fait raisonnable de s'attendre que les services de statistique nationaux brossent un tableau fiable, cohérent et pertinent des processus socio-économiques (lire Garonna, 1994; OCDE, 1976). Dans une large mesure, ils y parviennent grâce au Système de comptabilité nationale (SCN). On sait néanmoins que maintes faiblesses, qui plus est sérieuses, affligent le SCN, en particulier sur le plan des préoccupations sociales et des politiques. Ces faiblesses expliquent le désir de concevoir des séries ou des ensembles d'indicateurs sociaux susceptibles de recevoir la sanction internationale.

Jusqu'à présent, les efforts déployés pour créer des indicateurs sociaux d'une ampleur et d'une cohérence similaires à celles du SCN, et acceptés partout dans le monde comme celui-ci, se sont soldés par un échec. Pour cette raison, sans doute devra-t-on envisager des approches fondamentalement différentes si on veut satisfaire la demande de statistiques socio-économiques, notamment répondre aux besoins qui motivent depuis longtemps les travaux poursuivis dans le domaine des indicateurs sociaux.

Au sens large, on a envisagé trois grandes stratégies pour donner un cadre statistique à la sphère sociale. Une solution consisterait à prolonger le SCN, principalement sous forme de matrices de comptabilité sociale (MCS; lire Pyatt, 1990) ou de comptes satellites (lire Vanoli, 1994; Pommier, 1981). La seconde stratégie proposée prévoit la construction d'un canevas adapté à la statistique sociale - le plus connu et le mieux développé étant le système de statistiques sociodémographiques de Stone (SSSD; ONU, 1975; Stone, 1973). La troisième approche abandonne la structure et la cohérence d'un cadre explicite et les remplace par un ensemble spécial d'indicateurs statistiques faisant l'assentiment de tous. Une parfaite illustration de cette approche est le jeu d'indicateurs sociaux recommandé par l' OCDE (Moser, 1973; OCDE, 1976).

Aucune de ces trois stratégies n'a été mise en oeuvre à suffisamment grande échelle dans les pays industrialisés pour qu'on dispose de la base nécessaire à la production de données comparables à l'échelon international. Trois raisons expliquent cet échec. La première à trait à la faisabilité; la seconde résulte du manque d'intérêt de la part des gouvernements ou des organismes qui procurent les services de statistique; enfin, la troisième se rapporte au peu d'attention qu'engendre le sujet. (Manifestement, ces raisons partagent des liens entre elles.) L'expérience sur les indicateurs sociaux (OCDE, 1982) a néanmoins permis de dresser une liste opérationnelle et détaillée d'indicateurs - fruit d'un consensus entre experts et hauts fonctionnaires du gouvernement des pays membres. Malheureusement, les mécanismes de collecte des données qui faciliteraient la comparaison au niveau international n'ont pas encore été implantés pour bon nombre des indicateurs sur lesquels on s'est entendu (p.ex. vie saine, emploi du temps, revenu). Pourtant, on ne peut parler d'impossibilité technique puisque certains pays recourent déjà à des systèmes de données à des fins analogues.

Reste donc comme principale explication de cet échec un amalgame d'insignifiance et de priorités mal définies - entre autres la réticence d'investir les ressources voulues dans la collecte de statistiques. Le peu d'intérêt que soulèvent les indicateurs sociaux se situe plus au niveau de la comparabilité internationale qu'à celui de l'utilité locale. En effet, la plupart des pays utilisent déjà des statistiques sociales très variées; la difficulté est qu'il est généralement impossible de les comparer entre nations. Le manque d'intérêt pour un ensemble de statistiques sociales comparable à l'échelon international pourrait tout simplement venir de l'importance nettement plus grande accordée aux préoccupations de nature économique (comme en témoignent les efforts, couronnés de succès, visant à créer un SCN d'envergure internationale), comparativement aux préoccupations d'ordre social, du moins quand on les compare. Une autre raison pourrait être que quelques indicateurs sociaux (p.ex. espérance de vie, taux de chômage) permettent déjà une comparaison entre pays, et qu'on les juge suffisants.

Une des raisons soulignant l'utilité de données comparables sur le plan international est que les pays partagent des liens étroits dans le secteur en question. En ce qui concerne l'économie, les flux financiers internationaux rapprochent manifestement beaucoup les pays où on retrouve de surcroît des courants de pensée analogues sur la théorie macro-économique. Pareilles affinités sont à l'origine du SCN. Les liens pourtant tangibles au niveau social paraissent peut-être plus lâches (bien qu'on puisse douter de cette explication, étant donné les importants flux culturels et intellectuels qu'engendrent les médias de masse et les tendances similaires, observées partout dans le monde, relatives au terrorisme, au chômage,

à l'éclatement de la famille, à la dénatalité, à l'inégalité grandissante des salaires, etc.). De plus, faute d'une théorie commune, la comparabilité internationale des données repose sur une base plus fragile - remarque aussi valable pour les statistiques individuelles (à savoir, répartition du revenu entre les ménages) que pour la manière dont on regroupe diverses statistiques sociales (quand on le fait). Bref, l'hésitation des pays à investir dans la constitution des systèmes de collecte des données essentiels aux indicateurs sociaux de l'OCDE pourrait bien venir d'un manque d'intérêt pour la comparabilité d'une série particulière d'indicateurs.

Quoi qu'il en soit, l'incapacité d'implanter le SSSD ou les comptes satellites dans le domaine social doit avoir d'autres raisons qu'un simple manque d'intérêt pour des données comparables à l'échelon international; le peu d'intérêt que soulève le cadre théorique sous-entendu par ces dernières doit aussi faire partie du problème. Les causes du mal sont ici plus profondes, car il est difficile de trouver des structures générales cohérentes dont l'envergure et la mise en oeuvre se rapprochent même de loin du SCN pour les statistiques sociales, y compris à l'intérieur d'un pays.

L'échec des stratégies précitées décennie après décennie donne à penser que l'élaboration d'une structure cohérente pour les statistiques sociales doit être envisagée sous un nouvel angle. Des statistiques plus cohérentes contribueraient à répondre aux besoins d'une population d'utilisateurs toujours plus nombreux. D'une part, elles établiraient la base essentielle à la satisfaction des exigences du généraliste (par exemple, un aperçu sommaire des grandes tendances) et, d'autre part, elles permettraient une organisation des statistiques sociales complexes susceptible de venir en aide aux utilisateurs spécialisés, chez qui l'existence de plusieurs estimations hétérogènes du même élément, selon la provenance des données, pourrait semer la confusion (on comprend pourquoi).

2. Principes élémentaires

Avant de chercher de nouvelles approches à la construction d'un modèle pour les statistiques sociales, il convient de se fixer une série d'objectifs quantitatifs fondamentaux. En voici trois auxquels devrait se rallier, espérons-le, la majorité.

a. Issue générale - Un des principaux objectifs consiste à déterminer si la situation empire ou s'améliore. Les gens sont-ils mieux lotis que l'année ou la décennie antérieures ? Répondre à une telle question s'avère difficile, principalement parce qu'aucune approche sommaire permettant de jauger le bien-être des individus ne fait l'assentiment général. Le revenu, l'état de santé, le niveau de scolarité et la privation sociale sont autant d'éléments qui entrent dans une telle mesure. Cependant, on ne s'entend pas sur les autres facteurs dont il faudrait tenir compte, ni sur la façon de les combiner pour obtenir un indice général. En outre, le concept même de l'issue recherchée ne fait pas l'objet d'un consensus dans certains domaines comme la santé et l'éducation.

Une entente partielle de ce genre a d'importantes répercussions sur l'élaboration d'un modèle statistique. La première est qu'on a besoin d'une marge de manoeuvre. Si certains paramètres permettant de mesurer le bien-être global engendrent un consensus, il est essentiel de les inclure au programme statistique sous-jacent. Cependant, puisqu'il n'existe pas une seule et unique «bonne manière» de les combiner, les utilisateurs devraient disposer d'une certaine latitude à cet égard. On devrait pouvoir forger différents indices sommaires à partir des éléments fondamentaux - tant pour des aspects comme la santé et l'éducation, que sur un plan plus général.

Une deuxième implication est que les statistiques sociales ne devraient être ni asservies aux statistiques économiques, ni s'en dissocier, comme l'ont fait jusqu'à présent les trois grandes approches stratégiques au problème. En effet, la situation économique se trouve manifestement au coeur même de toute mesure générale servant à déterminer si la situation des gens s'améliore ou non. Par conséquent, on envisagera de préférence un cadre pour des statistiques *socio-économiques* plutôt que des statistiques purement sociales. Ainsi, contrairement à ce qu'en pensent certains spécialistes des comptes nationaux (lire Vanolli, 1994), bâtir un modèle pour les statistiques sociales à partir des principes qui régissent le SCN laisserait à désirer. On doit plutôt songer à des approches innovatrices qui tireraient parti de nouvelles

prémisses et engloberaient «l'économie sociale» dans son ensemble. De cette façon, le SCN deviendrait un important élément d'un plus vaste «système de statistiques socio-économique» (Wolfson, 1994; Ruggles et Ruggles, 1973).

Une troisième implication s'ajoute aux deux précédentes, soit que les méthodes sommaires qui permettent de comparer deux économies sociales ou davantage dans le temps ou entre divers pays ne doivent pas se résumer à une agrégation linéaire. Ainsi l'intérêt est un concept qui se prête parfois mal à l'agrégation à partir d'un seul numéraire, comme le fait le SCN avec les valeurs monétaires. Heureusement, les recherches sur certains sujets comme les méthodes servant à comparer la distribution du revenu entre les ménages, les distributions qui recourent de façon plus générale à des techniques graphiques (lire Easton et McCulloch, 1990) et les bases de données à architecture souple autorisant l'application des mathématiques à théorie figée, reposant sur des pointeurs, aux ensembles complexes de microdonnées longitudinales multivariées - désormais aisément utilisables grâce aux progrès de l'informatique - démontrent qu'une agrégation analogue à celle du SCN n'est pas essentielle. De fait, de telles approches se complètent et appuient le deuxième objectif que voici.

b. Diversité -- Un autre objectif fondamental de mesure consiste à aider les utilisateurs à percevoir la diversité de l'économie sociale. Le terme «diversité» englobe l'hétérogénéité sous de nombreuses formes -- par exemple, ne pas se borner aux agrégats et aux moyennes pour apaiser une critique qui revient constamment au sujet du SCN, soit que ce dernier ne dit rien des nantis, des démunis et de la répartition inégale du revenu. La diversité se reflète aussi dans la dispersion du niveau de scolarité et de la structure des ménages au sein de la population d'un pays.

Saisir cette diversité a des implications fondamentales pour la statistique. Essentiellement, un tel exercice requiert des bases de microdonnées explicites. Modèle statistique prédominant, le SCN a été conçu avant la révolution de l'ordinateur. En ce sens, il nuit à l'exercice de réflexion créateur au sujet d'un cadre utile pour les statistiques sociales et socio-économiques. À l'ère pré-informatique où la structure de base du SCN a été conçue, l'agrégation n'était pas seulement un fondement théorique, c'était une nécessité. Aujourd'hui, grâce aux nouvelles techniques d'exploitation des bases de données, l'agrégation n'entrave pas que l'expression précise de la diversité, elle devient inutile sur le plan pratique.

(L'idée de bases de microdonnées explicites dans le cadre de projets statistiques d'une telle envergure ne date pas d'hier : lire Organisation des Nations Unies, 1979; Ruggles, 1981. Pourtant, le principe de «l'agrégation» a tout envahi, comme on a pu le constater avec les efforts déployés par l'OCDE pour mettre au point des indicateurs sociaux. Cette organisation a jugé bon de définir une série «de désagrégations fondamentales des principaux indicateurs sociaux» *a priori* ; OCDE 1977. Le fait qu'on se soit entendu à cet égard n'est pas dépourvu d'utilité, mais un tel exercice n'est pas essentiel quand les analystes disposent déjà de bases de microdonnées appropriées et comparables à l'échelon international.)

c. Et si ? -- Le troisième objectif de mesure fondamental consiste à établir la base qui permettra de poser des questions du genre «et si ?» et d'y répondre correctement. Il existe deux raisons élémentaires à cela. La plus évidente est que les services gouvernementaux qui forgent les politiques et les décideurs du secteur privé, principaux utilisateurs des statistiques socio-économiques, s'efforcent justement de trouver une réponse à de telles questions, par exemple comment le revenu disponible se répartirait-il si on apportait telle ou telle modification aux politiques fiscales/de transfert ? Ou encore, combien dépenserait-on pour tel ou tel produit dans cinq ans si les tendances actuelles se maintenaient ?

Une raison moins apparente mais tout aussi importante est qu'en réalité, les indicateurs statistiques constituent la réponse à ces questions. L'espérance de vie en est la meilleure illustration. L'espérance de vie en 1990, par exemple, répond à la question hypothétique suivante: «combien de temps vivrait une cohorte de naissances si tous ses membres étaient constamment exposés au taux de mortalité observé en 1990 ?» L'espérance de vie n'est pas une donnée que l'on peut observer directement comme le nombre de décès selon l'âge et le sexe. Il s'agit plutôt du résultat d'une simulation numérique étroitement associée au nombre de décès et de personnes à risque pris respectivement comme numérateur et dénominateur après désagrégation.

Si l'espérance de vie est un élément artificiel impossible à mesurer directement, il s'agit aussi d'un concept intuitif, facile à saisir et pouvant servir de canevas à des séries d'indicateurs apparentées. On le

constate le plus clairement dans le secteur de la santé où un groupe spécial de chercheurs du REVES (Réseau pour l'espérance de vie en santé; Mathers et Robine, 1993) s'est constitué dans le but de mettre au point une telle série d'indicateurs et de susciter un consensus à leur sujet.

Lorsqu'on rassemble tous les éléments du débat, trois objectifs de quantification fondamentaux se dégagent:

- indicateurs généraux permettant de déterminer dans quelle mesure la situation de la population s'améliore;
- capacité d'illustrer la diversité et l'hétérogénéité de la population;
- instruments permettant de poser des questions du genre «et si ?» et d'y répondre.

Pour atteindre ces objectifs, le modèle statistique doit satisfaire aux exigences suivantes:

- il doit être souple;
- il doit englober les aspects social et économique;
- il doit reposer sur des bases de microdonnées explicites;
- il doit recourir aux techniques contemporaines d'informatique et d'exploitation des bases de données;
- il doit intégrer des modèles de simulation étroitement associés aux données.

À partir de telles prémisses, quels pourraient être les principaux éléments d'un modèle de statistiques socio-économiques ? On peut formuler les remarques suivantes:

- à un moment quelconque dans le temps, la meilleure façon de représenter la population consiste à prélever un échantillon d'individus, chacun caractérisé par un jeu d'attributs et de relations données;
- les attributs comprennent le revenu, le niveau de scolarité, la consommation, divers paramètres de l'état de santé et les tendances relatives au temps consacré à diverses activés;
- les relations se rapportent aussi bien aux liens de parenté classiques qu'à la cohabitation (à savoir, dans les bases de données ou sur le plan graphique-théorique, on peut représenter les relations de ce genre au moyen de pointeurs différents désignant d'autres personnes, qui font aussi partie de la base de données);
- le terme «relation» couvre également les liens avec les grandes institutions sociales, soit l'école, le travail et les programmes gouvernementaux. Les prises de contact, les relations ou les transactions entre individus et grandes institutions peuvent faire partie du jeu d'attributs personnels. Il pourrait s'agir de pointeurs se rapportant à la description des institutions - école, lieu de travail et programmes gouvernementaux - avec lesquelles le sujet a affaire;
- la base de données peut ensuite aisément être perçue comme une hiérarchie composée d'unités de type variable, à savoir individu, famille nucléaire, famille étendue et ménage;
- chaque unité (sujet, famille ou ménage) peut être décrite par un ou plusieurs attributs sommaires, par exemple le revenu disponible, les heures de loisir ou le degré de satisfaction personnelle;
- cela fait, il est possible d'évaluer la diversité de la population par application de statistiques sommaires à la distribution mixte multivariée des unités (p.ex., coefficient de Gini, quantiles);
- dans le temps, la meilleure façon de représenter la population consiste à utiliser une série de biographies, donc l'équivalent d'une vaste et longue enquête longitudinale à échantillon constant;
- grâce à une telle représentation longitudinale, la généralisation du concept de l'espérance de vie permettra de bâtir un ensemble cohérent d'indicateurs sommaires - notamment par division de l'espérance de vie en périodes cumulatives passées dans divers stades de l'existence.

Un tel modèle socio-économique intégrerait essentiellement un échantillon complet de micro-données longitudinales, véritable microcosme de la population réelle et de ses liens avec les principales institutions sociale et économiques. On pourrait aisément concevoir un vaste assortiment d'indicateurs statistiques à partir d'un tel microcosme, en réalité fait sans beaucoup plus d'efforts qu'enfoncer la touche <enter> présente sur tous les claviers d'ordinateur pour lancer l'algorithme approprié et faire analyser les données du microcosme par le logiciel.

De par leur construction, de tels indicateurs sommaires seraient cohérents, car ils dériveraient d'une base de microdonnées identique. Ils ne cacheraient pas la diversité et l'hétérogénéité de la population, la base de microdonnées étant toujours accessible pour une analyse approfondie (au déclic de la souris, par exemple, si on pense à l'aspect fonctionnel de l'informatique contemporaine).

La principale question que l'on peut se poser est la suivante: d'où viendrait ce microcosme ? Pour des raisons très pratiques (coût, fardeau pour les répondants et préoccupations relatives au respect de la vie privée), il ne pourrait venir d'une enquête longitudinale générale sur les ménages. En outre, on ne disposerait pas de cinquante ans ou davantage pour compléter une telle enquête, car au bout de ce laps de temps, de nombreuses choses auront changé de facon radicale. La conclusion inévitable est que le microcosme en question doit être artificiel.

La synthèse du microcosme prolongerait celle de la cohorte que sous-entendent déjà certains indicateurs comme l'espérance de vie. La méthodologie utilisée différerait néanmoins, car l'approche reposant sur une agrégation partielle ou la constitution de cellules, inhérente à la table de survie sous-jacente, est incompatible avec les bases explicites de microdonnées. On devrait plutôt recourir à une microsimulation.

De fait, on propose un hybride des idées de Stone sur le SSSD (ONU, 1975) et des bases de microdonnées intégrées explicites qu'a envisagées un groupe d'experts international subséquent (ONU, 1979). La première étape consiste à admettre que le SSSD repose implicitement sur des microdonnées longitudinales. Stone (1973) donne l'explication suivante :

> «Bien sûr, si l'on recueille les statistiques au moyen d'un ensemble de registres compatibles reliés entre eux ou, mieux encore, au moyen d'un jeu de données générales et individuelles, continuellement mises à jour (*à savoir microdonnées longitudinales)*, toute la question d'un ordre séquentiel (*c.-à-d. représentation des données en fonction de périodes de temps finies, chaînes du premier ordre de Markov)* n'a plus grande importance puisque l'information peut être combinés de toutes les façons voulues dans une vaste base de données informatisée. Néanmoins, s'il est possible que de telles méthodes de collecte des données statistiques voient le jour dans l'avenir, on n'y recourt pas encore pour l'instant, à de très rares exceptions près, si bien qu'il est sensé d'aborder la systématisation des statistiques sociales à partir de méthodes de collecte qui nous sont plus familières.» [TRADUCTION] (p.152, c'est nous qui écrivons en italique)

Vu sous cet angle, nous voici dans l'avenir. Inutile donc désormais de se plier aux contraintes de l'algèbre matricielle de Stone, des hypothèses restrictives du premier ordre de Markov et des «méthodes de collecte des données qui nous sont familières».

La deuxième étape ne fait que pousser plus loin l'idée de la création d'une base de données intégrée (BDI) synthétique par les méthodes de rapprochement statistique, énoncée si clairement il y a près de vingt ans par le groupe de travail de l'ONU sur les BDI (ONU, 1979). Les membres de ce groupe avaient reconnu la grande utilité des microdonnées à très grand nombre de variables et les limites pratiques à la collecte directe de telles données. Ils avaient donc recommandé que les microdonnées souhaitées soient synthétiques, même s'il fallait pour cela recourir à des enregistrements artificiels. Dans ces premiers travaux sur les BDI, on parlait généralement de microdonnées *transversales*.

Les microdonnées *longitudinales* synthétiques forment le trait d'union entre ces deux grands courants de pensée -- un modèle similaire au SSSD de Stone articulé sur la dynamique de l'analyse longitudinale et le rapprochement statistique artificiel de microdonnées. La différence est que la création de microdonnées longitudinales synthétiques exige plus que des techniques de rapprochement statistique, qui se prêtent mal à la combinaison de séries de microdonnées longitudinales disjointes. Les microdonnées longitudinales synthétiques doivent plutôt être générées par un modèle recourant à la microsimulation dynamique (idée qui, elle non plus, ne date pas d'hier; lire Ruggles, 1981). En un mot, ce ne sont pas les particularités des observations individuelles que l'on «apparie» dans les séries de données longitudinales, mais les tendances qui ressortent du comportement dynamique des séries d'observations dans chaque ensemble de microdonnées (ainsi qu'on le constatera plus loin).

Par ailleurs, la synthèse du microcosme par la microsimulation signifie qu'il ne faudra pas débourser grand-chose pour répondre aux questions du genre «et si ? ». Ainsi, une fois qu'on aura bâti une table de survie, calculer la réduction de l'espérance de vie attribuable à une cause précise ne nécessitera que très peu de travail supplémentaire. La construction du microcosme par microsimulation crée une situation analogue. Dès qu'on aura investi dans la genèse artificielle du microcosme «de base», la synthèse de «variantes» de ce microcosme s'avère relativement simple.

Enfin, comme on se rendra compte dans la description qui suit, cette approche microanalytique du cycle de vie signifie qu'on n'a plus besoin de choisir entre une comptabilité sociale chronologique ou démographique, ainsi qu'il en était question dans Juster et Land (1981). L'approche élaborée dans le présent document couvre ces deux aspects.

3. Implications à l'égard de la collecte des données

Le fait d'envisager un modèle de statistique socio-économique dans les grandes lignes décrites précédemment a d'importantes conséquences sur les aspects conceptuels et opérationnels des mécanismes de collecte des données pour les organismes nationaux chargés d'une telle mission. Ces conséquences pourraient néanmoins ne pas être très onéreuses (comparativement à ce que coûte la collecte des données primaires) et, dans la plupart des cas, sont relativement bénignes:

- les procédés de collecte des données ne peuvent être fait «sur mesure», ni être isolés les uns des autres;
- l'existence de définitions et de principes communs (à savoir, définitions identiques du niveau de scolarité et des méthodes servant à l'établir) constitue une forme de coordination entre les méthodes de collecte des données;
- un autre type de coordination consiste à veiller à ce que les données se chevauchent de la manière appropriée, soit à prévoir la nécessité d'un rapprochement statistique artificiel (ou de méthodes équivalentes);
- une microanalyse des données brutes s'avère beaucoup plus exigeante qu'une agrégation au niveau de la qualité des données.

En fait, tout cela signifie que les systèmes de collecte des données doivent être planifiés conjointement et que les normes applicables à la qualité des microdonnées doivent être plus sévères.

Une planification conjointe n'est pas une nouvelle exigence. L'élaboration du SCN a elle aussi exigé une certaine coordination au niveau de l'acquisition des données, ne serait-ce que pour s'assurer qu'on couvrait tous les secteurs de l'économie d'une manière quelconque. Cette forme de coordination coûte toutefois considérablement moins cher que celle nécessitée par la microsimulation, parce qu'il est possible d'éliminer les incohérences des systèmes de collecte des données révélées par le SCN à un palier «macroscopique». On modifie de vastes agrégats, sans se soucier si les changements suscitent d'autres incohérences entre divers agrégats du SCN et les microdonnées originales. Pour la microsimulation, par contre, la cohérence interne entre les séries de données originales au niveau microscopique est capitale.

La nécessité de données de qualité à l'échelon microscopique n'est pas neuve. On se heurte surtout à cette difficulté chaque fois qu'il faut créer un ensemble de microdonnées pour l'usage public. Sachant que les utilisateurs examineront et analyseront les données à la loupe (pour étudier les valeurs «aberrantes» de la régression, par exemple), ces dernières font l'objet de modifications et d'imputations extensives. Les fichiers de microdonnées sur le recensement de la population soulèvent des difficultés analogues quant à la qualité des données au niveau microscopique, bien que ces difficultés soient moindres, car même si le public ne peut les consulter, ces fichiers peuvent faire l'objet de demandes générales spéciales pour des tableaux de corrélation.

La qualité des microdonnées deviendra encore plus préoccupante avec un modèle de micro-analyse intégrée comme celui que nous allons décrire. S'assurer que chaque élément d'un ensemble de microdonnées est plausible et cohérent à l'interne par un processus de correction et d'imputation est une chose, veiller à ce que la totalité de multiples jeux de microdonnées soient cohérents est bien différent (par exemple s'assurer qu'une enquête sur la santé et une autre sur les incapacités donnent les mêmes

distributions d'incapacités en fonction de leur gravité selon l'âge et le sexe ou qu'une enquête longitudinale sur la dynamique du travail procure une estimation transversale de la participation de la population active qui concorde avec celle obtenue dans le cadre de l'enquête principale sur la population active, voire qu'une série chronologique de données administratives sur le nombre d'inscriptions dans les écoles coïncide avec les données du recensement sur le niveau de scolarité, selon l'âge et le sexe).

Pareille exigence de cohérence réciproque met en relief un problème soulevé par Wilk (1987), à savoir la faiblesse relative des méthodes statistiques pour résoudre les erreurs qui ne résultent pas de l'échantillonnage. Ainsi, le refus de répondre ou l'existence d'un biais quelconque lors des enquêtes auprès des ménages entraîne habituellement une sévère sous-déclaration de certaines sources de revenu. En général, les méthodes de correction et d'imputation classiques ne résolvent ce problème qu'en partie (c.à-d. on ne change pas les sources du revenu faussement signalées comme inexistantes). Les chercheurs qui élaborent les modèles de microsimulation pour les politiques fiscales/de transfert sont les seuls à avoir résolu ce problème, par nécessité (Citro et Hanushek, 1991; Bordt et al., 1990; Wolfson et al., 1989). En outre, la correction des données recueillies auprès des ménages ne présente à toutes fins pratiques aucune utilité pour l'arrondissement des réponses (à savoir, mentionner le revenu à la centaine ou au millier du dollars le plus près), même si on possède la preuve qu'un tel comportement de la part des répondants fausse autant certaines statistiques (à savoir les quantiles) que l'erreur classique d'échantillonnage (Rowe et Gribble, 1994).

Enfin, on admet de plus en plus l'importance des enquêtes longitudinales, de toute évidence essentielles à une description de la dynamique et au déchiffrage des enchaînements de causalité. Pour utiliser les microdonnées longitudinales à ces fins, on devra recourir à des méthodes d'inférence plus complexes que celles dont se servent couramment les organismes qui s'occupent de statistiques, notamment la régression des risques plutôt que les tableaux de corrélations. Il pourrait s'ensuivre un examen encore plus critique des données.

4. Le projet LifePaths

Nous allons maintenant passer à une illustration des points généraux qui précèdent. Le projet LifePaths est un projet ayant pour but la construction d'un modèle expérimental de statistique socio-économique. Ce projet se poursuit sous la direction de Statistique Canada au nom du ministère du Développement des ressources humaines du Canada, le nouveau «superministère» responsable, entre autres, du bien-être social, des pensions, de l'assurance-chômage et des politiques sur le marché du travail.

L'objectif fondamental du modèle LifePaths est de produire une série de vues multiples mais cohérentes de la situation socio-économique des Canadiens. Le modèle a été conçu en fonction des caractéristiques générales que nous venons de voir, soit la capacité de dégager la tendance générale, de refléter la diversité et de répondre aux questions du genre « et si ? ». Plus concrètement, le projet porte sur le temps que les Canadiens consacrent à diverses activités comme le travail, l'éducation, la famille, les programmes gouvernementaux et les loisirs.

La généralisation des tables de survie pour la population active est l'une des vues ou l'un des aspects principaux envisagés. Le tableau 1, par exemple, présente non seulement l'espérance de vie classique des cohortes de Canadiens de sexe masculin nés à différentes années, mais aussi l'âge moyen auquel ces sujets entreront dans la population active rémunérée et en sortiront. En examinant une série de cohortes de naissances (période), chacune représentant des décennies qui se succèdent, l'analyse illustre très clairement certaines tendances à long terme, soit consacrer plus de temps aux études, prendre sa retraite plus tôt et travailler généralement moins longtemps, dans le cas des hommes. La dernière colonne révèle les effets de ces tendances sur le coût des régimes de pension publics. (Notons que malgré leur ancienneté relative, ces données brossent apparemment le tableau estimatif le plus récent de la vie active).

Tableau 1 - Espérance de vie et espérance de vie active des hommes de 15 ans (cohortes historiques constantes)

Âge	entrée dans la population active	âge moyen à la retraite	âge au décès	années de travail	années de retraite	années de travail par année de retraite
1921	16,5	63,7	67,6	47,2	3,9	12,1
1931	17,0	64,0	68,4	47,0	4,4	10,7
1941	17,2	64,1	69,1	46,9	5,0	9,4
1951	17,5	63,9	70,4	46,4	6,5	7,1
1961	18,2	64,0	71,2	45,8	7,2	6,4
1971	19,8	63,3	71,3	43,5	8,0	5,4

Source: Gnanasekaran et Montigny (1975) et Wolfson (1979)

Le projet LifePaths étend les résultats de ce tableau fondamental sur la vie active dans plusieurs directions. Il approfondit l'analyse des tendances annuelles relatives au travail sans se limiter à une simple division entre années de vie active et inactive. Ainsi, il tient compte du travail à temps partiel, de la prolongation des congés payés et des vacances, de la modification du nombre typique d'heures de travail par semaine, des périodes de chômage ou de sortie de la population active de moins d'un an, des périodes où les personnes travaillent tout en poursuivant leurs études et de l'intérêt croissant pour le travail autonome. De plus, les aspects temporels du travail sont combinés à ses aspects économiques, notamment au revenu.

D'autres grandes catégories d'activités entrent aussi en ligne de compte. L'une d'entre elles est la poursuite des études; une autre, le milieu familial (à savoir, le fait de vivre seul ou avec d'autres membres de la famille). On peut dire que le modèle couvre les interactions avec les grandes institutions sociales, soit le travail, l'école et la famille. Le modèle LifePaths réunit donc les séquences «actives» (études et travail rémunéré) et «passive» (suite de groupements familiaux auxquels un individu adhère durant le cours de sa vie, p. 145), comme le dit Stone (1973) dans sa proposition de comptabilité démographique pour le SSSD, combinaison difficile à réaliser en pratique avec les méthodes matricielles qu'utilise cet auteur.

Le modèle planifie la participation à quelques grands programmes sociaux, entre autres les prestations d'assistance sociale (AS), d'assurance-chômage (AC) et d'indemnisation pour les accidents du travail (AT). En règle générale, on tient aussi mieux compte de l'emploi du temps, grâce aux données des enquêtes pertinentes - le régime de comptabilité du temps proposé, notamment par Juster et ses collaborateurs (1981). Les principales catégories d'activité n'intègrent donc pas seulement le travail et les études, mais aussi les tâches domestiques non rémunérées, l'hygiène personnelle, les soins dispensés à autrui, le sommeil, les déplacements urbains, la télévision, les autres loisirs passifs, les loisirs actifs, l'interaction avec les membres de la famille et d'autres formes de socialisation.

Le modèle LifePaths englobe toutes ces activités humaines, en adoptant un point de vue qui épouse le cycle de vie dans son entièreté, d'une manière cohérente et intégrée - si bien qu'on combine et enchasse les approches de comptabilité chronologique et démographique qu'analysent Juster et Land (1981). L'élaboration du modèle LifePaths est un exercice de haute voltige et les résultats présentés ici doivent être considérés comme expérimentaux.

Côté méthodologie, le modèle LifePaths innove sur plusieurs plans.

En premier lieu, aucun ensemble de données particulier ne renferme toute l'information requise, par exemple des données détaillées sur les activités humaines sous les angles économique et social. Les ensembles existants et ceux qui pourraient s'y ajouter ne sont que partiels et fragmentaires. En outre, tel qu'indiqué précédemment, pour des raisons d'ordre pratique (coût, fardeau pour les répondants et protection

de la vie privée), on ne pourra jamais recourir aux données entièrement intégrées des enquêtes sur les ménages. On devra inévitablement appliquer des processus d'intégration synthétique à des ensembles de données multiples.

Deuxièmement, le modèle doit couvrir le cycle de vie complet des sujets. Le faire avec les données longitudinales dont on dispose actuellement exigerait des décennies d'enquêtes subséquentes au terme desquelles bon nombre de choses auront changé. L'idée fondamentale consiste donc à généraliser le concept de l'espérance de vie pour la *période* et la table de survie sous-jacente. Par conséquent, l'analyse concentrera sur des cohortes réalistes mais hypothétiques.

Un troisième élément de l'objectif primaire a trait à l'expression détaillée de la diversité ou de l'hétérogénéité de la population, donc la capacité d'observer un phénomène distributif comme la répartition inégale du revenu. Une telle capacité exige comme base des microdonnées explicites. Puisqu'on ne peut recueillir de données sur la vie réelle d'un échantillon représentatif d'individus, les microdonnées nécessaires doivent être artificielles. Elles doivent cependant être assez réalistes pour qu'on ne puisse pas vraiment les distinguer des ensembles partiels de caractéristiques provenant des données réelles sur des échantillons de la population, notamment des enquêtes longitudinales.

Toutes ces contraintes signifient que le modèle LifePaths doit avoir pour âme un modèle de microsimulation, en d'autres termes, un aperçu réaliste mais artificiel de la vie de l'individu.

5. Les données synthétiques

Avant de vous présenter nos résultats initiaux, il est important d'expliquer dans quelle mesure le modèle LifePaths repose sur des données synthétiques et comment les résultats synthétiques reflètent raisonnablement la réalité.

Un être humain vit généralement 75 ans, environ. Toutefois, face à la rapidité relative avec laquelle évoluent maintes activités humaines, procéder à des observations socio-économiques cohérentes et soutenues pendant un tel laps de temps est à toutes fins pratiques irréalisable. Les statistiques reconnnues depuis des décennies n'existaient tout simplement pas il y a 75 ans (qu'on songe au taux de chômage, au PIB par habitant et aux mesures concernant les loisirs). De même, il se peut fort bien que dans 75 ans d'ici, en 2070, les statistiques fondamentales, dont l'importance est acquise de nos jours, soient remplacées par d'autres valeurs que l'on peut à peine imaginer aujourd'hui.

Pourtant, les indicateurs statistiques qui reflètent les processus couvrant la vie humaine dans son entièreté suscitent beaucoup d'intérêt. Le plus connu est sans doute l'espérance de vie. Néanmoins, il en existe d'autres comme la proportion de mariages qui devraient se terminer par un divorce, le nombre d'emplois différents qu'une personne pourrait connaître durant sa carrière professionnelle, la pertinence ou non des pensions publiques face aux revenus antérieurs à la retraite et la partie de la vie moyenne qu'une personne passe en santé ou malade. Il existe manifestement des indicateurs qui s'appliquent à la vie humaine, et ceux-ci sont plus ou moins largement acceptés. Le modèle LifePaths les généralise.

Même si on ne le l'admet pas de manière générale, l'espérance de vie est une statistique «artificielle». En un sens, on peut la comparer à une déclaration que l'on ferait sur la destination d'une automobile dont on connaît la position et la vitesse, mais pas l'accélération. L'espérance de vie repose sur le taux de mortalité spécifique selon l'âge (et le sexe). Comme c'est le cas pour la vitesse du véhicule, elle s'appuie donc sur des données réelles. Cependant, l'espérance de vie (période) s'applique à un individu hypothétique qu'on retire du passage du temps et qui passera le reste de sa vie exposé au taux de mortalité en vigueur au début des années 1990. Bref, on ne tient pas compte de l'accélération positive ou négative du taux de mortalité.

On sait, bien sûr, que les taux de mortalité ont généralement diminué au cours des dernières décennies et on s'attend largement à ce que cette tendance se poursuive. Par conséquent, bien qu'en soi elle néglige la tendance suivie par les taux de mortalité, l'espérance de vie, par les tendances qui lui sont propres, nous renseigne fort commodément sur les changements des taux de mortalité sous-jacents, car elle suit dans le temps une forme de moyenne pondérée du taux de mortalité spécifique selon l'âge (et le

sexe). Évidemment, on pourrait toujours examiner le taux de mortalité spécifique selon l'âge sous-jacent, mais essayer de saisir l'évolution ne serait-ce que d'une centaine de valeurs est une tâche passablement complexe. (Or, ces valeurs deviennent beaucoup plus nombreuse dès qu'on répartit les taux de mortalité selon le sexe, l'état civil et l'âge.) L'espérance de vie garde néanmoins son utilité comme indicateur, précisément parce qu'elle comprime des centaines de données en un indicateur intuitivement accessible, indicateur dont la variation dans le temps épouse raisonnablement celle des taux de mortalité spécifiques selon l'âge sous-jacents.

Le modèle LifePaths est concu pour être entièrement analogue. Néanmoins, il repose sur une multitude de processus et de descriptions statistiques des états marquant la transition de l'individu entre différents stades de la vie. Par exemple, outre la mortalité, le modèle tient explicitement compte des paramètres démographiques comme l'état civil et les transitions connexes que sont le passage du célibat au concubinage ou au mariage, voire la rupture du couple par la séparation ou le divorce. Pareillement, on a tenu compte d'autres classifications de la situation socio-économique, notamment le fait de travailler ou de poursuivre des études, en fonction des données réelles sur les taux récents de distribution et de transition pertinents.

Pour réussir une telle généralisation de l'espérance de vie, on a dû généraliser le concept sous-jacent de la table de survie. Le degré de précision le plus élevé de la table de survie est le groupe de sujets - par exemple défini par le sexe et l'âge. On présume que les membres du groupe sont homogènes. Pareil degré de précision ne suffit pas pour le modèle LifePaths. En effet, on doit explicitement tenir compte des sujets hétérogènes que caractérisent de nombreux attributs si l'on veut effectuer le meilleur usage du modèle et illustrer aussi précisément que possible les résultats de l'analyse des schémas de comportements dynamiques observés avec les vastes ensembles de microdonnées longitudinales.

En un sens, non négligeable, tout cela signifie que le modèle LifePaths donne des résultats beaucoup plus réalistes que l'espérance de vie obtenue avec la table de survie classique. Ainsi, le modèle ventile les taux de mortalité d'après l'état civil, plus l'âge et le sexe. À son tour, l'état civil repose sur un ensemble complexe de facteurs comme le niveau de scolarité, les antécédents de fécondité et la durée des séjours au sein de la population active.

D'un autre côté, étant donné l'aspect artificiel des «données», le modèle produira inévitablement des résultats plus explicites qu'une table de survie. En effet, alors que la table de survie classique s'appuie sur un groupe de sujets, ces derniers demeurent implicites en soi - on se borne à compter le nombre d'individus de la cellule ou de la catégorie, soit selon l'âge et le sexe. Le modèle LifePaths, quant à lui, tient explicitement compte de la vie de chaque élément.

Dans ce cas, quel sens devrait-on accorder à un résultat du modèle, par exemple la ventilation de l'espérance de vie entre le nombre d'années qu'une personne peut s'attendre à consacrer au travail et aux études ? Un tel résultat devrait donner une interprétation analogue à celle de l'espérance de vie traditionnelle - c'est-à-dire constituer une sorte de sommaire du taux démographique récent. Les résultats du modèle LifePaths illustrent ce qui se produirait si les taux de transition les plus récents entre différents états socio-économiques (dépendant des attributs des sujets hétérogènes) demeuraient constants.

6. Résultats initiaux

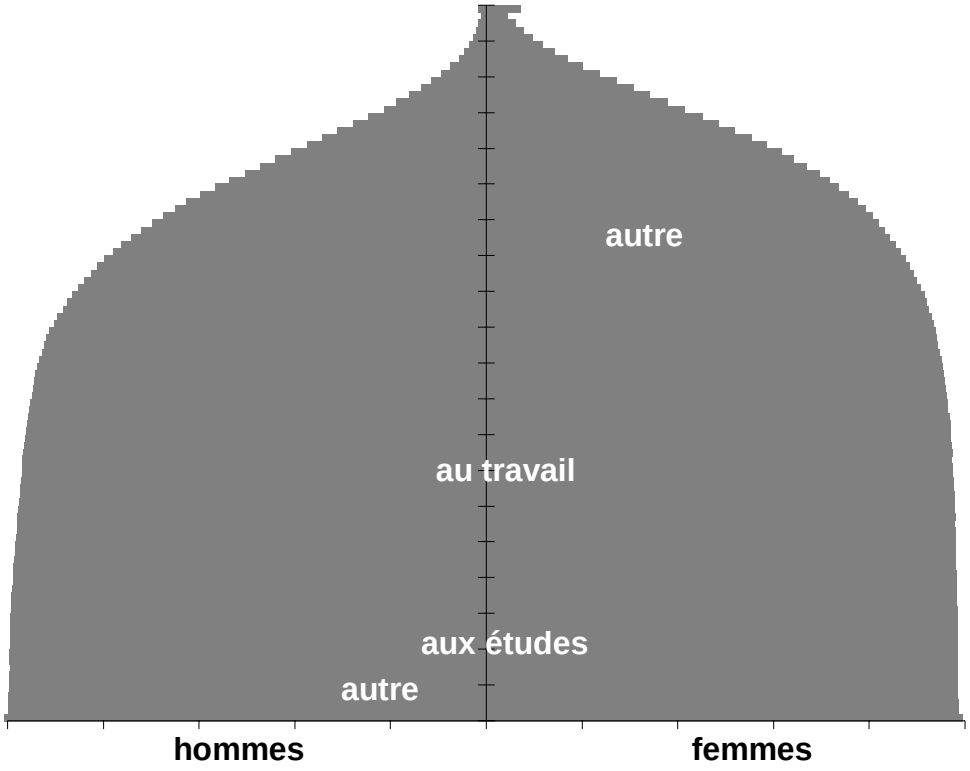
Le modèle LifePaths se compose essentiellement d'un échantillon du cycle de vie complet (synthétique) de plusieurs sujets. Cette base de microdonnées longitudinale articulée sur un échantillon de vies est malheureusement beaucoup trop complexe pour permettre une analyse directe. C'est pourquoi nous ne présenterons ici que quelques «vues» sommaires du microcosme sous-jacent, vues obtenues dans la plus pure tradition d'une analyse démographique.

Soulignons que ces «vues» se limitent à des indicateurs scalaires comme le PIB; elles dévoilent simultanément plusieurs paramètres démographiques fondamentaux. Néanmoins, on ne doit pas voir là une lacune, comme l'impossibilité de passer à une mesure générale unique avec le SCN. Chaque «vue» doit plutôt être perçue comme une illustration de la puissance des logiciels d'infographie contemporains, qui permettent une évaluation plus poussée de l'économie sociale qu'un simple indice.

Débutons par une des représentations démographiques les plus simples, la pyramide des âges. La figure 1 montre la pyramide des âges d'une table de survie de base. Le nombre de sujets de sexe féminin apparaît à droite, sur l'axe horizontal, et le nombre de sujets de sexe masculin, à gauche, jusqu'à l'âge de 100 ans, indiqué sur l'axe vertical. Un tel diagramme repose sur les probabilités de transition pour la *période* (fin des annéees 1980 et début des années 1990), sur lesquelles on reviendra. Ainsi qu'on peut s'y attendre, la courbe de survie des femmes diminue plus lentement que celle des hommes quand l'âge augmente, illustrant l'espérance de vie supérieure des femmes (ou plus exactement la raison à l'origine de ce phénomène). (La coche à 99 ans est attribuable à l'intervalle, qui correspond à ≥99 ans.)

La figure 1 répartit la même population en trois catégories socio-économiques - soit les «travailleurs», les personnes aux «études» et les «autres». La vie d'un «étudiant» débute avec la première année, de telle sorte que les enfants en garderie et à la maternelle font partie du groupe «autres». Puisque le modèle LifePaths suit chaque sujet tout au long de sa vie, on a dû prendre certaines décisions arbitraires lorsqu'une personne poursuit plusieurs activités la même année. Plus exactement, pour qu'une personne fasse partie des «travailleurs», elle doit travailler au moins 15 heures par semaine et consacrer la majeure partie de l'année au travail, au même rythme. Une personne qui passerait donc 5 mois à étudier, 4 à travailler au moins 15 heures par semaines et les trois derniers mois à travailler hebdomadairement moins de 15 heures (ou pas du tout) se retrouverait dans le groupe «aux études» cette année-là; si on substitue les périodes de 5 et de 4 mois cependant, elle compterait parmi les personnes «au travail». (L'utilisateur du modèle a tout pouvoir sur ces définitions.) Le diagramme révèle qu'à de rares exemptions près, chacun poursuit des études à l'âge de 8 ans et que quelques sujets commencent à quitter le milieu scolaire à 16 ans; que la majorité des gens ont terminé leurs études à 20 ans, mais qu'une poignée d'hommes et de femmes les poursuivent au-delà de la vingtaine.

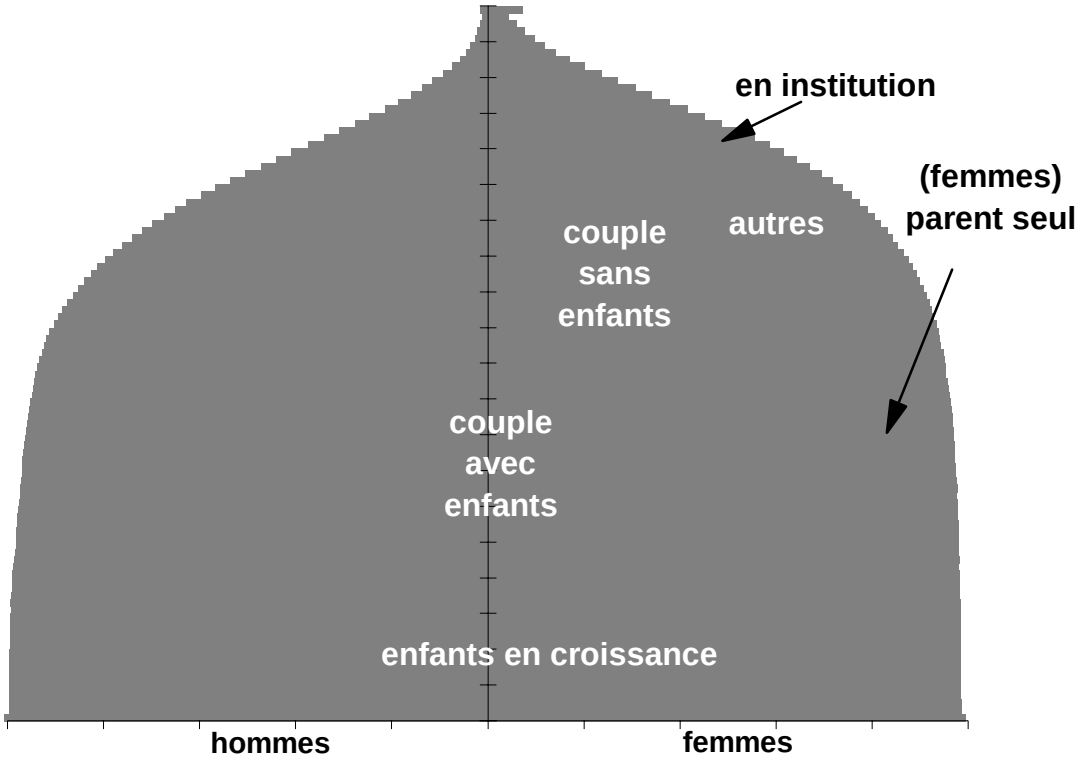
Figure 1 - Population du modèle LifePaths (années-personnes) selon le type d'activité, l'âge et le sexe



Personne ne semble passer directement des études au travail, mais nous en reparlerons en examinant un autre diagramme. On notera peut-être une proportion surprenante de sujets dans le groupe «autres», qui comprend les personnes sans emploi et celles qui ne font pas partie de la population active (à savoir, ménagères, pensionnés). Comme c'était prévisible, toutes proportions gardées, la probabilité que les hommes travaillent à un âge quelconque est plus élevé que pour les femmes. La courbe des femmes de 20 à 25 ans détenant un emploi s'affaisse malgré la tendance relative à une plus grande participation à la population active, car cette période représente les principales années de procréation. Ensuite, la courbe augmente légèrement entre 25 et 35 ans. Dans le cas des hommes, le taux de participation diminue considérablement pour le groupe des 60 à 65 ans.

La figure 1 correspond à la «séquence active» dont parle Stone (c'est-à-dire le passage des études au travail), alors que la figure 2 donne un apercu de la «séquence passive». La figure 2 reprend la pyramide des âges de la figure 1 et se rapporte exactement à la même population synthétique sous-jacente du modèle LifePaths, mais ventile les individus en fonction d'une autre dimension, soit l'état civil. Par définition, tous les sujets de moins de 18 ans sont des «enfants» à moins qu'ils soient mariés ou aient eux-mêmes un enfant. Lorsqu'un couple se sépare, on suppose que les enfants restent avec leur mère. Cette hypothèse explique pourquoi il y a des parents seuls de sexe féminin mais pas de sexe masculin. (Les versions ultérieures intégreront des données plus réalistes sur les ententes de garde.)

Figure 2 - Population du modèle LifePaths (années-personnes) selon l'état civil, l'âge et le sexe



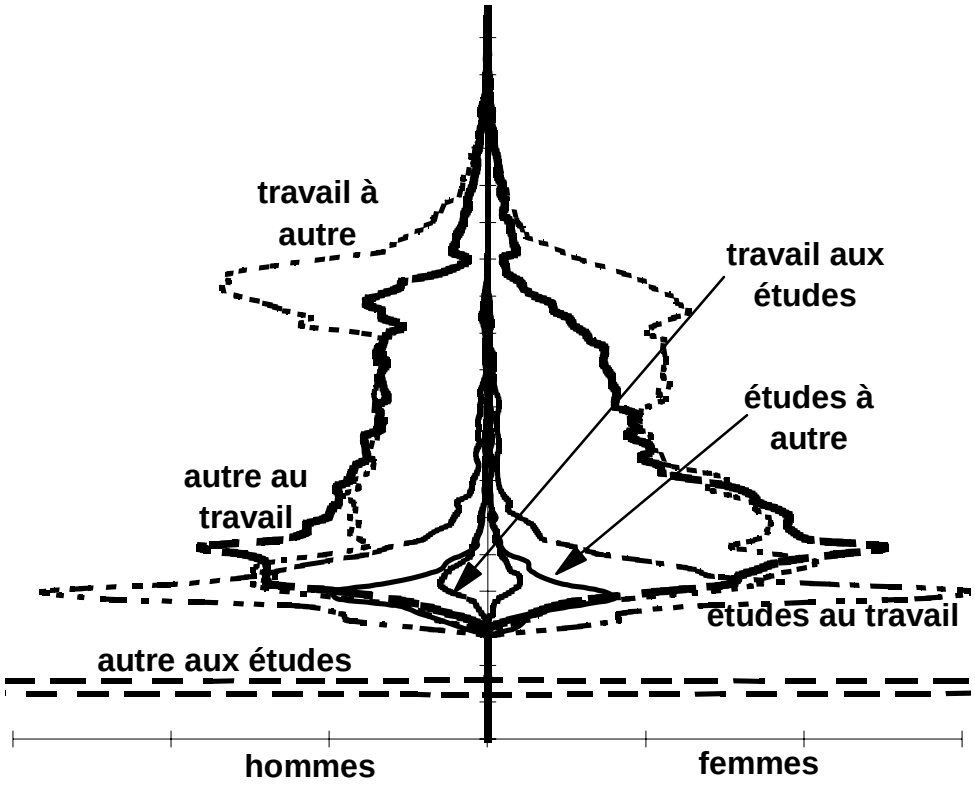
Quand on compare la courbe des gens mariés (couples avec ou sans enfants) des deux sexes, on constate que celle des hommes est décalée de quelques années vers le haut. Ce résultat reflète la tendance générale selon laquelle le mari a quelques années de plus que son épouse. Le diagramme révèle aussi qu'il existe plus de veuves que de veufs. On peut y voir la conséquence de la différence d'âge moyenne positive entre mari et femme, et de la plus grande espérance de vie des personnes de sexe féminin. Enfin, le diagramme montre que beaucoup de femmes vivent en institution (principalement des foyers de soins infirmiers ou des établissements de traitement des maladies chroniques), encore une fois en raison de leur plus grande longévité et de la plus forte prévalence des problèmes de santé à un âge avancé,

doublées au fait que les hommes souffrant d'une incapacité similaire ont souvent une épouse en mesure de prendre soin d'eux à la maison.

Les figures 1 et 2 ne donnent qu'un aperçu des possibilités du modèle LifePaths, des «vues» (tableaux de corrélation dans le cas présent) du microcosme complet sous-jacent, soit un ensemble de microdonnées longitudinales pour une cohorte de naissances artificielle du «début des années 1990». On peut classer le même ensemble de microdonnées longitudinales sous-jacent de façon à obtenir la figure 3 indiquant le *passage* entre divers états, plutôt que la *population* de chaque état. Dans ce cas, le graphique de la figure 3 reproduit les passages d'un état à l'autre pour les sujets de la figure 1. L'axe horizontal correspond au nombre de personnes qui passent par telle ou telle transition chaque année, une fois de plus sous forme de pyramide des âges. L'axe vertical commun donne l'âge, les femmes se retrouvant à la droite, sur l'axe horizontal, et les hommes, à la gauche. (Dix-huit pour cent de la population se retrouvent aux extrémités de l'axe horizontal, si bien qu'une cohorte de 100,000 sujets autorise jusqu'à 9 000 transitions aussi bien pour les hommes que pour les femmes, par année).

La première transition se fait du groupe «autre» (petite enfance ou pré-maternelle) à celui «aux études». La figure 1 révèle que tous les enfants de sexes masculin et féminin effectuent ce passage à l'âge de 6 ou de 7 ans. La grande transition suivante survient au terme des «études», le passage au groupe des «travailleurs» atteignant son point culminant vers l'âge de 20 ans, aussi bien pour les hommes que pour les femmes. Un plus petit nombre de sujets, qui atteint également un sommet vers 20 ans, accomplit la transition entre les «études» et une «autre» activité. On se rappelera que cette dernière catégorie correspond aux années-personnes qui n'ont consacré la majeure partie de leur année (soit légèrement plus du tiers) ni aux études, ni à un travail les occupant plus de 15 heures par semaine.

Figure 3 - Transitions de la population du modèle LifePaths (années-personnes), selon l'âge et le sexe

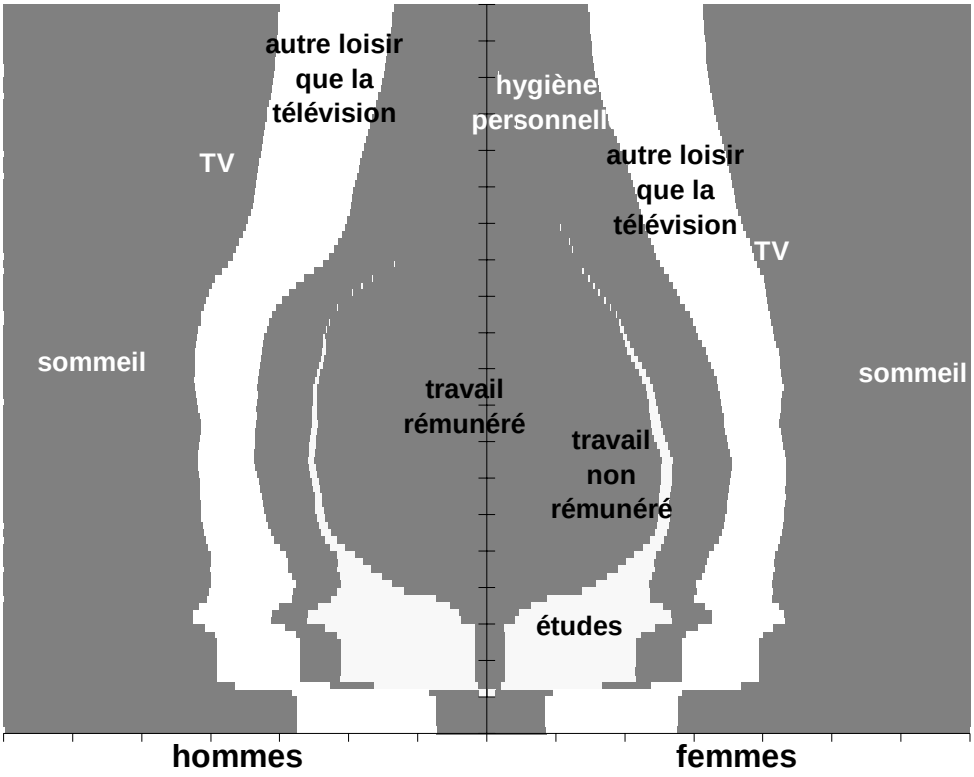


Du début de l'âge adulte à la soixantaine, les principales transitions surviennent entre les stades «travail» et «autre». Précisons que les passages correspondent à une valeur brute et non à une valeur nette. On remarque d'ailleurs que le passage net entre les catégories «travail» et «autre» (établi par comparaison des transitions brutes) change de direction vers le groupe «autre» pour les sujets de sexe féminin de 40 à 45 ans, alors qu'il ne varie pratiquement pas pour les hommes jusqu'à l'âge de 50 ans. Enfin, un nombre maximal de personnes prennent leur retraite lorsqu'elles atteignent de 55 à 65 ans, le pic étant plus prononcé dans le cas des hommes.

Outre le nombre de personnes recensées dans les différentes catégories d'activité et le nombre de sujets passant d'un groupe à l'autre, le modèle LifePaths permet de produire des tableaux montrant la durée de séjour, c'est-à-dire le temps que les individus passent dans tel ou tel stade de vie. Le tableau 1, plus haut, tendait déjà dans cette direction puisqu'il donnait une première estimation de l'espérance de vie active. Une autre capacité du modèle, d'une grande importance compte tenu des microdonnées qui en forment implicitement la base, concerne la production de distributions à une ou deux variables du séjour pour tel ou tel groupe - par exemple la distribution combinée du nombre d'années passées aux études et au travail pour les hommes et les femmes. (Faute d'espace nous ne pourrons présenter ces diagrammes.)

La figure 4 présente une autre image obtenue grâce à la simulation de base du modèle - il s'agit cette fois d'une classification des activités selon un axe horizontal différent. Au lieu d'indiquer le nombre d'années-personnes tirées de la table de survie d'une cohorte de naissances pour une période donnée, comme les figures 1 et 2, la figure 4 prend comme axe horizontal les activités principales, soit le nombre d'heures consacrées à chacune d'elles pendant une semaine normale (168 heures), selon le sexe et l'âge. Ainsi, on remarque que les hommes passent en moyenne près de 40 heures par semaine au travail entre l'âge de 30 à 55 ans.

Figure 4 - Emploi du temps (nombre d'heures moyen par semaine) selon le modèle LifePaths, par grande activité selon l'âge et le sexe



En surface, la figure 4 imite exactement les données que l'on pourrait obtenir directement d'une enquête sur l'emploi du temps. Sur le plan de la validation, les données des deux sources devraient correspondre étroitement. Pourtant, ce diagramme a été produit par simulation au moyen du modèle LifePaths et diffère quelque peu des données de l'enquête sur l'emploi du temps sous-jacente, principalement parce qu'on a accru la cohérence des données. Ainsi, le taux de participation annuel à la population active selon l'âge et le sexe de la figure 4 est cohérent avec les taux sous-entendus à la figure 1, de même qu'avec les tendances démographiques de la figure 2, en raison de la façon dont le modèle est construit.

Une des impressions qui se dégage du diagramme est qu'en moyenne, les hommes et les femmes passent une partie relativement faible de leur vie à travailler contre rénumération - domaine d'élection du SCN. Quand on l'examine sous l'angle du nombre moyen d'heures qu'on y consacre chaque semaine (plutôt qu'en fonction du fait qu'on travaille plus de 15 heures par semaine pendant plus du tiers de l'année, comme à la figure 1), on se rend compte que le travail ne monopolise qu'une très petite fraction de la vie (la vie éveillée, s'entend). Bien sûr, le travail non rémunéré, de même que les aspects de l'hygiène personnelle et des loisirs, présentent aussi une grande importance pour l'économie, mais le SCN n'en tient pas compte au-delà de la valeur monétaire agrégative du taux de consommation personnel par produit.

La même figure révèle les limites des ratios de dépendance démographiques classiques - qui reposent sur l'utilisation du nombre brut de personne d'âge productif (à savoir de 20 à 64 ans) comme dénominateur. L'examen de la figure 4 révèle que ces ratios sous-estiment manifestement la dépendance économique de nombreuses personnes face à la société. Le même diagramme donne à penser qu'il faut reproduire de façon plus explicite les mécanismes qui permettent au reste de la population d'acquérir son pouvoir d'achat, surtout par le temps consacré au «travail productif». Ces mécanismes comprennent les transferts intra-familiaux ainsi que les programmes fiscaux et de transfert des gouvernements. D'une manière plus générale, en combinant l'emploi du temps aux paramètres démographiques plus classiques, le modèle LifePaths permet de construire une série cohérente de tableaux statistiques qui reflètent beaucoup mieux l'activité sociale et économique.

Les diagrammes LifePaths comme celui de la figure 4 montrent clairement que la vie ne se résume pas à ce que le SCN permet de capturer avec son orientation économique. Il s'ensuit qu'une publication régulière des résultats statistiques de ce genre pourrait avoir une incidence appréciable sur l'élaboration des politiques publiques, car on replacerait les paramètres économiques dans un cadre plus général et attirerait l'attention sur les effets beaucoup plus étendus des politiques en matière de chômage, retraite, redistribution du revenu, éducation , garde des enfants, désinstitutionalisation et semaine de travail, pour n'en citer que quelques-unes.

Soulignons encore une fois que les résultats du modèle LifePaths ont toujours essentiellement une valeur d'illustration. La base de microdonnées longitudinales synthétique sous-jacente exige encore des améliorations. Comme on le verra dans la partie qui suit, ces données reposent sur une série d'enquêtes et d'analyses récentes, donc sur des données réelles, mais les analyses concernées ne portent toujours en partie que sur des résultats préliminaires.

7. Méthodes sous-jacentes

Le modèle LifePaths que l'on vient d'illustrer exploite surtout deux séries de données récentes et les résultats de près d'une décennie de recherches sur les modèles de microsimulation connexes. Les deux séries de données précitées sont celles de l'Enquête sociale générale (ESG) de 1992, en vertu de laquelle on a posé aux répondants des questions détaillées sur leur emploi du temps au cours des 24 dernières heures, et l'Enquête sur l'activité (EA), qui a servi à recueillir des données longitudinales détaillées sur la dynamique du marché du travail entre 1988 et 1990. Le modèle de microsimulation LifePaths est un hybride de ces deux enquêtes, du modèle de microsimulation DEMOGEN (Wolfson, 1989) adapté à l'environnement du tout nouveau logiciel de microsimualtion ModGen C++, et du nouveau modèle de calcul du remboursement en fonction du revenu applicable aux prêts pour l'éducation post-secondaire mis au point par le ministère du Développement des ressources humaines du gouvernement canadien.

La présente partie survole très rapidement les processus qui ont débouché sur la synthèse de la cohorte de naissances du modèle LifePaths, âme même du modèle. En règle générale, cette synthèse nécessite, d'une part, une architecture générale raccordée à divers mécanismes économiques et socio-démographiques et, d'autre part, une analyse détaillée des données en mesure de produire une description statistique de chaque processus de manière empirique (par exemple, la dynamique du comportement).

Comme la table de survie classique, le modèle LifePaths débute avec une population spécifique, par exemple 100 000 naissances. Contrairement à elle cependant, il suit chaque sujet toute sa vie jusqu'à la mort. (Une table de survie suit un groupe d'individus que l'on suppose homogènes.) À divers moments dans le temps, chaque sujet a la possibilité d'effectuer un changement dans sa vie (transition). Compte tenu de la série de paramètres existants, il pourrait s'agir du passage au marché du travail ou d'un changement d'état civil. Le nombre de transitions possibles dépend de la gamme d'états explicitement envisagés. Dans la version actuelle du modèle, les membres de la cohorte sont caractérisés par les attributs de base qui suivent, à chaque moment de leur vie:

- âge -- en tant que variable continue
- fécondité -- âge à la naissance des enfants, présence d'enfants au foyer familial
- état civil -- célibat, concubinage ou mariage, séparation ou divorce
- situation relative à l'emploi -- y compris, participation à la population active et type d'emploi (nombre d'heures par semaine, nombre de semaines dans l'année)
- éducation -- année et type d'établissement si le sujet est encore aux études, niveau de scolarité
- revenu -- rémunération horaire, hebdomadaire et annuelle
- emploi du temps -- catégories de la figure 4 avec désagrégation plus poussée
- participation aux programmes -- notamment assistance sociale, assurance-chômage, régime de pension public
- paramètres du conjoint -- y compris âge, niveau de scolarité, expérience du marché du travail

Ces attributs fondamentaux, permettent également d'en obtenir d'autres, secondaires, fort variés, comme les variables qui apparaissent aux figures 1 à 4.

Connaissant cette série d'attributs, nous pouvons maintenant décrire les processus qui ont servi à générer la trajectoire de chaque attribut dans le modèle. En voici une brève description.

Démographie - Dans le modèle, la fécondité est une conséquence de la conception, elle-même modélisée sous forme d'une suite de taux de probabilité finis et constants, subordonnés à l'âge, à l'état civil et au nombre antérieur de naissances vivantes. Les principales sources de données sont les enregistrements de naissances, plus l'Enquête sur la famille de 1983. De cette façon, on peut tenir compte du biais attribuable à la conception durant le célibat ou le concubinage, suivie par le mariage avant l'accouchement. Le taux de mortalité est associé à l'âge, au sexe et à l'état civil et s'appuie sur les avis de décès. Dans les deux cas, c'est le recensement de la population qui sert de dénominateur.

La formation et la dissolution du couple sont illustrées par une série de fonctions de probabilité. Partant du célibat , on note des probabilités concurrentes de passer au concubinage ou au mariage. La rupture du couple génère des probabilités de séparation et de divorce qu'on estime séparément pour les hommes et les femmes et qui dépendent des antécédents, d'une manière assez complexe. Par exemple, la «probabilité» qu'une femme s'engage dans une union est positivement corrélée à celle d'être enceinte et atteint sa valeur la plus élevée peu après l'entrée au sein de la population active. La probabilité d'une séparation est plus forte pour les femmes si le couple n'a pas d'enfants en bas âge à la maison, si la femme s'est mariée lorsqu'elle était adolescente et si elle a travaillé récemment.

Éducation - Les taux de transition illustrant le passage de l'école primaire au cours secondaire ont été bâtis de manière à être conjointement aussi cohérents que possible avec les taux de fréquentation scolaire des enfants d'âge pertinent, obtenus lors des recensements de 1986 et de 1991. Le passage aux études post-secondaires (collège, institut technique, université) repose sur les taux de probabilité estimatifs qui dérivent de l'Enquête nationale auprès des diplômés (END), des données administratives sur le nombre

d'inscriptions dans les écoles et de l'Enquête sur l'activité (EA) lorsque les jeunes quittent leur emploi pour retourner aux études et poursuivre celles-ci.

Travail - L'expérience sur le marché du travail est simulée en deux grandes étapes: le fait d'avoir ou non un emploi et la rémunération provenant de l'emploi en question. Dans le premier cas, on estime les passages à la vie active et inactive grâce à l'EA pour les hommes et les femmes pris séparément, ainsi que séparément pour un premier emploi, un second, les emplois subséquents, et l'abandon du marché du travail. L'entrée initiale sur le marché du travail est représentée par une distribution du temps d'attente, alors que les autres transitions sont illustrées par des fonctions de probabilité à variables multiples. Le sexe et le niveau de scolarité sont d'importants facteurs en ce qui concerne la période d'attente qui précède l'obtention du premier emploi. La probabilité d'une réinsertion dans la population active dépend du sexe, du niveau de scolarité et de la durée de la période courante de non-emploi; dans le cas des femmes, elle dépend aussi de l'existence d'enfants en bas âge, ce paramètre ayant un effet à la baisse supplémentaire.

La rémunération repose sur la situation relative à l'emploi décrite précédemment et sur des modèles distincts pour le nombre d'heures de travail hebdomadaires et le salaire horaire. Après l'entrée initiale sur le marché du travail, on attribue au hasard un nombre d'heures de travail hebdomadaires à partir d'une distribution selon l'âge, le sexe et le niveau de scolarité. Cette distribution s'appuie sur les données combinées de l'END, de l'EA et de l'Enquête sur les finances des consommateurs (EFC - enquête annuelle sur la distribution du revenu dans les ménages). La variable «heures hebdomadaires» est ensuite corrigée d'après l'âge, le sexe, le nombre d'heures de travail hebdomadaires de l'année antérieure et le niveau de scolarité. Chaque sujet reçoit un rang-centile pour le salaire horaire avec le nombre d'heures de travail hebdomadaires. Le taux de rémunération horaire est subséquemment «repris» en fonction de distributions selon l'âge, le sexe et le niveau de scolarité. Le rang-centile est corrigé annuellement d'après le classement ordinal qui «dérive» de l'EA.

Emploi du temps - L'Enquête sociale générale (ESG) de 1992 a permis d'interroger environ 9000 personnes, uniformément réparties selon l'âge, le sexe, la journée de la semaine et le mois de l'année, sur leur emploi du temps pendant 24 heures. Elle a aussi servi à recueillir des données de base sur le niveau de scolarité, la situation relative à l'emploi et l'état civil. Après analyse approfondie des données, on a créé un module LifePaths qui impute à chaque journée-personne simulée un vecteur représentant le temps consacré à chaque activité durant une période de 24 heures, y compris au niveau d'agrégation le plus élevé des catégories indiquées à la figure 4. (On a formulé des hypothèses spéciales pour les enfants de moins de 15 ans et les personnes agées résidant en institution car ils n'étaient pas touchés par l'ESG.)

L'analyse statistique révèle que l'âge, le sexe, le jour de la semaine, l'état civil, l'existence de jeunes enfants, le niveau de scolarité et l'activité principale (à savoir, études, travail rémunéré ou travail autonome, autre) sont tous associés significativement à ces tendances vectorielles. On s'est donc servi des attributs produits par d'autres processus du modèle LifePaths pour l'imputation. Le processus d'imputation a également été concu pour imiter les tendances variables relatives à l'emploi du temps observées chez les sujets qui présentaient les mêmes attributs, essentiellement grâce à une distribution des résidus vectoriels d'une analyse de régression à variables multiples.

8. Validation et qualité des données

Valider le modèle LifePaths est fondamentalement impossible, tout simplement parce que le modèle crée un échantillon à partir d'une cohorte de naissances hypothétique. Par conséquent, on ne pourra jamais en comparer les résultats avec la «réalité». Néanmoins, en raison de sa construction, le microcosme artificiel de vies devrait reproduire les principales distributions communes marginales qui lui servent de point de départ, par exemple le taux de participation à la population active, le taux de fécondité, le taux de mortalité, le taux d'union et de dissolution des couples, le taux d'inscriptions à l'école et la distribution du revenu provenant d'un travail, selon l'âge et le sexe.

On a constamment vérifié les comparaisons de ce genre durant la construction du prototype du modèle décrit ici. Dans une large mesure, il existe une bonne concordance. Les principales discordances surviennent lorsque les sources de données sous-jacentes manquent elles-mêmes de cohérence, signe qu'il

existe des erreurs dans les données originales. En fait, le modèle LifePaths fournit un cadre en partie analogue à celui du SCN, pour les microdonnées socio-économiques, cadre qui rend les données de diverses sources cohérentes, donc met en relief les incohérences.

9. Conclusion

Nous avons débuté en soulignant les besoins des utilisateurs pour des renseignements plus complets et plus cohérents sur le plan socio-économique et proposé une explication à l'échec des efforts déployés antérieurement à l'échelon international pour satisfaire ces besoins. Une nouvelle approche a été suggérée, approche supposant un usage beaucoup plus important de base de microdonnées multivariées et de méthodes de microsimulation. Les particularités fondamentales de cette nouvelle approche, entre autres sa cohérence et sa complétude, ont été mises en relief grâce aux résultats préliminaires venant du modèle en cours d'élaboration à Statistique Canada.

L'espace ne nous permet pas d'illustrer les autres caractéristiques du modèle par des graphiques, notamment les microdonnées explicites qui en forment la base et permettent d'analyser la diversité. Des recherches plus poussées sont nécessaires pour dégager d'autres grandes caractéristiques comme les indicateurs sommaires (à savoir, distribution du revenu durant la vie) et les simulations du type «et si ?». Les résultats présentés dans ce document constituent néanmoins une importante «preuve par construction» de la faisabilité pratique et technique d'une telle approche.

La même approche fait apparaître des lacunes et des faiblesses au niveau des données de statistique socio-économique existantes, examinées sous l'angle de la micro-analyse. L'approche LifePaths susciterait des exigences beaucoup plus sévères à l'égard de la cohérence et de la qualité des enquêtes ainsi que des méthodes de collecte de données socio-économiques. Dans la mesure où l'on reconnaît les avantages d'une méthode semblable au modèle LifePaths pour l'analyse des statistiques socio-économiques, pareille méthode pourrait constituer la base d'un exercice quelconque de planification stratégique pour les organismes nationaux qui dispensent des services de statistique.

Bibliographie

Bordt, M.,G.Cameron, S.Gribble, B.Murphy, G.Rowe, et M.Wolfson (1990), «The Social Policy Simulation Database and Model: An Integrated Tool for Tax/Transfer Policy Analysis», Canadian Tax Journal, 38:48-65.

Citro, C.F. et E.A.Hanushek (1991), Improving Information for Social Policy Decisions, The Uses of Microsimulation Modeling, National Academy Press, Washington, D.C.

Easton, G.S. et R.E.McCulloch (1990), «A Multivariate Generalization of Quantile-Quantile Plots», Journal of the American Statistical Association, juin, vol. 88, n° 410, Theory and Methods, pp376-386.

Garonna, P. (1994), «Statistics facing the concerns of a changing society», Statistical Journal of the United Nations ECE, vol. 11, n° 2, pp147-156

Gnanasekaran, K.S. et G. Montigny (1975), Tables de vie active des hommes au Canada et dans les provinces, 1971, Statistique Canada, n° 71-524F au catalogue occasionel, Ottawa.

Juster, F.T. et K.C.Land (1981), «Social Accounting Systems: An Overview» in F.T.Juster et K.C.Land (sous la dir. de), Social Accounting Systems -- Essays in the State of the Art, Academic Press, New York.

Juster, F.T., P.N.Courant et G.K.Dow (1981), «The theory and measurement of Well-Being: A Suggested Framework for Accounting and Analysis» in F.T.Juster et K.C.Land (sous la dir. de), Social Accounting Systems -- Essays in the State of the Art, Academic Press, New York.

Mathers, C. et J-M Robine (1993), «Health expectancy indicators: a review of the work of REVES to date», in J-M Robine, C.D.Mathers, M.B.Bone, I.Romieu (sous la dir. de), Calculation of Health Expectancies: Harmonization, Consensus Achieved and Future Perspectives, INSERM / John Libby Eurotext Ltd., vol. 226.

Moser, Sir C. (1973), «Social Indicators -- Systems, Methods and Problems», Review of Income and Wealth, Series 19, n° 2, juin, pp133-141.

OCDE (1976), Measuring Social Well-Being, Paris.

OCDE (1977), «Basic Disaggregations of Main Social Indicators», D.F.Johnston, Special Studies No. 4, The OECD Social Indicator Development Programme, Paris

OCDE (1982), The OECD List of Social Indicators, Paris.

Pommier, P. (1981), «Social Expenditure: Socialization Expenditure? The French Experience with Satellite Accounts», Review of Income and Wealth, décembre.

Pyatt (1990), «Accounting for Time Use», Review of Income and Wealth, Series 36, n° 1, mars, pp 33-52

Rowe, G. et S. Gribble (1994), «Income Statistics from Survey Data: Effects of Respondent Rounding», à venir dans Proceedings of the American Statistical Association, Section on Government Statistics.

Ruggles, N. et R. Ruggles (1973), «A Proposal for a System of Economic and Social Accounts», in M. Moss (sous la dir. de), The Measurement of Economic and Social Performance, National Bureau of Economic Research, New York.

Ruggles, R. (1981), «The Conceptual and Empirical Strengths and Limitations of Demographic and Time-Based Accounts», in F.T.Juster et K.C.Land (sous la dir. de), Social Accounting Systems -- Essays in the State of the Art, Academic Press, New York.

Stone, R. (1973), «A System of Social Matrices», Review of Income and Wealth, Series 19, n°.2, juin, pp143-166..

Organisation des Nations Unies (1975), Towards a System of Social and Demographic Statistics (SSDS), Studies in Methods, Series F, n° 18, ST/ESA/STAT/SER F/18, New York.

Organisation des Nations Unies (1979), The Development of Integrated Data Bases for Social, Economic, and Demographic Statistics (IDBs), Studies in Methods, Series F, n° 27, ST/ESA/STAT/SER F/27, New York.

Vanoli, A. (1994), «Extension of National Accounts: opportunities provided by the implementation of the 1993 SNA», Statistical Journal of the United Nations ECE, vol. 11, n° 3, pp183-191.

Wilk, M.B. (1987) «The Concept of Error in Statistical and Scientific Work», document présenté à la U.S. Bureau of the Census Third Annual Research Conference, Baltimore.

Wolfson, M.C. (1979), «Épargner pour la retraite: mais combien ?» , volume II, annexe 18 in Le système de retraite au Canada: problèmes et possibilités de réforme, Groupe d'étude sur la politique de revenu de retraite, ministère des Finances, Ottawa.

Wolfson, M.C. (1989), «Divorce, Homemaker Pensions, and Lifecycle Analysis», Population Research and Policy Review, 8: 25-54.

Wolfson, M.C., S.Gribble, M.Bordt, B.Murphy et G.Rowe (1989), «The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration», Survey of Current Business, 69, 36-40.

Wolfson, M.C. (1994), "Implications of Evolutionary Economics for Measurement in the SNA, Towards a System of Social and Economic Statistics", document présenté à la vingt-troisième conférence générale de l'International Association for Research in Income and Wealth, St.Andrews,Nouveau-Brunswick, août, 21-27, 1994, miméographié, Statistique Canada, Ottawa.