Définir et réussir la Révolution des données : le point de vue des démographes

Union internationale pour l'étude scientifique de la population

15 octobre 2014

1. Introduction

L'Union internationale pour l'étude scientifique de la population (UIESP) a organisé à Paris, du 9 au 10 octobre 2014, une réunion de 22 éminents démographes originaires à la fois de pays du Sud et du Nord, pour discuter du rôle que les démographes, et les compétences démographiques, pourraient jouer dans la Révolution des données.

Au cours de nos discussions, nous avons reconnu que la Révolution des données n'est pas une fin en soi. Il s'agit d'un mécanisme visant à améliorer la vie des populations dans les dix prochaines années et au-delà. Nous pensons que la probabilité de succès du programme de développement de l'après-2015 se trouvera renforcée par des plans et des stratégies fondés sur l'expertise scientifique et reposant sur des données de plus grande qualité, ce qui favorisera en retour de meilleurs choix politiques.

La démographie a une histoire longue et unique. Les démographes sont particulièrement compétents pour comprendre les liens systémiques entre les stocks et les flux de population dans le temps et dans l'espace. Cette compréhension profonde place les démographes en position de force pour évaluer de manière réaliste les possibilités offertes par des données collectées sur des populations, les limites de ces données, et la validité des résultats. L'évaluation de la qualité des données, la capacité d'associer et de traiter des données tirées d'une multiplicité de sources disparates, et de considérer les données comme partie d'un cadre systémique plus large, sont des aspects clés de cette compétence.

S'agissant plus particulièrement de la Révolution des données, les démographes sont bien placés pour évaluer des indicateurs relatifs à la population tels que les indicateurs des Objectifs pour le développement durable(ODD), dont le but est de suivre les progrès dans la réalisation de résultats chiffrés, et s'assurer que ces indicateurs sont cohérents, corrects et opérationnels. Les numérateurs et/ou dénominateurs d'un grand nombre de ces indicateurs ODD dépendent également de données démographiques, et la panoplie d'outils couramment utilisés par les démographes pour produire des estimations et des projections de variables de population plus fiables à partir de données limitées, déficientes et défectueuses peut aisément être mise au service des ODD.

Comprendre la **qualité** et de l'**interopérabilité** des **données** est un aspect essentiel du processus de production, d'analyse et d'interprétation des données pour atteindre l'objectif final de la Révolution des données. Ce document identifie les apports spécifiques de la communauté mondiale des démographes et des spécialistes de la population à la Révolution des données. Ces apports sont de deux types : tout d'abord, le rôle que les démographes pourraient jouer pour garantir la mesurabilité, la validité et l'utilité des ODD ; ensuite, un éclairage plus général sur diverses dimensions de la Révolution des données du point de vue des démographes.

Nous exhortons les organisations impliquées dans l'élaboration de la Révolution des données et des ODD à :

- identifier les domaines d'intervention pour lesquels les démographes peuvent contribuer au programme de développement de l'après-2015, et particulièrement les questions évoquées dans la suite de ce document ;
- recommander des investissements importants dans les méthodes et la formation ;
- envisager la constitution de groupes de surveillance ou de comités de haut niveau pour examiner plus avant les questions soulevées ici.

2. Garantir la mesurabilité, la validité et l'utilité des ODD

La probabilité de succès de la Révolution des données et du programme de développement pour l'après-2015 pourrait être renforcée par l'implication des démographes dans l'élaboration et l'évaluation des indicateurs proposés pour mesurer le progrès dans la réalisation des ODD, particulièrement en ce qui concerne les cinq points suivants.

a. Estimations ponctuelles et incertitude

Les démographes demandent aux rédacteurs des ODD de garder à l'esprit les limites des estimations ponctuelles en tant qu'indicateurs du développement. Si les estimations ponctuelles fournissent un chiffre global, la part substantielle d'incertitude qui entoure souvent ces estimations doit être clairement énoncée.

De plus, lorsque l'on produit des estimations et des indicateurs fortement désagrégés (comme l'implique le principe du « ne laisser personne derrière ») à partir de données issues de populations relativement petites, le degré d'incertitude est considérablement plus important. Les tendances déduites d'une série d'estimations ponctuelles de ces indicateurs pourraient être fallacieuses car la tendance observée se trouve souvent dans la marge d'erreur. Par conséquent, nous proposons que les démographes, qui sont habitués à travailler avec des échantillons de données représentant de petits groupes, soient plus directement impliqués dans l'élaboration des directives en matière de désagrégation.

b. Objectifs, indicateurs et arbitrage

Les démographes s'inquiètent du fait que les contraintes techniques de l'évaluation des risques pour certains des indicateurs surpassent les bénéfices que l'on pourrait tirer de leur mesure. L'accent mis sur le recueil des données pour une série d'indicateurs associés aux ODD peut effectivement amener à porter moins d'attention à d'autres informations essentielles, telles que les déterminants sous-jacents de ces indicateurs ou phénomènes, plus difficiles à définir en termes d'objectifs. Mesuré par un indicateur, un objectif de réduction de la mortalité maternelle peut ainsi se trouver confondu avec l'indicateur lui-même. Etant donné les difficultés liées à l'obtention de données adéquates, si cet indicateur est désagrégé, le travail se concentrera peut-être seulement sur la mesure de risques de mortalité. Avec des ressources limitées, on peut se retrouver face à un choix entre le recueil de données nécessaires au *calcul* des indicateurs et la collecte des informations nécessaires à *l'étude de ses déterminants causaux*, ici, les informations directement utiles au développement de politiques et stratégies appropriées à la lutte contre la mortalité maternelle. Certains autres phénomènes d'une grande importance pour l'amélioration du bien-être des populations sont difficiles à définir en termes d'objectifs, et risquent également d'être négligés. Il s'agit notamment de la croissance de la population et de sa structure par âge (phénomènes directement liés aux investissements dans le capital humain, le développement économique et la durabilité environnementale) et la migration. Là aussi, nous recommandons une plus grande implication des démographes, qui sont conscients des difficultés de la collecte de données sur des thèmes comme la mortalité maternelle, et conscients également des coûts financiers et des coûts d'opportunité.

c. Estimations empiriques ou estimations fondées sur un modèle

Il faut tout mettre en œuvre pour s'assurer que les indicateurs ODD s'appuient sur des données empiriques plutôt que sur des estimations fondées sur un modèle. S'il est évident que les modèles démographiques, mathématiques et statistiques sont utiles dans certains cas, une trop grande dépendance envers les estimations fondées sur un modèle induisent des risques importants. Tout d'abord, indépendamment du soin apporté à la construction de tels modèles, il existe toujours un risque non négligeable que ces derniers soient erronés, biaisés ou incorrectement paramétrés.

Ensuite, les résultats obtenus avec ces modèles reflètent dans une large mesure les hypothèses utilisées pour leur construction et leur paramétrage. Si les estimations fondées sur le modèle sont nécessaires (notamment les estimations de population annuelles dans les périodes intercensitaires), il convient de veiller à ce que ces modèles soient aussi solides que possible, et que la méthode de modélisation et toutes les hypothèses associées soient transparentes.

Dans le contexte de la production d'indicateurs fiables, la fiabilité des estimations utilisées pour la production de dénominateurs est d'une importance capitale. Les données empiriques sur la taille de la population ne s'obtiennent généralement que tous les dix ans, lors d'un recensement. La question qui se pose alors est de déterminer la meilleure façon de projeter la taille et la composition de la population durant les périodes intercensitaires. Dans ce domaine aussi les démographes possèdent une expertise scientifique substantielle, et cette connaissance et cette compétence devraient être employées à la mesure des indicateurs ODD.

d. L'importance d'un renforcement des statistiques nationales d'état-civil et de recensement

Si les estimations basées sur un modèle doivent être évités, il faut dans le même temps renforcer les systèmes statistiques nationaux (SNN) qui fourniront une grande partie des données utilisées pour les indicateurs ODD. L'amélioration de la régularité et de l'exhaustivité des données nationales d'état-civil est d'ores et déjà reconnue comme étant une priorité urgente dans les pays en développement.

Si la communauté des démographes et des spécialistes de la population souscrit pleinement à cet engagement, nous nous préoccupons aussi de l'amélioration de la qualité et de la couverture des données de recensement. Ces données de recensement fournissent des chiffres de référence pour de nombreux indicateurs, particulièrement à des niveaux de granularité fine (par exemple pour de petites zones, ou lorsqu'elles sont désagrégées selon des variables qui ne sont pas collectées dans le cadre du système d'enregistrement de l'état-civil), et constituent souvent, par ailleurs, la base de sondage pour tirer des échantillons représentatifs au niveau national pour de nombreuses enquêtes et d'autres exercices de collecte de données (y compris la gestion des biais de sélection pour ce que l'on appelle communément le *Big Data*). En outre, il ne faut pas négliger le besoin de métadonnées et d'une documentation détaillée sur l'élaboration et la base de calcul des indicateurs ODD.

Impliquer la communauté des démographes et des spécialistes de la population dans la conception et l'élaboration des indicateurs contribuera à garantir la cohérence des indicateurs ODD. Les démographes peuvent également jouer le rôle de conseillers sur les différentes méthodes et approches couramment utilisées pour évaluer les indicateurs à l'échelle des populations à partir de données limitées et défectueuses.

e. Définition des régions

Enfin, si les démographes sont conscients de la nécessité de produire des mesures pour de grandes régions du monde, nous appelons à la plus grande prudence dans leur construction et leur interprétation. En premier lieu, il convient d'utiliser un ensemble normalisé de définitions régionales afin d'éviter les problèmes associés aux différences de classification des pays d'une organisation à

l'autre. En second lieu, les utilisateurs de statistiques et d'indicateurs régionaux doivent être conscients de la composition de ces statistiques régionales, qui sont le plus souvent des agrégats pondérés. Ainsi, si une région se compose d'un pays avec une population importante et plusieurs autres avec des populations proportionnellement moins importantes, les statistiques régionales reflèteront principalement celles du pays le plus peuplé. Dans ces cas, il est de toute évidence dangereux d'appliquer de la même façon un indicateur régional à tous les pays.

3. La Révolution des données vue par les démographes

Au-delà des recommandations spécifiques telles que détaillées plus haut sur les ODD et le processus d'élaboration d'indicateurs mesurables, les participants à la réunion de Paris ont réfléchi à plusieurs aspects plus généraux de la Révolution des données. Pour réussir la Révolution des données, les experts devront prendre compte des enjeux suivants liés au domaine des sciences de la population :

a. Qualité des données

Pour que la Révolution des données soit un succès, la qualité des données collectées est d'une importance primordiale. A cet égard, les démographes ont identifié les enjeux suivants.

i. Base scientifique pour la qualité des données

Une des principales préoccupations des démographes consultés porte sur le besoin d'encourager l'élaboration d'une base scientifique pour la mesure et l'évaluation de la qualité des données. Il conviendrait en particulier d'accorder une plus grande attention aux difficultés méthodologiques liées à l'intégration de « nouvelles » formes de données (par exemple le *Big Data*, les données administratives, les données satellite, etc.) aux formes « classiques » de données, et à la question de la sélection et de l'incertitude des échantillons aux niveaux les plus fins de désagrégation géographique. Les démographes seraient bien placés pour contribuer à identifier les nouvelles opportunités offertes par ces données, ainsi que leurs limites et leur apporter une éventuelle caution scientifique.

Plus spécifiquement, il est crucial de s'assurer que toutes les données sont validées avant utilisation ; si c'est bien le cas dans une certaine mesure avec les formes « classiques » de données, nous ne sommes pas convaincus qu'il soit accordé autant ou suffisamment d'attention aux mêmes aspects des « nouvelles » données. Cette validation devrait être à la fois interne (en garantissant une cohérence interne des données) et externe (par rapport à d'autres données similaires). En outre, pour permettre cette validation, il convient de faciliter l'accès à une documentation régulièrement mise à jour ainsi qu'aux métadonnées. Cette documentation devrait inclure des informations détaillées sur les processus employés pour le nettoyage, la correction et toute autre manipulation effectuée sur les données (par imputation statique ou dynamique, par exemple) ainsi que sur les effets de ces manipulations sur les données.

Une des pistes possibles pour qu'une Révolution des données puisse répondre à ces préoccupations serait la création d'un **système commun de notation de la qualité des données**, fondé sur une évaluation indépendante et impartiale, et prenant en compte l'incertitude associée aux estimations dérivées des données. Par exemple, en disposant d'un indicateur obtenu à partir de différents ensembles de données, les chercheurs et les statisticiens seraient à même d'en évaluer la fiabilité, permettant de ce fait un usage prudent des données concernées et, le cas échéant, le calcul d'un indicateur composite plus précis.

Enfin, il est important de souligner que la plupart des indicateurs nécessiteront une bonne compréhension des estimations de population, particulièrement pour l'estimation des dénominateurs. A cet égard, les démographes ont un rôle essentiel à jouer en tant que spécialistes des projections démographiques. En effet, la plupart des outils usuels des démographes ont une réelle valeur pour l'ensemble des ODD, au-delà des variables strictement « démographiques ». Les efforts visant à favoriser l'augmentation des taux de scolarisation, par exemple, qui doivent être fondés sur des

projections de population d'âge scolaire et des schémas d'abandon scolaire, constituent un sujet qui peut être étudié de manière plus efficace avec la méthode des tables de mortalité.

ii. Des indicateurs représentatifs au plus haut niveau de granularité

En théorie, nous reconnaissons qu'il est nécessaire de disposer d'indicateurs ODD détaillés et valides à des niveaux de granularité fine afin d'assurer un suivi régulier des estimations désagrégées. En pratique, toutefois, nous insistons sur le fait qu'un compromis devra être trouvé entre ce qui est recherché et ce qui peut être réalisé sans affecter la solidité des estimations ainsi dérivées. Aucune approche « universelle » ne peut être déterminée : les indicateurs d'événements ou de résultats relativement rares (comme la mortalité maternelle) ne seront pas utilisables à des niveaux de granularité aussi fine que des indicateurs d'événements ou résultats plus courants (comme la mortalité infantile). Les démographes peuvent contribuer à déterminer les niveaux de granularité appropriés pour chaque indicateur sélectionné.

iii. Accessibilité des données

Dans le cadre de la Révolution des données et en lien avec l'objectif de permettre que les données soient utilisées pour placer les systèmes de pouvoir face à leurs responsabilités et encourager la transparence, toutes les données utilisées pour la détermination des indicateurs devraient être rendues accessibles rapidement et de façon régulière, avec des normes claires spécifiant les paramètres pour différents types de données. Les ensembles de données devraient être complétés par des métadonnées appropriées sur les données elles-mêmes, ainsi que sur leur processus de production ; et elles devraient être codées et diffusées dans un format standardisé pour en faciliter l'utilisation (par exemple, les normes du Data Documentation Initiative (DDI) pour les microdonnées dans les domaines sociaux, comportementaux et économiques).

b. Interopérabilité

Une Révolution des données réussie doit garantir autant que possible l'adhésion aux principes et aux normes d'interopérabilité. Nous avons identifié trois aspects pour cette question.

i. Accès aux données ouvert par palier

En matière d'accessibilité aux données, l'accès par palier constitue une des approches possibles car elle permet de garantir l'accès aux données tout en prenant en compte les questions de confidentialité. Un système par palier modulera les quantités de données qui seront rendues accessibles selon le niveau de désagrégation : plus l'indicateur sera désagrégé, plus le volume de données nécessaires sera important. Dans certains cas, par exemple, un micro-échantillon public de 10 % sera suffisant ; dans d'autres, la totalité des données sera nécessaire. Lorsque des questions de confidentialité se posent de manière légitime, des mécanismes alternatifs d'accessibilité aux données doivent être proposés. Des centres de données sécurisés dans les pays hôtes peuvent accorder un accès complet aux données lorsque c'est nécessaire, ou bien en prenant soin de masquer l'identité des individus et des ménages concernés (les enquêtes EDS effectuent ainsi de très légères modifications des coordonnées GPS afin de rendre impossible l'identification de ménages spécifiques).

ii. Données aux niveaux de granularité fine

Les données utilisées pour les estimations d'indicateurs, ou servant de base aux interventions des pouvoirs publics posent problème aux niveaux de granularité fine. Dans de nombreux cas, les données issues d'une seule source (une enquête, par exemple) peuvent ne pas être solides ou fiables à des niveaux de granularité fine. Il ne faut jamais perdre de vue les limites de ce type de données. A ce niveau, les estimations et les indicateurs devront souvent être complétés par des données harmonisées et couplées issues d'autres sources (y compris, peut-être, du *Big Data*).

Pour ce faire, il reste à développer et tester des méthodes fiables. De même, il convient de définir des normes de géocodification appropriées (y compris les décisions relatives à l'exactitude des

coordonnées géocodées). Lorsque les paramètres impliquent un calcul de distance (par exemple d'une source, d'une clinique ou d'une école), un haut niveau de précision est nécessaire pour produire des résultats significatifs. Les données géocodées permettront également de relier plus facilement les ensembles de données entre eux, et de fournir des informations sur des localités, ce qui est essentiel pour évaluer l'impact des politiques. Mais cela peut avoir des implications pour la confidentialité des données. Il convient d'étudier avec soin la manière dont cette difficulté peut être surmontée. Dans tous ces débats, les démographes pourraient apporter une contribution significative.

iii. Calibrage des « nouvelles » formes de données

Les données collectées à l'aide de méthodes et technologies nouvelles, y compris le *Big Data*, doivent être calibrées avant d'être fusionnées avec des données existantes afin de garantir autant que possible que tout biais (par exemple dans la sélection de l'échantillon) aura été éliminé de ces données avant leur intégration. L'évaluation de ces données a déjà été traitée dans les sections précédentes, mais l'exercice de calibrage est tout aussi important et doit avant tout porter sur l'échantillonnage, les cadres d'échantillonnage ainsi que le lissage des données, activités pour lesquelles les données de recensement sont généralement les plus appropriées. Encore une fois, de nouvelles normes, méthodes et techniques devront être élaborées pour s'assurer que le plus grand soin sera apporté en la matière.

4. Garantir la capacité institutionnelle

Lors des débats sur la Révolution des données auxquels nous avons pris part, le rôle des systèmes statistiques nationaux (SSN) et des instituts nationaux de statistique (INS) a reçu une attention particulière. Cependant, si les SSN doivent être considérés comme un élément clé d'une Révolution des données, il faut que les compétences et l'expertise adéquates existent au sein de ces SSN (et surtout des INS) pour collecter, évaluer et analyser les données. A cet égard, l'UIESP se préoccupe depuis quelque temps de l'érosion des compétences et des connaissances démographiques au sein des SSN et des INS. L'heure est venue de prendre des mesures significatives pour améliorer ces compétences, ces capacités et ces connaissances en investissant dans la formation du personnel des SSN et INS (y compris aux méthodes et théories démographiques fondamentales) et en élaborant des stratégies efficaces pour garantir le maintien de ce personnel bien formé au sein de ces organisations.

Conclusions

Parmi toutes les sciences sociales, la démographie est la discipline qui se consacre de plus près à la question de la qualité des données et à la façon de tirer au mieux parti de données limitées et imparfaites. L'approche démographique, structurée et systématique, à laquelle vient s'ajouter une batterie de méthodes éprouvées que les démographes ont mises au point au fil des années pour mesurer et analyser les différentes dimensions des stocks et des flux de population (santé et mortalité, fécondité et santé de la reproduction, migrations et intégration des immigrés, croissance de la population et structure par âge, qui affectent les perspectives de croissance économique et l'environnement), présente un intérêt évident aussi bien pour l'examen des indicateurs ODD proposés que pour leur mesure. De notre point de vue, la démographie a beaucoup à apporter, particulièrement pour ce qui concerne la qualité des données et leur interopérabilité, et nous espérons que des spécialistes de la population seront appelés à jouer un rôle de premier plan dans les activités futures liées à la Révolution des données.

Remerciements :

Ce document d'information est un résumé des conclusions de la Réunion d'experts de l'UIESP sur la démographie et la révolution des données pour l'après-2015 qui s'est tenue à Paris du 9 au 10 octobre 2014. Il a été préparé par Tom Moultrie (University of Cape Town ; tom.moultrie@uct.ac.za), Tom LeGrand (Université de Montréal; tk.legrand@umontreal.ca) et Emma Samman (Overseas Development Institute, UK; e.samman@odi.org.uk). Cette réunion a été organisée avec le soutien de la Fondation William and Flora Hewlett.