Une approche historique du recueil de données dans les pays en développement

Jean-Jacques Droesbeke

Université libre de Bruxelles

jjdroesb@ulb.ac.be.

A historical approach to data collection in developing countries

Résumé

Nous vivons dans un monde où le recueil d'informations est devenu essentiel dans l'organisation de la vie politique et sociale de nos sociétés. Si l'on parle souvent des développements de la « science des données » dans les pays industrialisés, qu'en est-il pour les pays en développement ? Nous abordons cette question par une démarche historique.

Mots-clés : pays en développement ; techniques statistiques ; données.

Summary

We live in a world where the collection of information has become essential in the organization of the political and social life of our societies. If we often talk about the developments of "data science" in industrialized countries, what about developing countries? We analyse this question through a historical approach.

Key words : developing countries; statistical techniques; data.

1. Introduction

La manière dont les données dans les pays en développement sont recueillies a donné lieu à de nombreuses évaluations contradictoires. Comme nous l'avons déjà souligné antérieurement (Droesbeke 2011), les méthodes utilisées dans les pays développés sont, pour certaines, difficilement applicables dans les pays en développement pour diverses raisons (manque de structures administratives, absence ou manque de statisticiens spécialisés, difficultés pour recueillir l'information auprès des individus des populations concernées, absence de bases de sondage …). Il peut aussi arriver que les méthodes retenues ne soient pas appropriées. On évoque encore la difficulté de traiter valablement les données recueillies.

Tous les pays ont été à un moment de leur histoire « en développement ». Aussi est-il utile d'examiner la manière dont on a recueilli des données numériques dans le monde jusqu'à ce jour pour examiner ce qui s'est passé dans les pays en développement jusqu'à nos jours.

2. Les leçons du passé

Pour tenter de résumer l'histoire des principales méthodes de recueil de données, il est intéressant de distinguer quatre périodes entre la naissance de l'écriture et notre époque (voir, par exemple, Droesbeke 2015 ou encore Droesbeke et Vermandele 2018a). La première précède le dix-septième siècle, la deuxième couvre les dix-septième et dix-huitième siècles, la troisième est le dix-neuvième siècle et enfin la dernière concerne le vingtième siècle et le début du suivant.

La première période se caractérise essentiellement par trois catégories de données observées : celles qui résultent d'un *dénombrement*, l'enregistrement de *productions de biens* et de *transactions commerciales* et enfin les résultats de l'*observation de phénomènes astronomiques et terrestres*.

Les dénombrements ont toujours constitué une opération importante de l'activité humaine, que ce soit au niveau familial, de la tribu, des États. Effectuer le relevé des habitants d'une ville ou d'une région et établir la liste de leurs biens, constituent deux opérations qui ont permis d'évaluer très tôt le nombre d'hommes pouvant être enrôlés dans les armées ou encore d'apprécier la capacité des individus à contribuer au train de vie des monarques et à la réalisation de leurs projets. Que ce soit dans le Croissant fertile où elle prend sa source écrite, l'empire chinois, les Indes, l'Egypte ancienne ou encore l'Empire romain, cette technique a été utilisée pendant de très longues périodes dans l'Antiquité. Le déclin de l'Empire romain et le Haut Moyen Âge n'ont par contre pas constitué une période propice pour l'organisation de recensements. Ce n'est qu'entre le 14^{e} et le 16^{e} siècle que l'on ressent à nouveau le besoin d'informations de ce type, que ce soit au niveau des rôles fiscaux ou à celui des relevés d'ordre religieux. Soulignons aussi que durant cette première tranche de notre découpage, les données sont *primaires* et *individuelles*, c'est-à-dire fournies directement par les individus concernés par le recueil.

Rappelons qu'un autre domaine propice à l'éclosion de données numériques, plus proche de la vie au jour le jour, est le commerce. Produire, vendre, consommer sont des activités avides de données même réduites à leur plus simple expression, que ce soit pour établir des listes de produits agricoles, de biens fabriqués par des individus ou des collectivités.

Passons à la deuxième période de notre découpage. Le 17^{e} siècle fait apparaître trois courants distincts dans la recherche d'informations statistiques en Europe : la *Staatkunde* allemande, les *enquêtes* de l'administration française et l'*arithmétique politique* anglaise.

La *Staatkunde* allemande trouve ses racines dans les travaux d'Aristote. Pour ses défenseurs, la statistique est la *science de l'État*. Purement descriptive, elle ne fait pratiquement jamais appel à des données chiffrées. Son influence perdurera jusqu'au 19^{e} siècle, surtout en Europe centrale.

En France, on plaide plutôt pour les dénombrements comme outils de gouvernement. Deux hommes se sont particulièrement illustrés dans ce domaine. Le premier, Jean-Baptiste Colbert (1619-1683), en fait usage tant en France que dans ses colonies. C'est ainsi que :

« le premier relevé d'habitants, au Canada, se rapportait à la fondation de Port-Royal, dans la nouvelle Ecosse, en 1605. On dispose encore d'autres relevés, dont celui datant de la fondation du Québec, en 1628. En 1663, la population de la Nouvelle-France est évaluée à 2500 habitants, dont 800 à Québec. Mais le premier recensement nominatif des temps modernes est entrepris au Canada en février-mars 1666. Il donna 3215 habitants, répartis selon le sexe, l'âge, l'état matrimonial, la profession » (Hecht 1987, p.45).

Un second personnage qui s'est aussi illustré dans ce domaine est Sébastien Le Prestre, marquis de Vauban (1633-1707) ; il est l'auteur, en 1686, d'une *Méthode généralle et facille pour faire le dénombrement des peuples* très intéressante.

En Angleterre, les difficultés de mise en œuvre d'un recensement (réactions de méfiance des enquêtés, coûts de mise en œuvre trop élevés) permettent l'émergence d'un mouvement novateur dénommé *arithmétique politique.* Il est dû principalement au mercier londonien John Graunt (1620-1674) et à son ami, l'économiste William Petty (1623-1687). Comme le dira William Davenant, émule de Petty, « *l'arithmétique politique est l'art de raisonner par des chiffres sur des objets relatifs au gouvernement* ». On y trouve les fondements de la *méthode du multiplicateur* qui a marqué les techniques de recueil de données aux 17^{e} et 18^{e} siècles, provoquant une mise à l'ombre certaine de la Staatkunde allemande en Europe occidentale.

La *méthode du multiplicateur* repose sur l'idée suivante : il existe des quantités qui sont en rapports simples et relativement constants avec la population d'un pays. Ainsi, dès la fin du Moyen Âge, on a vu apparaître en Occident la notion de *feu* apparenté au concept de maison ou de logement (déjà utilisé en Chine depuis longtemps), plus facile à dénombrer. Pour obtenir une estimation du nombre d'individus dans une population, il suffit de multiplier le nombre de feux par un *multiplicateur* adéquat. À l'époque, on possède aussi des registres contenant le nombre de baptisés dans l'année. L'usage d'un autre multiplicateur permet aussi d'estimer la taille de la population à partir de ces registres. L'utilisation d'un multiplicateur pose cependant des problèmes de précision des estimations obtenues. Il n'est donc pas étonnant de constater que les recensements sont revenus en force au 19^{e} siècle avant de connaître une stagnation puis un déclin au 20^{e} siècle, dû notamment à l'introduction de registres administratifs performants et au développement des techniques de sondage.

Cette deuxième période est aussi celle qui concerne la troisième source de données numériques mentionnée ci-dessus, à savoir l'astronomie. Les Babyloniens avaient déjà observé les mouvements du soleil et des planètes à intervalles réguliers, obtenant ainsi plusieurs observations d'un même phénomène. Dans cette situation, la difficulté la plus importante est de les remplacer par une « valeur de compromis ». Que ce soit durant la première période considérée ci-dessus ou au début de la deuxième période, les faibles progrès techniques réalisés dans la recherche d'une plus grande précision des instruments de mesure ont fait croire longtemps qu'une « bonne mesure » était meilleure qu'une *agrégation* dont on ne soupçonnait pas l'intérêt. Il faut attendre le 18^{e} siècle pour voir apparaître une théorie des erreurs d'observation permettant notamment de justifier l'usage d'une moyenne arithmétique pour agréger des observations distinctes d'un même phénomène (voir, par exemple, Droesbeke et Vermandele 2016, 2018a et 2018b).

Le 19e siècle occupe une place très importante dans l'histoire des données numériques. Reprenons quelques points forts de cette histoire durant ce siècle (Droesbeke et Vermandele 2018a) :

1) La *statistique* devient un outil de gestion important des États, au niveau économique et social (Desrosières 1993).
2) L'application à l'étude des populations et de leurs caractéristiques humaines, des outils utilisés en astronomie, permet à Adolphe Quetelet (1796-1872) de créer une *théorie des moyennes* aux accents multiples (Académie royale de Belgique 1997, Desrosières 1993 ou Droesbeke et Vermandele 2016).
3) La *démographie* devient un domaine spécifique de l'étude des populations humaines et de leur dynamique.
4) Les *données individuelles* cohabitent avec les *données agrégées* qui acquièrent un statut à part entière.
5) Les *tables statistiques* et les *représentations graphiques* deviennent des outils importants d'analyse et de communication.
6) Le rôle central joué par la recherche d'une moyenne est remplacé par celui de la mesure d'une *dispersion* autour de cette valeur ou d'un autre milieu, surtout dans la seconde moitié du siècle.
7) Le centre de gravité de la statistique se déplace vers Londres qui voit l'émergence des concepts de *corrélation* et de *régression* (Droesbeke et Tassi 2015 ou Droesbeke et Vermandele 2016) dans un contexte d'évolutionnisme et d'eugénisme.

Pour ce qui concerne le recueil des données, trois méthodes sont utilisées ou initiées : les *recensements* qui deviennent récurrents, les *monographies* et, à la fin du siècle, les *sondages* (Desrosières 1993, Droesbeke et Tassi 2015 ou Droesbeke et Vermandele 2016).

Recourir à des recensements ou des sondages permet la plupart du temps d'étudier des faits généraux. Par contre, la nécessité d'aborder des questions plus fines demande des méthodes plus adaptées comme les *monographies*, proposées au 19e siècle par Pierre Guillaume Frédéric Le Play (1806-1882) pour qui « *Rien de tel que de recourir à des réseaux de familiarité* ». Cette idée lui était venue à l'esprit pour décrire les habitudes de vie des ouvriers dans cette Europe qui s'ouvrait à l'industrialisation. Il a demandé aux notables des villages de désigner des ouvriers « typiques » dont il a étudié en détail l'existence (Desrosières 1988). Il ne faut évidemment pas confondre cette démarche avec un sondage ou un recensement. Les objectifs ne sont pas identiques, la façon d'interpréter les résultats de l'étude non plus.

Au début du 20e siècle, début de notre quatrième période, l'*inférence statistique* est au centre des préoccupations, avec deux problèmes centraux : l'*estimation de paramètres d'une population* et les *tests d'hypothèses* réalisés à partir d'un échantillon. De nouvelles méthodes de recueil de données voient le jour, permettant de mettre en œuvre des méthodes « inférentielles ». Citons en particulier les *plans d'expérience*, les méthodes de *stratification* en théorie des sondages (voir ci-dessous) ou encore l'étude des *données longitudinales*. On recourt de plus en plus à l'usage de modèles sous-jacents (Dehon *et al.* 2015, Droesbeke et Vermandele 2016 ou Saporta 2011). Il y a des modèles pour *comprendre* et des modèles pour

prédire (Breiman 2001, Donoho 2015 ou Saporta 2017). Les premiers facilitent souvent l'usage des seconds. Parallèlement des *stratégies d'analyse* voient le jour ainsi que des *procédures de diffusion des résultats d'analyse* appropriées.

Les données sont devenues *multivariées,* portant simultanément sur plusieurs *caractères* ou *variables.* Elles sont *quantitatives* ou *qualitatives,* selon qu'elles sont constituées de nombres ou pas. Il en est de *manquantes* et d'*extrêmes,* c'est-à-dire très différentes de la majorité d'entre elles. Elles deviennent de plus en plus *nombreuses* ce qui pose de nouvelles questions sur la manière de les recueillir, leur stockage, leur traitement ou encore le temps mis à les analyser. En ce début de vingt-et-unième siècle, l'analyse des *données massives* — ou *Big Data* — et l'*intelligence artificielle* constituent les nouveaux domaines de recherche des « data scientists ».

3. La situation dans les pays en développement

Afin de ne pas nous disperser, nous limiterons notre propos aux pays africains. Ce qui caractérise la plupart de ceux-ci, c'est qu'ils ont été colonisés à partir de la fin du 19^{e} siècle par des pays européens. Jusque dans les années 1950-1970, les structures mises en place dans ces pays ont été construites sur les modèles utilisés dans les pays colonisateurs. L'accès à l'indépendance des pays concernés n'a pas modifié fortement la situation si ce n'est au niveau des difficultés rencontrées par la mise en œuvre des avancées méthodologiques de l'époque. Des accords de coopération entre pays antérieurement liés ont peut-être permis de tenir compte de ces dernières, mais pas sans difficultés. Une conséquence importante de cette situation est que ces difficultés ont permis des avancées méthodologiques nouvelles dans les techniques de recueil de l'information. Selon Théodore (1995, p. 9) :

> « *A posteriori*, l'amélioration de la situation à partir des années 1950 peut être expliquée par quatre facteurs exogènes. Trois d'entre eux sont liés à la situation nationale de l'époque :
>
> - La diffusion de la méthode des sondages.
> - Les progrès de la comptabilité nationale.
> - Les besoins du Fonds d'investissement et de développement économique et social.
> - La prise de conscience de besoins de connaissances chiffrées par l'Organisation des Nations Unies… ».

Le premier facteur cité ci-dessus est le recours aux sondages, en plein développement méthodologique et pratique à l'époque. Pour en parler, rappelons quelques éléments de base relatifs à cette méthode de recueil d'informations, tirés de Droesbeke et Vermandele (2019).

3.1 Quelques éléments pour comprendre les sondages

La figure 1 présente les différentes étapes d'une enquête par sondage (Droesbeke et Vermandele 2019, p. 22). On y constate la nécessité de définir le plus correctement possible la *population-cible* dans laquelle on veut prélever un échantillon.

On distingue essentiellement deux grandes familles de méthodes de sondage : celle des méthodes *aléatoires* (ou *probabilistes*) — où le *hasard* régit la procédure de sélection de

l'échantillon— et celle des méthodes dites *empiriques* (ou *à choix raisonné*) utilisées davantage dans le domaine commercial (voir Droesbeke et Vermandele 2019).

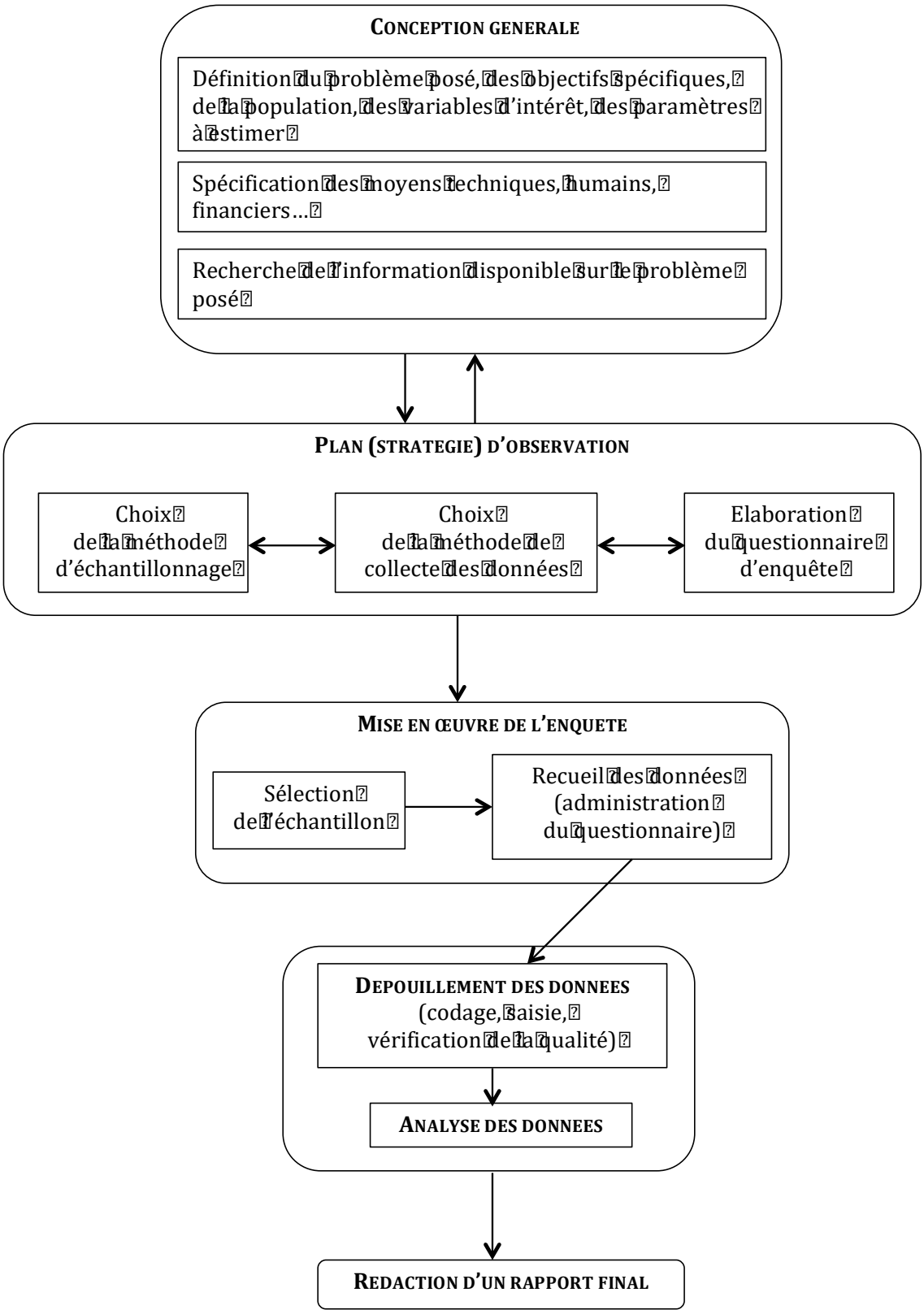

Figure 1 : *Les principales étapes d'une enquête par sondage*

Une méthode de sondage aléatoire nécessite tout d'abord de disposer d'une *base de sondage*, c'est-à-dire d'une liste exhaustive — complète, à jour et sans double compte — des unités statistiques de la population, représentées chacune par un identifiant unique. « L'un des facteurs de réussite d'une enquête par sondage tient à la base de sondage », écrit Philippe Brion (1995). C'est dans cette base que les individus interrogés sont sélectionnés selon un *plan de sondage* adapté à la situation sur le terrain.

La détermination du plan de sondage permet d'affecter à chaque unité ou individu de la population-cible une probabilité *connue* de faire partie de l'échantillon qui sera sélectionné ; cette probabilité porte le nom de *probabilité d'inclusion*. Lorsqu'on fait appel à une méthode de sondage *aléatoire*, c'est le hasard qui décide si une unité de la population fera partie de l'échantillon sélectionné et on connaît à l'avance avec quelle probabilité cela risque de se produire (voir, par exemple, Clairin et Brion 1997 ou encore Droesbeke et Vermandele 2019).

La méthode de sondage aléatoire la plus simple est certainement celle qui consiste à prélever un échantillon en effectuant un nombre fixé de tirages successifs « au hasard » (à l'aveugle) dans la population. On l'appelle le *sondage aléatoire simple.* Tous les individus de la population ont la même probabilité d'être sélectionnés pour faire partie de l'échantillon. Il s'agit donc d'un *sondage à probabilités égales*. À titre d'exemple, le tirage du lotto peut être assimilé à un sondage aléatoire simple de six nombres parmi les quarante-cinq premiers nombres entiers. Chacun de ces derniers a une probabilité d'inclusion égale au *taux de sondage* défini comme étant le rapport entre la taille de l'échantillon (6) et celle de la population (45).

Le but d'un sondage est d'*estimer* la valeur d'un *paramètre de la population* auquel on s'intéresse, comme la moyenne des âges des individus d'une population, le revenu total des habitants d'un pays ou encore la proportion d'individus ayant fait des études supérieures. Comme on n'utilise qu'une partie de la population pour faire cette estimation, il est important d'en connaître les qualités et de pouvoir disposer d'un moyen d'apprécier la *précision* de cette estimation (est-elle proche ou non de la valeur inconnue du paramètre auquel on s'intéresse ?). Un sondage aléatoire simple permet d'atteindre ces objectifs.

Une manière intéressante de procéder consiste à compléter l'estimation fournie par un échantillon au moyen d'un *intervalle de confiance*, familièrement dénommé *fourchette* dans le jargon des sondeurs. Cet intervalle est généralement construit en le centrant en l'estimation du paramètre fournie par l'échantillon ; quant à ses bornes, elles sont définies de telle sorte qu'il y ait de très grandes chances — le plus souvent, on choisit 95 chances sur 100, mais c'est purement arbitraire — pour que l'intervalle obtenu à partir des observations réalisées dans l'échantillon prélevé recouvre bien la valeur exacte du paramètre ; on spécifie cette propriété de l'intervalle de confiance en disant qu'il lui est associé un *niveau de confiance* de 95%. Une estimation sera d'autant plus précise que cet intervalle est de petite largeur. On appelle *marge d'erreur* la moitié de cette dernière.

On peut montrer que la précision d'une estimation par sondage aléatoire simple dépend de trois facteurs :

- les deux premiers sont fournis par la taille de l'échantillon et le taux de sondage. Plus la taille de l'échantillon est grande et se rapproche de la taille de la population, plus la précision de l'estimation est élevée. Cette propriété est naturelle : elle exprime simplement le fait que plus l'échantillon est grand, plus on peut avoir confiance dans l'estimation qui en résulte ;

- le troisième facteur est lié au caractère homogène ou hétérogène de la population vis-à-vis de la variable d'intérêt à laquelle on s'intéresse (âge, revenu…). L'hétérogénéité de la population induit un effet négatif sur la précision de l'estimation.

Si la méthode de sondage aléatoire simple permet de comprendre aisément la manière dont un sondage peut être interprété, elle n'est pas la plus efficace. De nombreuses méthodes plus ou moins complexes ont été proposées pour améliorer cette efficacité et pour résoudre des problèmes pratiques comme l'optimisation des coûts d'enquête, les difficultés de disposer d'une base de sondage accessible… Il n'est pas question de les présenter toutes ici (la lectrice ou le lecteur intéressé peut se reporter à Droesbeke et Vermandele (2019) ou encore aux ouvrages mentionnés dans la bibliographie pour en examiner les plus courants). Nous nous limiterons donc aux trois méthodes suivantes.

a) Le sondage stratifié

Un *sondage stratifié* consiste à découper la population en sous-populations, appelées *strates*, selon les modalités de l'une ou l'autre variable souvent qualitative, puis à réaliser un sondage aléatoire simple dans chacune d'elles. Chacun des échantillons ainsi prélevés constitue un *sous-échantillon* de l'échantillon global. Nous venons de mentionner qu'à taille d'échantillon fixée, un sondage aléatoire simple est plus efficace lorsqu'il est appliqué dans un ensemble d'unités homogène plutôt que dans un ensemble hétérogène. On a dès lors clairement intérêt à découper la population en strates les plus homogènes possible : chaque sondage partiel, réalisé dans une strate particulière, jouira alors d'une assez bonne précision et l'assemblage des sondages partiels dans les différentes strates donnera lieu à des résultats plus fiables qu'un sondage aléatoire simple de même taille effectué dans l'ensemble de la population, sans découpage préalable de celle-ci. Notez que la stratification définit généralement un sondage aléatoire à probabilités inégales, les individus des strates homogènes ayant une probabilité d'inclusion plus faible que ceux des strates hétérogènes. Elle répond souvent aussi à un objectif de réduction des coûts d'enquête ou d'optimisation de sa gestion. C'est en particulier le cas lorsqu'on utilise un critère géographique, comme la région par exemple, pour le découpage de la population : cela permet d'organiser l'administration de l'enquête région par région et de diminuer ainsi les frais de déplacement des enquêteurs.

b) Le sondage à deux degrés

Supposons que la population-cible soit partitionnée en un nombre relativement élevé de sous-ensembles d'individus. Ces sous-ensembles sont généralement appelés les *unités primaires* de la population. Pour réaliser un *sondage à deux degrés*, on prélève aléatoirement un échantillon d'unités primaires ; puis, dans chaque unité primaire sélectionnée, on prélève un échantillon d'individus. La construction de l'échantillon se fait donc en *deux étapes* successives.

L'utilisation de cette méthode est essentiellement motivée par des considérations de faisabilité, et la mauvaise qualité, voire l'inexistence, d'une base de sondage pour l'ensemble de la population à sonder. On ne peut dès lors pas prélever un échantillon par sondage aléatoire simple. Le sondage à deux degrés peut alors être une solution intéressante, pour autant que la population soit découpée en unités primaires dont il est aisé de dresser la liste et qu'il soit possible de constituer ensuite une liste exhaustive

des individus ou unités statistiques qui composent chaque unité primaire sélectionnée au premier degré du sondage. On peut bien sûr généraliser la procédure à *plusieurs degrés.*

L'utilisation d'un sondage à deux ou plusieurs degrés permet, dans certains cas, de réduire drastiquement les coûts liés aux déplacements que doivent faire les enquêteurs pour aller à la rencontre des personnes enquêtées. Considérez le cas où la population à sonder est celle d'une région relativement vaste. Si vous optez pour un sondage à deux ou plusieurs degrés pour lequel les unités primaires correspondent à des zones géographiques de dimensions relativement faibles, vous allez fortement limiter la zone de travail de chaque enquêteur et réduire ainsi significativement les coûts — et les pertes de temps — liés à ses déplacements.

c) Le sondage en deux phases

Le *sondage en deux phases* est une méthode particulière de sondage à probabilités inégales pour laquelle on procède en deux temps : dans un premier temps (phase 1), on tire un échantillon — généralement selon une méthode aléatoire simple et relativement peu coûteuse — dans la population entière ; dans un second temps (phase 2), on prélève un échantillon dans l'échantillon prélevé au cours de la première phase, selon un plan de sondage à probabilités d'inclusion inégales définies sur la base d'informations individuelles récoltées auprès des individus de l'échantillon de phase 1. On peut aussi généraliser la méthode à trois phases ou plus.

3.2 Deux exemples développés en Belgique

Si le texte de Théodore cité ci-dessus concerne la France, il est aussi valable en grande partie pour d'autres pays. Prenons le cas de la Belgique en évoquant une enquête particulière à titre d'exemple de sondage appliqué dans un pays en développement au milieu du siècle dernier (Romaniuk 2006). Dans les années 1950, la République Démocratique du Congo, qui s'appelait alors « Congo belge », était composée de 6 provinces, 23 districts, 135 territoires. Ces derniers étaient eux-mêmes divisés en circonscriptions, groupements de villages, villages et petits centres. Pour chaque territoire, on disposait, en principe, dans chaque village du nombre d'habitants, de l'appartenance tribale du village et de certaines informations économiques. En général, chaque territoire avait un chef-lieu pour lequel on disposait de fichiers d'habitations. On possédait aussi des cartes relativement complètes des territoires avec des plans de ville. Mais des contraintes financières et l'incertitude sur certaines données recueillies ont, comme dans d'autres pays, favorisé le recours à des enquêtes par sondage.

Disposant d'un recensement administratif récent, les enquêteurs souhaitaient, d'une part, vérifier la qualité de ce dernier et, d'autre part, rechercher des informations complémentaires. Il était bien sûr nécessaire de disposer d'une base de sondage relativement correcte, s'appuyant sur une équipe d'enquêteurs autochtones bien acceptés sur le terrain et posséder un centre de traitement des données bien organisé. Les responsables de l'enquête que nous citons dans cet exemple ont dès lors procédé à un sondage stratifié de la population congolaise avec des taux de sondage par strate variant entre 10% et 15%. Cette enquête par sondage présentait des qualités qui l'ont rendue crédible.

Mais le sondage ne fut pas le seul type de recueil utilisé à l'époque. À titre d'exemple, citons une autre enquête menée dans le même pays, qui s'apparente davantage à une monographie (CEMUBAC 1972). Elle avait pour objectif d'analyser très finement la population d'un petit village (Fuladu) d'une trentaine d'habitants. La manière dont elle a été menée repose sur des choix clairement exprimés, avec un souci de la précision dans le protocole d'enquête et une clarté dans la présentation des résultats qui méritent d'être signalés, même si le recours fréquent à la comparaison de pourcentages peut sembler excessif. Ici aussi, l'intervention d'enquêteurs autochtones a contribué au bon fonctionnement de l'enquête.

3.3 Le recueil des données après les années 1970

Les recensements menés en Afrique dans les années 1970 ont été appréciés de manière diverse. Les résultats affichés n'étaient pas tous exempts de falsifications et de questionnements divers (voir, par exemple, Locoh et Omoluabi 1995), tant il était difficile d'appréhender l'ensemble des unités statistiques à recenser via un ratissage de terrain. Deux voies ont été suivies pour tenter d'améliorer leur qualité : accentuer la formation des personnes chargées du recueil et du traitement des données, et compléter le travail au moyen de sondages plus accessibles financièrement. De grandes enquêtes ont été menées au niveau international. Après l'*enquête mondiale sur la fécondité* (EMF) qui s'est déroulée de 1972 à 1984 dans 60 pays — dont 20 européens et 15 africains — un autre programme d'envergure a pris corps : les *enquêtes démographiques et de santé* (EDS) qui ont concerné plus de 20 pays, dont 15 entre 1985 et 1990, 13 de 1990 à 1993 et 10 de 1993 à 1995 pour le continent africain (voir Cantrelle 1995).

Que valaient ces différentes enquêtes par sondage ? Devant la difficulté de mise en œuvre des enquêtes dans les pays en développement, qu'elles soient exhaustives ou non d'ailleurs, quelques statisticiens, « isolés mais astucieux » selon l'expression utilisée par Philippe Brion, ont mis au point des méthodes nouvelles qui leur ont permis d'être en avance sur leur temps. Prenons, par exemple, le recensement des nomades mauritaniens réalisé en 1977, en complément de celui de la population sédentaire de 1976. Si ce dernier a pu se réaliser de manière « classique », celui de 1977 a posé de nombreuses questions « de définition, de repérage sur le terrain et de traitement statistique » qui ont engendré des innovations méthodologiques intéressantes, notamment dans la façon de constituer une base de sondage facilement utilisable (Paccou et Blanc 1979 ; voir aussi Clairin 1988).

Un autre exemple significatif de cette constatation est l'approche originale développée à partir de 1987 pour mesurer et analyser l'économie informelle dans les pays en développement. L'économie informelle est articulée autour de deux concepts : le *secteur* informel et *l'emploi* informel. Le secteur informel est « l'ensemble des entreprises individuelles (en général non agricoles) produisant au moins en partie pour le marché, qui opèrent à petite échelle (en-deçà d'un certain seuil d'emplois ; souvent 5 employés) » (Nordman et Roubaud 2010). L'emploi informel comprend essentiellement deux composantes : les emplois dans le secteur informel et l'emploi non protégé dans le secteur formel. Des « enquêtes 1-2-3 » ont été menées dans de nombreux pays à partir de 1987 selon un processus dont la version complète comporte trois phases.

« La première phase […] est une enquête sur l'emploi, le chômage et les conditions de travail des ménages et des individus. Elle permet de documenter et d'analyser le fonctionnement du marché du travail et est utilisée comme filtre pour la deuxième phase où un échantillon représentatif des unités de production est enquêté. […] Enfin dans la troisième phase, une enquête sur les revenus et les dépenses est administrée à un sous-échantillon de ménages sélectionnés lors de la phase 1, pour estimer le poids des secteurs formel et informel dans la consommation des ménages, par produits et par types de ménage. La phase 3 permet également l'estimation du niveau de vie des ménages et de la pauvreté monétaire, basée sur les revenus ou les dépenses » (Nordman et Roubaud 2010).

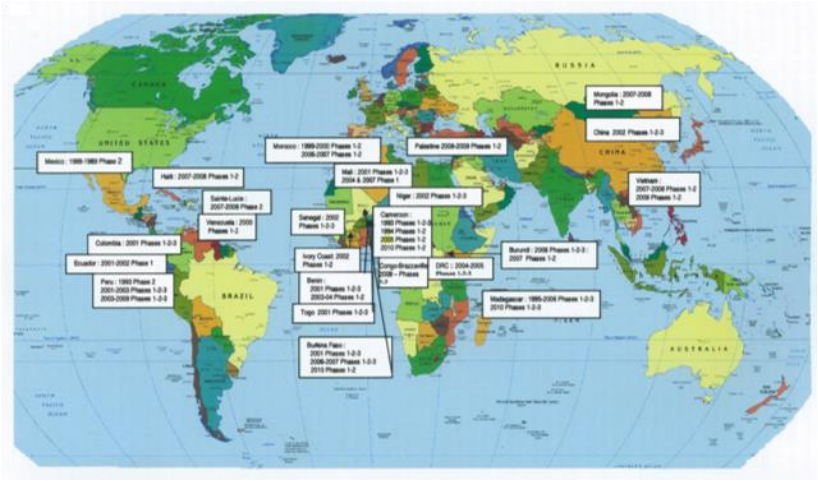

Figure 2 : *Les enquêtes* 1-2-3 *dans le monde* (Nordman et Roubaud 2010, p. 6)

Mais la statistique publique africaine a eu aussi d'autres défis à relever dès la fin du siècle dernier, sollicitée pour participer à la construction de bases de données internationales dans de nombreux domaines. La tâche était ambitieuse mais cette opération s'est faite « dans un contexte de raréfaction des ressources humaines et financières allouées par l'État depuis les années 1980 et l'instauration de programmes d'ajustement structurel » (Bédécarrats *et al.* 2016, s'appuyant sur Devaradjan 2013). Ces défis ont donné lieu à la création de structures internationales. Citons, entre autres, AFRISTAT, constitué à Abidjan le 21 septembre 1993 et regroupant au départ 14 États ou encore, plus récemment, l'institut panafricain de statistique STATAFRIC, opérationnel depuis 2017.

Le recueil des données dans les pays en développement a pris une dimension nouvelle au début du 21e siècle avec les *Objectifs du millénaire pour le développement* (OMD), suivis un peu plus tard par les *Objectifs de développement durable* (ODD). De nouvelles stratégies ont vu le jour : les politiques basées sur des données probantes, la gestion par les résultats… Elles ont engendré la publication de *Stratégies nationales pour le développement de la statistique* dans les pays en développement.

Il en est résulté, en 2015, un agenda pour le développement durable 2015-2030, adopté par l'ONU. Dans ce contexte, un appel à une « révolution des données » a été lancé dès 2013 dans cette institution internationale (Moultrie 2016). Les enjeux de ce programme sont énormes pour tous les pays, en particulier pour ceux d'Afrique (Bédécarrats *et al.* 2016): 17

objectifs, 169 cibles et 229 indicateurs statistiques sont proposés. On imagine la difficulté, voire l'impossibilité, de participer efficacement pour de nombreux pays. Il est à espérer que les développements de la nouvelle *science des données*, incluant notamment l'usage de données ouvertes et massives, pourront aider à atteindre les objectifs de ce programme.

4. En guise de conclusion

Il faut veiller à ce que la révolution des données dans laquelle nous sommes actuellement engagés n'agrandisse le fossé entre pays développés et les autres. Il ne s'agit pas seulement d'améliorer le recueil des données ; il faut aussi chercher à assurer leur qualité et un traitement adéquat. La porte est malheureusement ouverte à de nombreuses imprécisions, des erreurs de tout type, voire des falsifications. Cette situation n'est pas nouvelle, mais bien sa dimension.

Le pessimisme n'est pas de mise. Osons espérer que les progrès actuels réalisés dans le recueil et le traitement des données numériques permettent de limiter ces dangers, tant dans les pays en développement que dans les autres. La chasse aux erreurs et aux falsifications est encore plus nécessaire qu'auparavant car elles n'ont plus nécessairement la même origine. Et ce défi posera certainement de très nombreux problèmes dans les pays en développement.

Nous ne voudrions cependant pas terminer cet article par ce constat. Nous avons évoqué les trente millions de Nigérians manquants dans le recensement organisé en 1992 (Locoh et Omoluabi 1995). Il est une autre histoire qui ne manque pas de sel et qui concerne l'ex-« Congo belge ». Nous ne résistons pas au plaisir de reprendre ici le texte de Droesbeke et Vermandele (2018a), page 84 :

> **«** En 1885, Henry Morton Stanley (1841-1904) publie *The Congo and the founding of his free state* dans lequel il raconte ses missions pour Léopold II, roi des Belges. Il se rend compte qu'il lui est absolument nécessaire d'y mentionner une estimation du nombre d'habitants de ce nouvel état et, pour y arriver, il recourt, probablement sans le savoir, à la méthode du coefficient multiplicateur [voir ci-dessus]. Il estime dans son ouvrage avoir observé environ 806 000 habitants sur les rives situées de part et d'autre du fleuve Congo et de certains de ses affluents sur lesquels il a navigué. Après avoir calculé que ces rives s'étendent sur 2030 miles, il émet une hypothèse de travail selon laquelle les habitants qu'il a observés proviennent d'un village situé au maximum à 10 miles de la rive. Connaissant la superficie totale du pays, une petite règle de trois lui fait écrire que la population totale est de 42 608 000 habitants. Ce nombre rond devient « la » référence pour les spécialistes, surtout anglo-saxons. Cet ouvrage est traduit en français par Gérard Harry à Bruxelles, au siècle suivant. Sans être un grand mathématicien, ce dernier s'aperçoit d'une petite erreur de calcul de Stanley. Pour arriver aux 2030 miles que représente la longueur des rives où se trouvaient les populations locales, il a simplement multiplié par 2 (il y a en effet 2 rives de part et d'autre d'un cours d'eau) la distance totale parcourue par son bateau : 1515 miles. Harry se rend compte que 2 fois 1515 ne vaut pas 2030 mais bien 3030. Et comme le dit si gentiment Stengers, en 2007 : « *discrètement, sans un mot d'avertissement au lecteur, le traducteur rectifie le calcul de Stanley pour aboutir ainsi à une population de 27 694 000 habitants* », ce deuxième

chiffre devenant dès lors la référence dans les pays de langue française. De nombreux ouvrages ont cité ces estimations, parfois la première, parfois la seconde selon qu'on avait lu la version originale du livre de Stanley ou sa traduction. L'histoire aurait pu rester anecdotique si en 1999 n'avait paru un best-seller d'Adam Hochschild dont le titre est évocateur : *Les fantômes du roi Léopold II. Un holocauste oublié*. Un document filmé a même été tiré de ce dernier, produit par la *BBC* et diffusé sur les petits écrans avec un certain succès. L'effet aurait peut-être été moins ravageur si la différence entre population en 1885 et population au début du vingtième siècle n'avait pas été basée sur l'estimation initiale de Stanley ! Il faut reconnaître que la disparition de près de quinze millions d'habitants due à une erreur de multiplication, ce n'est pas si fréquent ! ».

Remerciements

Nous tenons à remercier vivement Philippe Brion, Marc Christine et Jean-Pierre Cling pour les informations et les remarques qu'ils nous ont aimablement transmises.

Bibliographie

Académie Royale de Belgique (1997), *Actualité et universalité de la pensée scientifique d'Adolphe Quetelet,* Actes du Colloque des 24 et 25 octobre 1996, textes rassemblés sous la direction scientifique de J.-J. Droesbeke, *Mémoire de la Classe des Sciences,* 3e série, tome **13.**

Bédécarrats Fl., Cling J.-P. et Roubaud Fr. (2016), Révolution des données et enjeux de la statistique en Afrique. Introduction thématique, dans Bédécarrats Fl., Cling J.-P. et Roubaud Fr. (eds), *Gouverner par les nombres en Afrique* (dossier), *Afrique contemporaine,* 9-23.

Breiman L. (2001), Statistical modeling: The two Cultures, *Statistical Science,* **16**, 199-215.

Brion Ph. (1995), Base de sondage : entre rigueur et bricolage, dans Vallin J. (éd.), *Clins d'œil de démographes à l'Afrique et à Michel François*, Paris, Centre français sur la population et le développement,117-124.

Cantrelle P. (1995), Quarante ans d'enquêtes démographiques en Afrique, dans Vallin J. (éd.), *Clins d'œil de démographes à l'Afrique et à Michel François*, Paris, Centre français sur la population et le développement, 101-115.

CEMUBAC (1972), *Enquête de Fuladu, 1959 : L'emploi du temps du paysan dans un village Zande du Nord-Est du Zaïre*, Université libre de Bruxelles, 89, Edition Cemubac.

Clairin R. (1988), Le recensement des nomades, dans Lohle-Tart L. et Clairin R. (eds), *De l'homme au chiffre. Réflexions sur l'observation démographique en Afrique*, Paris, CEPED, 169-174.

Clairin R. et Brion Ph. (1997), *Manuel de sondages. Applications aux pays en développement*, 2e édition, Paris, Centre français sur la population et le développement.

Dehon C., Droesbeke J.-J. et Vermandele C. (2015), *Eléments de statistique*, sixième édition corrigée et augmentée, Bruxelles, Editions de l'Université de Bruxelles, Paris, Ellipses.

Desrosières, A. (1988), La partie pour le tout : comment généraliser? La préhistoire de la contrainte de représentativité, *Journal de la Société de Statistique de Paris*, **129**, 96-115.

Desrosières A. (1993), *La politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte.

Devaradjan S. (2013), Africa's Statistical Tragedy, *Review of Income and Wealth*, **59**(1), 9-15.

Donoho D. (2015), *50 Years of Data Science*, Tukey Centennial Workshop, http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Droesbeke J.-J. (2011), Les techniques d'enquête dans les pays en développement : la démarche historique est-elle porteuse de leçons ?, dans Tremblay M.-E., Lavallée P. et El Haj Tirari M. (éds.), *Pratiques et méthodes de sondage*, Paris, Dunod, 1-10.

Droesbeke J.-J. (2015), La donnée : des tablettes sumériennes aux Big Data, *Variances*, **53**, 16-21.

Droesbeke J.-J. et Tassi Ph. (2015), *Histoire de la statistique*, Que-sais-je ?, 2e édition rééditée, Paris, Presses Universitaires de France.

Droesbeke J.-J. et Vermandele C. (2016), *Les nombres au quotidien. Leur histoire, leurs usages*, Collection *La statistique autrement*, Paris, Technip.

Droesbeke, J.-J. et Vermandele, C. (2018a), *Histoire(s) de(s) données numériques*, Collection *Le monde des données*, Paris, EDP Sciences.

Droesbeke, J.-J. et Vermandele, C. (2018b), Faciliter l'enseignement de la moyenne arithmétique en se servant de son histoire, *Statistique et Enseignement*, **9**, 1, 99-120.

Droesbeke J.-J. et Vermandele C. (2019), *Ce que nous disent les sondages*, Collection *L'Académie en poche*, Bruxelles, Académie royale de Belgique.

Hecht J. (1987), L'idée de dénombrement jusqu'à la révolution, dans Affichar, J. (éd.), *Pour une histoire de la statistique*, **1**, Paris, Economica, 21-81.

Locoh Th. Et Omoluabi E. (1995), Où sont donc passés les 30 millions de Nigérians manquants ?, dans Vallin J. (éd.), *Clins d'œil de démographes à l'Afrique et à Michel François*, Paris, Centre français sur la population et le développement, 57-75.

Moultrie T.A. (2016), Démographie, démographes et « révolution des données » en Afrique, dans Bédécarrats Fl., Cling J.-P. et Roubaud Fr. (eds), *Gouverner par les nombres en Afrique* (dossier), *Afrique contemporaine*, 25-39.

Nordman C.J. et Roubaud F. (2010), Une approche originale en économie du développement : 20 ans d'efforts pour mesurer et analyser l'économie informelle dans les pays en développement, *Dialogue*, 31, 2-9.

Paccou Y. et Blanc R. (1979), Le recensement des nomades mauritaniens, *Populations*, 34 (2), 343-377.

Romaniuk A. (2006), *Démographie congolaise au milieu du 20e siècle*, Louvain la Neuve, Presses universitaires de Louvain.

Saporta G. (2011), *Probabilités, analyse des données et statistique*, 3e édition révisée et augmentée, Paris, Technip.

Saporta G. (2017), Quelle statistique pour les Big Data ?, entretien avec Gilbert Saporta, *Statistique et Société*, **5**, 1, avril 2017.

Stengers J. (2007). *Congo. Mythes et réalités*? 2e édition, Bruxelles, Racine.

Théodore G. (1995), Cinquante ans après, dans Vallin J. (éd.), *Clins d'œil de démographes à l'Afrique et à Michel François*, Paris, Centre français sur la population et le développement, 7-27.