RAF-AFCAS-21-F-54 – FAO **F**

Novembre 2021

منظمة الأغذية والزراعة للأمم المتحدة | 联合国粮食及农业组织 | Food and Agriculture Organization of the United Nations | Organisation des Nations Unies pour l'alimentation et l'agriculture | Продовольственная и сельскохозяйственная организация Объединенных Наций | Organización de las Naciones Unidas para la Agricultura y la Alimentación

COMMISSION AFRICAINE DES STATISTIQUES AGRICOLES

Vingt-septième session

15 – 18 novembre 2021, Hôte virtuel – Dakar, Sénégal

POINT 5 DE L'ORDRE DU JOUR

Lignes directrices sur la désagrégation des indicateurs des ODD à partir des données d'enquête

NOMS DES AUTEURS
Piero Demetrio Falorsi, Statisticien principal
Clara Aida Khalil, Statisticien
Stefano Di Candia, Statisticien junior
Pietro Gennari, Chef Statisticien
ORGANISATION
Bureau du Chef Statisticien de la FAO

SYNTHÈSE

Les « ***Lignes directrices sur la désagrégation des indicateurs des ODD à partir de données d'enquête*** » (FAO, 2021) donnent un aperçu détaillé des méthodes et des outils d'enquête pour générer des estimations désagrégées relatives aux indicateurs des ODD en utilisant les enquêtes comme principale source de données. La publication des Lignes directrices constitue l'une des mesures prises par la FAO pour accompagner les pays membres dans le calcul des indicateurs des ODD désagrégés par groupes de population et par zones territoriales pertinents.

Les Lignes directrices permettent de surmonter les principales contraintes de la plupart des enquêtes par sondage que sont le fait d'avoir des échantillons qui ne sont pas assez larges pour assurer des estimations directes fiables pour toutes les sous-populations, ou qui ne couvrent pas tous les domaines de désagrégation envisagés. Les Lignes directrices définissent premièrement un cadre pour la promotion d'une approche globale relative à la désagrégation des données en décrivant des approches normalisées et innovantes pour faire face à ces contraintes lors des différentes phases du processus de production statistique. À l'étape de plan d'échantillonnage, elles décrivent une série de stratégies d'échantillonnage alternatives (le suréchantillonnage, la stratification plus poussée, l'échantillonnage à plusieurs phases avec filtrage des répondants et la stratification marginale) permettant d'assurer un nombre « suffisant » d'unités d'échantillonnage pour chaque domaine de désagrégation, mais qui s'accompagnent souvent d'une augmentation du coût et de la complexité des opérations statistiques. Lors de la phase d'analyse, en s'appuyant sur d'autres sources de données ou d'autres domaines, les Lignes directrices proposent une série d'approches d'estimation indirecte en vue de surmonter la contrainte associée au nombre limité d'informations disponibles sur ce qu'on appelle les petits domaines. À cet effet, les Lignes directrices présentent une approche d'estimation indirecte assistée par modèle qui permet d'intégrer les données de différentes enquêtes et recensements. L'estimateur décrit est opérationnalisé pour la production d'estimations synthétiques désagrégées relative à l'indicateur 2.1.2 des ODD : la Prévalence d'une insécurité alimentaire

modérée ou grave, évaluée selon l'échelle de mesure de l'insécurité alimentaire vécue (FIES*)*. Des méthodes et des logiciels d'évaluation de l'exactitude des estimations désagrégées sont fournies aussi bien pour les approches d'estimation directe qu'indirecte. Finalement, les Lignes directrices s'achèvent par un aperçu général des méthodes d'estimation de petits domaines (SAE) en présentant les étapes clés de leur application.

1. INTRODUCTION

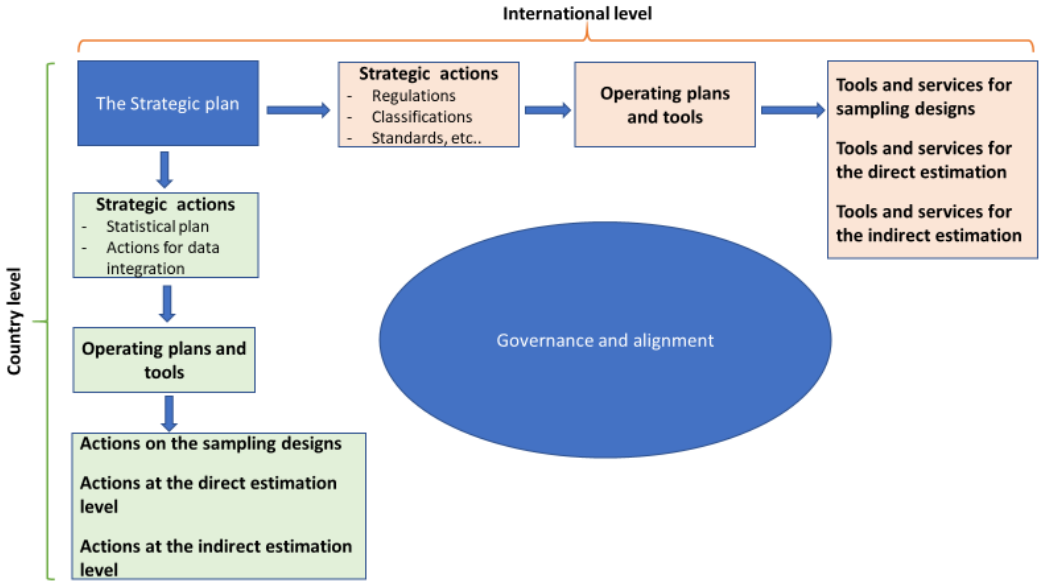
Avec l'adoption du principe « Ne laisser personne derrière » (LNOB) comme objectif transversal du Programme de développement durable à l'horizon 2030, les États membres des Nations Unies se sont engagés à éradiquer la pauvreté sous toutes ses formes, à éliminer la faim, à mettre fin à la discrimination et à réduire les inégalités et les vulnérabilités. Il est nécessaire de mesurer avec précision les progrès réalisés et de définir des indicateurs des Objectifs de développement durable (ODD) désagrégés par dimensions multiples pour suivre tous les objectifs au niveau de tous les groupes de population et de toutes les zones géographiques concernés. À cet effet, en déclarant que : « *Les indicateurs des ODD devraient être ventilés, le cas échéant, par revenu, sexe, âge, race, appartenance ethnique, statut migratoire, handicap et le lieu de résidence, ou d'autres caractéristiques, conformément aux Principes fondamentaux de la statistique officielle* », la Commission de statistique des Nations Unies (CSNU) – en charge de l'élaboration du cadre global pour mesurer les ODD - a adopté un principe général de désagrégation des données dans le cadre de l'élaboration du cadre mondial des indicateurs pour le suivi des ODD et de leurs cibles.

Néanmoins, la production d'estimations désagrégées de haute qualité relatives aux indicateurs des ODD représente un défi majeur pour les systèmes statistiques nationaux (SSN), en termes d'exigences en matière de données et de complexité opérationnelle. C'est pour cela que, lors de sa quarante-septième session, le Conseil de sécurité des Nations Unies a instruit le Groupe Inter-agences et d'Experts des Nations Unies sur les indicateurs relatifs aux objectifs de développement durable de former un groupe de travail (GT) sur la désagrégation des données, dans le but de renforcer les capacités nationales et d'élaborer des normes et outils statistiques nécessaires à la production de données désagrégées. Cela a conduit, entre autres résultats, à la compilation par les institutions garantes des principales catégories de désagrégation des indicateurs des ODD, ainsi qu'à l'identification d'un ensemble de politiques prioritaires ciblant les populations les plus vulnérables.

Les Lignes directrices de la FAO sur la désagrégation des indicateurs des ODD à partir des données d'enquête (FAO 2021) constituent, en ce sens, l'une des mesures prises par la FAO – en tant que membre du groupe de travail sur la désagrégation des données – pour soutenir les pays membres dans la production d'indicateurs relatifs aux ODD désagrégés par différents groupes de population et zones territoriales. Ainsi, ces lignes directrices proposent des conseils méthodologiques et pratiques pour produire des estimations désagrégées directes et indirectes relatives aux indicateurs des ODD en utilisant les enquêtes comme la source de données principale ou préférée et pour l'évaluation de l'exactitude des estimations.

Les Lignes directrices encouragent une approche détaillée en matière de désagrégation des données (Figure 1.1). Cette approche permet d'impliquer aussi bien les acteurs nationaux qu'internationaux dans la formulation d'un plan stratégique convenu pour une utilisation intégrée des diverses approches, méthodologies et outils statistiques suivant les différentes étapes de la chaîne de production statistique. Ces plans stratégiques influencent et orientent toutes les actions menées à un niveau plus technique, telles que celles associées aux étapes d'échantillonnage et d'estimation. Les bureaux nationaux de statistiques (ONS) sont les acteurs clés de tous les plans stratégiques relatifs à la désagrégation des données au niveau national. Au même moment, les organisations internationales peuvent renforcer la coordination en encourageant l'adoption de normes statistiques et de méthodologies communes pour garantir une meilleure qualité et comparabilité des statistiques désagrégées.

Figure 1.1. Approche détaillée en matière de désagrégation de données



Les Lignes directrices débutent avec un exposé sur les défis que constitue la désagrégation des données dans le contexte de l'Agenda 2030 des ODD. Ensuite, elles proposent des solutions techniques pour définir des stratégies d'échantillonnage pour l'estimation directe du domaine et des méthodes s'appuyant sur l'utilisation d'informations auxiliaires. Elles proposent également des plans d'échantillonnage qui garantissent un nombre suffisant d'unités d'échantillonnage pour chaque sous-population ou pour chaque domaine nécessitant des données désagrégées, permettant ainsi le calcul des estimations directes désagrégées. En outre, elles fournissent des méthodes pour mesurer la précision de l'échantillonnage. L'estimation et la diffusion d'indicateurs de qualité permettant d'évaluer l'exactitude des estimations représentent une étape fondamentale dans la production d'estimations désagrégées et pourraient accroître la transparence des BNS et, par conséquent, renforcer la confiance du public par rapport aux statistiques officielles. De plus, les estimations directes contenant des erreurs d'échantillonnage importantes sont une preuve qu'il faudrait soit faire recours à des techniques d'estimation des petits domaines, ou soit revoir le plan d'échantillonnage adopté.

Une grande partie des Lignes directrices est consacrée à la présentation d'une approche indirecte de production d'estimations désagrégées en s'appuyant sur l'utilisation intégrée de deux enquêtes indépendantes. Cette méthode permet d'intégrer une enquête de petite taille, mesurant une cible variable et ayant une petite erreur de mesure, à une enquête plus étendue, collectant des variables à des fins générales, dont au moins une est étroitement liée à la cible variable (variable approximative).

Les Lignes directrices se terminent par un aperçu des techniques d'estimation des petits domaines (SAE), comme étant l'une des approches pouvant être utilisées pour produire des estimations indirectes désagrégées. Étant donné que les approches SAE sont essentiellement basées sur des hypothèses de modèle, la validation et l'interprétation de leurs résultats peuvent être difficiles.

Suite à la publication de ces lignes directrices, le Bureau du statisticien en chef de la FAO continue d'élaborer des approches de désagrégation des données et d'estimation indirecte à travers des études de cas pratiques supplémentaires en matière de désagrégation des données relatives aux indicateurs des ODD dont la FAO est l'institution garante. Le rapport technique « *Utilisation de l'estimateur de projection pour la désagrégation des données relatives aux indicateurs des ODD sur la base des données d'enquête* » *(FAO, 2021b)*[1] présente en particulier une étude de cas – basée sur des micro-données du Guatemala, du Malawi et de l'Afrique du Sud, élargissant ainsi l'exercice pratique présenté par les Lignes directrices sur la désagrégation des données et offrant un guide séquentiel pour sa reproduction. À chaque étape, les routines logicielles pour son utilisation sont listées et expliquées. En outre, le rapport technique vient enrichir l'application empirique présentée dans les Lignes directrices, en fournissant les outils pour évaluer l'exactitude des estimations désagrégées indirectes.

[1] Le rapport technique est en cours de révision et sera publié dans les semaines à venir.

2. PLANIFICATION DE LA DÉSAGRÉGATION DES DONNÉES LORS DE LA PHASE DE CONCEPTION DE L'ENQUÊTE

Pour produire des estimations désagrégées directes, le plan d'échantillonnage doit avoir une taille d'échantillon planifiée dans chaque domaine de désagrégation. L'existence d'unités d'échantillonnage dans chaque domaine de désagrégation favorise également la production d'estimations indirectes grâce à la réduction significative du biais du modèle. Comme l'indique Kalton (2009), la sélection de la taille requise de l'échantillon du domaine est relativement simple lorsque la base d'échantillonnage permet de déterminer l'appartenance à une sous-population (ou domaine) rare. Dans de tels cas, le problème majeur serait comment déterminer l'étendue du suréchantillonnage à appliquer pour atteindre le niveau cible d'exactitude des estimations dans chaque domaine de désagrégation. L'échantillonnage et le suréchantillonnage de domaines rares dont les membres ne peuvent pas être identifiés à l'avance représentent un défi majeur. Diverses méthodes ont été utilisées dans de telles situations. Outre la sélection à grande échelle, ces méthodes comprennent l'échantillonnage stratifié disproportionné, l'échantillonnage à deux phases, l'utilisation de cadres d'échantillonnage à bases multiples, l'échantillonnage multiple et l'échantillonnage par emplacement. Les techniques d'échantillonnage traditionnelles abordent la question de la désagrégation des données en suréchantillonnant ou en introduisant une stratification plus poussée. Des techniques plus complexes permettent d'améliorer les plans d'échantillonnage en répartissant géographiquement les unités d'échantillonnage (Gräfstorm, Lundström and Schelin, 2012) et en réduisant le niveau de regroupement. Cela permettrait d'atteindre les sous-populations isolées ou rares. En général, les techniques d'échantillonnage traditionnelles présentent des contraintes dans le cas des sous-populations rares (Kalton, 2009). Kish (1987) a proposé une classification des domaines de désagrégation sur la base de leur taille relative par rapport à la population totale. Il identifie comme « domaines majeurs » ceux qui représentent approximativement au moins 10 % de la population totale. Un plan d'échantillonnage traditionnel devrait normalement produire des estimations fiables dans le cas des domaines majeurs. Les « petits domaines » sont ceux représentant 1 à 10 % de la population totale et ils requièrent des approches d'échantillonnage spéciales pour garantir une taille d'échantillonnage suffisante. Les « mini-domaines » comprennent ceux représentant 0,1 à 1 % de la population totale et nécessitent l'utilisation de modèles statistiques pour des estimations fiables. Enfin, les « domaines rares », comprenant ceux qui représentent moins de 0,1 % de la population totale, et ne peuvent pas faire l'objet d'estimation en utilisant des méthodes d'échantillonnage d'enquête.

Il existe également des contraintes liées aux populations difficiles à atteindre (les sans-abris, les migrants ou les populations nomades) ou hors d'atteinte. La pertinence de cette contrainte en ce qui concerne la désagrégation des données est mise en évidence par l'indicateur 2.3.1 (Volume de production par unité de travail, en fonction de la taille de l'exploitation agricole, pastorale ou forestière) devant être ventilé, entre autres, par type d'exploitation (agricole/pastorale/foresterie/halieutique). Toutefois, certaines des contraintes majeures sont associées au fait que très souvent, les données relatives aux activités forestières, halieutiques et pastorales ne sont pas collectées lors des enquêtes agricoles. Il serait alors nécessaire d'harmoniser différentes sources de données. La conception d'une enquête relative aux activités pastorales ou les incluant, impliquerait dans de nombreux pays en développement, la collecte de données relatives aux populations nomades – qui sont souvent très difficiles à localiser (FAO, 2016). De nouvelles approches récemment publiées et relative à l'échantillonnage permettent de surmonter certaines des contraintes mentionnées ci-dessus. Ce sont, entre autres, l'échantillonnage indirect ou multi-source (FAO, 2014 et 2015 ; Lavallée, 2007 ; Singh et Mecatti, 2011) ou l'échantillonnage à stratification marginale (Falorsi et Righi, 2008, et Falorsi Lavallée et Righi 2019). Les Lignes directrices présentent en détail ces différentes approches (FAO, 2021).

Les sections 2.1 et 2.2 ci-dessous donnent un bref aperçu de certaines des techniques d'échantillonnage abordées par les Lignes directrices.

2.1. Techniques traditionnelles d'échantillonnage pour la désagrégation des données

Suréchantillonnage

Le suréchantillonnage permet de définir une taille plus importante de l'échantillon global initial. Ceci entraine à son tour un échantillon de taille plus large au niveau du domaine. Si la taille initiale de l'échantillon augmente d'une proportion Δ, cela devrait avoir un impact sur l'augmentation de la taille de l'échantillon du domaine qui serait égal à $n\Delta P_d$, avec $P_d = N_d/N$, la taille relative du domaine d. Le Tableau 2.1 représente l'augmentation

de la taille de l'échantillon du domaine n_d causée par une augmentation du pourcentage Δ de la taille globale de l'échantillon de 10.000 ménages par différentes proportions de sous-populations.

Tableau 2.1. Augmentation de la taille de l'échantillon du domaine n_d causée par une augmentation Δ de la taille initiale de l'échantillon (10.000 ménages) par la taille relative P_d du domaine

Δ	P_d%			
	0,05%	**1%**	**5%**	**10%**
10%	5	10	50	100
50%	25	50	250	500
100%	50	100	500	1.000

Nous constatons que le suréchantillonnage peut être utile dans le cas des domaines majeurs (Kish, 1986). Par contre, cette approche est moins pérenne lorsqu'il s'agit des petits domaines et des mini-domaines. En outre, si le domaine de désagrégation n'est pas pris en compte lors de la phase de plan d'échantillonnage, le résultat du suréchantillonnage n'est pas certain, puisque la taille de l'échantillon du domaine obtenue peut être différente des prévisions. Le Tableau 2.2 illustre la taille globale n requise de l'échantillon pour garantir la taille minimum acceptable n_d^*, pour différentes valeurs de n_d^* et de la proportion P_d de la sous-population. Il est évident que pour atteindre les n_d^* requises pour les sous-populations rares, ($P_d \leq 1$ %), la taille globale de l'échantillon devrait être considérablement plus grande et impossible à gérer par la plupart des enquêtes nationales.

Tableau 2.2. La taille *n* d'échantillon requis pour assurer le seuil minimum n_d^* des différentes valeurs de la proportion P_d de la sous-population

n_d^*	%P_d			
	0.05%	**1%**	**5%**	**10%**
30	62,000	31,000	6,200	3,100
50	102,000	51,000	10,200	5,100
100	202,000	101,000	20,200	10,100

Stratification plus poussée

La stratification par domaine de désagrégation est la stratégie traditionnelle adoptée pour contrôler la taille n_d de l'échantillon durant l'étape du plan d'échantillonnage. Cela implique l'inclusion des variables d'appartenance au domaine γ_{di} (avec $\gamma_{di} = 1$ if $i \in U_d$ et $\gamma_{di} = 0$) parmi celles devant être utilisées pour la stratification. Toutefois, dans de nombreuses situations pratiques la classification croisée des variables de la stratification est inappropriée, car elle requiert la sélection d'un certain nombre d'unités d'échantillonnage devant être au moins approximativement aussi grand que le produit du nombre de catégories des variables de stratification. De plus, pour obtenir des estimations non biaisées de la variance d'échantillonnage, il faut sélectionner au moins deux unités par strate. Cochran (1977) a bien illustré ce problème à travers un exemple clair d'un plan de classification croisée irréalisable. Une combinaison de stratification explicite et implicite est souvent utilisée dans les enquêtes pour prendre en compte des variables supplémentaires que la stratification traditionnelle ne pourrait pas prendre en compte. Dans le cas des domaines majeurs non planifiés, la stratification implicite peut faciliter l'estimation.

Falorsi et Righi (2015) donnent l'exemple de stratégies d'échantillonnage optimales avec des informations a priori (incertaines) sur le taux des populations rares dans les strates. Cette stratégie a permis d'identifier la solution la moins coûteuse qui consiste à faire le suréchantillonnant uniquement dans les strates susceptibles d'avoir une plus grande proportion de la sous-population rare. Ces stratégies peuvent être mises en œuvre en utilisant le logiciel Mauss-R, qui permet l'allocation multivariée d'unités dans les enquêtes par sondage.

Échantillonnage à plusieurs phases avec sélection des répondants

La stratégie basée sur une stratification plus poussée nécessite la disponibilité des variables d'appartenance au domaine γ_{di} dans le cadre d'échantillonnage. Cela peut être le cas pour les variables géographiques, mais pas pour de nombreuses autres variables de désagrégation telles que le quintile de revenu, le statut de migrant ou d'autochtone.

Une stratégie d'échantillonnage traditionnelle pour surmonter ce problème consiste à sélectionner un échantillon de première phase $S_{(1)}$ de taille $n_{(1)}$. Les variables d'appartenance γ_{di} sont ensuite collectées à partir des unités d'échantillonnage de S_1. Un échantillon stratifié $S_{(2)}$ est alors sélectionné afin de garantir les tailles finales prévues n_d $(d = 1, \dots, D)$ des échantillons. Le coût de la sélection devient une préoccupation majeure puisqu'il faut des échantillons de sélection de très grande taille pour produire une taille adéquate d'échantillon de domaine lorsque l'un (ou plusieurs) des domaines ciblés est une population rare. Plusieurs stratégies peuvent être utilisées pour maintenir les coûts à un niveau raisonnable (Kalton, 2009) : (i) utiliser un mode de collecte de données peu coûteux, tel que des systèmes d'interview par téléphone ou en ligne, pour la sélection ; (ii) permettre la collecte d'informations relatives à la sélection auprès d'unités non incluses dans l'échantillon de sélection ; et (iii) Sélectionner une grande taille d'échantillon dans chaque groupe pour en augmenter l'efficacité lorsque la sélection est effectuée par entretien en face à face dans un plan à plusieurs phases.

2.2. Techniques d'échantillonnage innovantes pour la désagrégation des données

Plans de stratification marginale

Les publications relatives aux plans d'échantillonnage proposent diverses méthodes pour contrôler la taille de l'échantillon dans toutes les catégories de variables de stratification sans utiliser un plan de classification croisée. Ces méthodes sont généralement appelées techniques de stratification multidimensionnelle et ont été développées selon deux approches principales : (i) les schémas des carrés latins ou en treillis (Jessen 1978) ; et (ii) des problèmes d'arrondi contrôlé grâce à la programmation linéaire (Lu et Sitter, 2002). Ces deux approches présentent des inconvénients qui ont limité l'utilisation des techniques de stratification multidimensionnelle comme solution standard lors de la planification de plans d'échantillonnage lors d'enquêtes réelles. La faiblesse principale de l'approche de programmation linéaire est la complexité de son calcul. La stratégie d'échantillonnage proposée ci-dessous, basée sur un échantillonnage équilibré, n'est pas affectée par les inconvénients des méthodes susmentionnées et permet de contrôler la taille de l'échantillon de divers domaines cibles de désagrégation tels que définis par différentes partitions de la population de référence. En outre, elle veille à ce que les erreurs d'échantillonnage des estimations de domaine soient inférieures à un seuil prédéfini. Pour définir l'échantillonnage équilibré dans le plan d'échantillonnage ou dans l'approche assistée par modèle, introduisons la définition générale du plan d'échantillonnage comme une distribution de probabilité p $(\cdot)$ sur l'ensemble $\mathcal{S}$ de tous les échantillons S de la population U. Soit x_i un vecteur de variables auxiliaire x disponible pour chaque unité de population. Le Plan d'échantillonnage $p(S)$ avec des probabilités d'inclusion $\pi = \{\pi_i : i = 1, \dots, N\}$ est dit équilibré par rapport aux variables auxiliaires si et seulement s'il respecte les équations d'équilibre

$$\sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i \quad (2.1)$$

avec tous les $S \in S$ tel que $p(S) > 0$. Supposons qu'un vecteur des probabilités d'inclusion π conforme aux échantillons de distribution marginale n_d $(d = 1, \dots, D)$ soit disponible, c'est-à-dire

$$\sum_{i \in U} \pi_i \gamma_{di} = n_d \ (d = 1, \dots, D), (2.2)$$

avec D le nombre total de domaines pour lesquels l'on doit produire de données désagrégées. Les plans de stratification multidimensionnelle constituent un cas spécial de plans équilibrés dans lequel pour une unité i, le vecteur de variable auxiliaire est :

$$x_i = \pi_i \gamma_i \ (2.3)$$

avec γ_i le vecteur D des variables $\gamma_i = (\gamma_{1i}, \dots, \gamma_{di}, \dots, \gamma_{Di})'$. d'appartenance au domaine.
En définissant le vecteur x_i tel qu'exprimé en 2.3, si la condition définie au 2.2 est respectée, alors la sélection d'échantillons respectant le système des équations d'équilibre de 2.3 garanti que les valeurs de n_d ne soient pas des quantités non aléatoires.
Le côté gauche de l'équation d'équilibre 2.1 est :

$$\sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in U} \frac{\pi_i}{\pi_i} \gamma_{di} \lambda_i = \sum_{i \in U} \gamma_{di} \lambda_i = n_d. \ (d = 1, \dots, D)$$

Le côté droit est :

$$\sum_{i \in U} x_i = \sum_{i \in U} \pi_i \gamma_{di} = n_d \ (d = 1, \dots, D).$$

Tillé (2020) a proposé la méthode du cube qui permet de sélectionner des échantillons équilibrés (ou approximativement équilibrés) pour un grand ensemble de variables auxiliaires et par rapport à différents vecteurs de probabilités d'inclusion. En particulier, Deville et Tillé (2005) montrent qu'avec des vecteurs x_i respectant le 2.3, l'équation d'équilibre en 2.1 peut être résolue avec précision. La méthode du cube s'applique grâce à un algorithme amélioré pour les grands ensembles de données (Chauvet et Tillé, 2006) disponible dans un code logiciel gratuit.[2]

Il est important de noter que l'échantillonnage équilibré constitue la base de la définition de grande classe de plans d'échantillonnage. Par exemple, les plans d'échantillonnage stratifié exigent que :

$$\sum_{d=1}^{D} \gamma_{di} = 1,$$

chaque U_d est appelé une strate. La section 3.5.4 des Lignes directrices (FAO, 2021) montre comment élaborer des plans de stratification marginale pour les plans d'échantillonnage à deux phases ou à deux étapes qui constituent les deux stratégies les plus utilisées dans les enquêtes réelles.

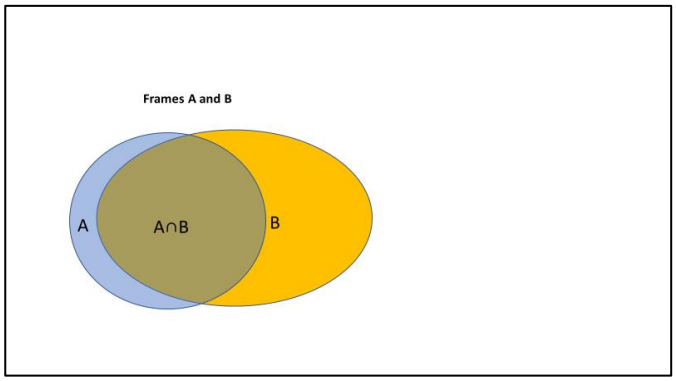
Échantillonnage indirect

Dans toute enquête conventionnelle, la sélection aléatoire d'un échantillon nécessite la disponibilité d'une base de sondage actualisée qui enregistre toutes les unités de la population cible éligibles à l'enquête ; chacune d'elles est identifiée par une étiquette. Lorsque la base d'échantillonnage est disponible, une préoccupation statistique cruciale est l'évaluation de la couverture réelle fournie par cette liste de la population cible. Une base d'échantillonnage est parfaite lorsque ses éléments sont associés un à un aux éléments de la population cible. Cependant, dans la pratique statistique, il est rare d'avoir des bases parfaites et il y a toujours des problèmes qui perturbent la correspondance idéale. Par exemple, il peut exister au sein d'une base d'échantillonnage un sous-dénombrement ou un sur-dénombrement, voire les deux. On parle de sous-dénombrement lorsqu'une base disponible est incomplète, c'est-à-dire qu'elle ne comprend qu'une partie de la population cible. Par conséquent, les éléments manquants n'apparaissent dans aucun échantillon provenant de la base. Par contre, il y a sur-dénombrement lorsqu'une base contient des doublons de la même unité ou des unités n'appartenant pas à la population cible. Dans la pratique, une base d'échantillonnage peut présenter également des imperfections d'autres types : par exemple, dans certaines circonstances, l'ensemble d'unités souhaitées ne serait pas disponible, mais l'on aurait plutôt un cadre d'unités lié d'une manière ou d'une autre à la liste des unités cibles. En outre, il est possible qu'une base soit disponible, mais dans un environnement dynamique, elle devient rapidement obsolète, représentant ainsi une situation probablement très différente de la réalité. Dans tous ces cas, la stratégie suivante peut être adoptée : à partir de l'observation d'une population, les unités de la population liée et ciblée font l'objet d'une enquête suivant leurs liens avec les unités de la première population. Ainsi, comme cela se produirait avec une approche d'échantillonnage indirect, les populations cibles peuvent être considérées comme échantillonnées à partir d'une base de sondage imparfaite, c'est-à-dire la base se référant à la population initiale. Pour identifier ces liens, les questionnaires doivent être structurés de manière appropriée. La FAO (2015) illustre les modules et les règles opérationnelles pour l'application de l'échantillonnage indirect dans les enquêtes agricoles.

L'échantillonnage à sources multiples est également une approche pratique dans le cas des bases de sondage imparfaites, en particulier lorsque la population cible se retrouve à l'intersection de deux bases voire plus de deux. Ce cas est illustré par la figure 2.1 ci-dessous, qui montre deux cadres qui se chevauchent partiellement. Un exemple pertinent ici est celui des enquêtes agricoles portant sur les exploitations du secteur ménage et non-ménage. Dans certaines circonstances, certaines exploitations peuvent relever de deux cadres différents, celui des ménages et celui des personnes morales.

[2] Disponible sur: http://www.insee.fr/fr/nom_df_met/outils_stat/cube/accueil_cube.htm

Figure 2.1. Exemple d'échantillonnage à plusieurs sources: la population cible couverte par l'intersection des deux sources



Comme l'on peut le voir sur la Figure 2.1, si un échantillon S^A est sélectionné dans la base A et un échantillon indépendant S^B est sélectionné dans la base B, les unités qui se retrouvent à l'intersection A ∩ B des deux bases peuvent être incluses dans les deux échantillons. La FAO (2014) propose une approche méthodologique qui étend l'utilisation de l'échantillonnage indirect (Lavallée, 2007) pour couvrir la production d'estimations intégrées relative à plusieurs populations cibles dans le cadre des enquêtes à base multiple (Hartley, 1974 ; Singh et Mecatti, 2011). Les techniques proposées sont relativement flexibles. En outre, dans des conditions générales, elles permettent la production de statistiques non biaisées, surmontant ainsi la plupart des problèmes causés par les bases de sondage imparfaites. Les Lignes directrices (FAO, 2021) montrent comment ces approches peuvent être combinées grâce au concept de multiplicité qui fut introduit pour la première fois par Birnbaum et Sirken (1965) lorsqu'ils présentaient l'échantillonnage en réseau comme une stratégie d'enquête relative aux populations rares ou inaccessibles.

3. LA DÉSAGRÉGATION DES DONNÉES LORS DE LA PHASE D'ANALYSE

3.1. L'estimateur de projection

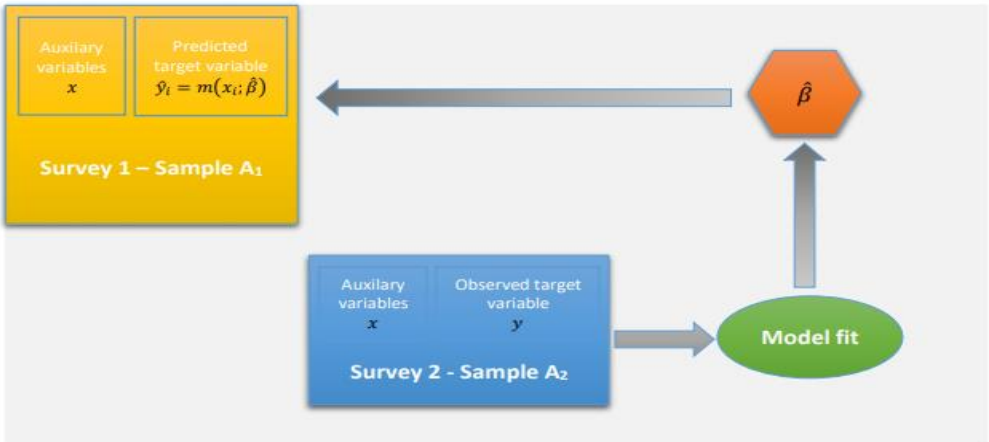
Lors de la phase d'analyse, la désagrégation des données peut se faire en adoptant des approches d'estimation indirecte afin de surmonter le manque d'informations disponibles sur ce que l'on appelle les petits domaines. Ce faisant, l'on tire parti de la force des domaines complémentaires. En particulier, l'utilisation intégrée de différentes sources de données permet d'avoir une approche efficace pour assurer le niveau de désagrégation désiré en préservant l'exactitude des estimations. Les sources de données typiques pouvant être intégrées aux données d'une enquête agricole et/ou sur les ménages sont : 1) d'autres enquêtes ; 2) les recensements ; 3) les registres administratifs ; 4) les informations géospatiales ; et 5) les mégadonnées. Les approches d'estimation indirecte vont des approches axées sur un modèle aux approches assistées par modèle.

Des différentes méthodes disponibles pour la production d'estimations indirectes, les Lignes directrices (FAO, 2021) présentent et appliquent ce que l'on appelle « l*'estimateur de projection* » (Kim et Rao, 2012). Cette approche assistée par modèle (Figure 3.1) permet d'intégrer les données de deux enquêtes indépendantes par sondage – ou d'une enquête par sondage et d'un recensement – à la première enquête caractérisée par un échantillon A_1 de grande taille sans toutefois collecter des informations auxiliaires ou des variables d'usage général (exemple : variables socio-économiques) ; tandis que la deuxième enquête a un échantillon plus petit A_2 mais permet la collecte d'informations sur la variable cible y, ainsi que le même ensemble de variables auxiliaires disponibles dans A_1. Dans ce cadre statistique, le total de la variable y dans le domaine de désagrégation d peut être obtenu par la formule :

$$\hat{Y}_{PR,d} = \sum_{i \in A_1} w_{i1} m(x_i;\, \hat{\beta}) \gamma_{id}\,, (3.1)$$

w_{i1} étant le poids d'échantillonnage de l'unité i dans l'enquête A_1, $m(x_i;\, \hat{\beta})$ la valeur prédite de la variable y en utilisant le paramètre de régression $\hat{\beta}$ estimé à partir de l'enquête A_2,et γ_{id} le domaine d'appartenance au domaine, c'est-à-dire une variable muette dont la valeur est égale à 1 si l'unité i appartient au *d-ième* domaine.

Figure 3.1: Mise en œuvre de l'estimateur de projection



Ce cas couvre un grand nombre d'éventuelles situations empiriques pertinentes pour la désagrégation des données. En effet, la plupart des pays disposent d'au moins une enquête à grande échelle collectant des variables d'usage général, telles que les recensements, les enquêtes auprès des ménages, mais aussi les registres administratifs. D'autre part, l'utilisation d'enquêtes à grande échelle pour mesurer certaines des variables cibles à ventiler dans le contexte des ODD est trop coûteuse. Dans ces circonstances, une solution éventuelle pourrait être de mesurer le phénomène pertinent à l'aide d'une enquête à petite échelle pour ensuite améliorer la précision des estimations en s'appuyant sur des informations auxiliaires recueillies par une enquête à plus grande échelle. La seule condition de mise en œuvre de cette approche est que les deux enquêtes doivent avoir le même ensemble de variables auxiliaires utilisées pour ajuster le modèle de régression.

Dans de nombreux cas, les indicateurs ODD basés sur des données d'enquête ont la forme fonctionnelle suivante :

$$R_d = \frac{Y_d}{Z_d}$$

avec

$$Z_d = \sum_{i=1}^{N} z_i \gamma_{di},$$

z_i étant la valeur de la variable z sur l'unité i, la variable z est observée dans l'enquête A_1. Dans tous ces cas, l'estimateur de projection peut être aussi exprimé sous forme de ratio :

$$\hat{R}_{PR,d} = \frac{\hat{Y}_{PR,d}}{\hat{Z}_d} \quad (3.2)$$

avec $\hat{Y}_{PR,d}$ défini dans la formule 3.1 et

$$\hat{Z}_d = \sum_{i \epsilon A_1} \omega_{i1} z_i \gamma_{di}$$

Est l'estimation directe du total Z_d obtenu à partir de l'enquête A_1. Lorsque $z_i = 1$, la formule en 3.2 donne l'estimateur de projection d'une proportion. Afin d'étudier les propriétés asymptotiques de l'estimateur (3.2), nous considérons son approximation linéaire, donnée par le premier ordre de l'approximation en série de Taylor :

$$\hat{R}_{PR,d} = R_d + \frac{1}{Z_d}\left[\left(\hat{Y}_{p,d} - Y_{p,d}\right) - R_d\left(\hat{Z}_d - Z_d\right)\right] + o_i \quad (3.3)$$

avec o_i un reste d'un ordre mineur. À partir de l'expression en 3.3. et considérant la formulation de la variance de Kim et Rao (2012) la variance de l'échantillon de $\hat{Y}_p$ peut être exprimée comme suit :

$$Var\left(\hat{R}_{PR,d}\right) = Var\left(\sum_{i \epsilon A_1} w_{i1}\, t_{di}\right) + Var\left(\sum_{i \epsilon A_2} w_{i2}\left[y_i - m(x_i; \beta_0)\right]\right) \quad (3.4)$$

avec β_0 désignant l'estimation de β lors de l'observation de l'ensemble de la population, à savoir l'estimation que nous aurions obtenue en utilisant les données du recensement, et t_{di} la transformation de *Woodruff* (1971) :

$$t_{di} = \frac{1}{Z_d}\gamma_{di}[m(x_i;\beta_0) - R_d z_i].$$

L'estimateur plug-in sans biais asymptotique $Var\left(\hat{Y}_p\right)$ peut être dérivé en remplaçant la valeur de la super-population β_0 avec l'estimation $\hat{\beta}$, comme indiqué ci-dessous :

$$\hat{V}ar\left(\hat{R}_{PR,d}\right) = \hat{V}ar\left(\sum_{i\epsilon A_1} w_{i1}\,\hat{t}_{di}\right) + \hat{V}ar\left(\sum_{i\epsilon A_2} w_{i2}\left[y_i - m(x_i;\,\hat{\beta})\right]\right)$$

où $\hat{V}ar(\cdot)$ d désigne l'estimation d'échantillonnage de $Var(\cdot)$, et

$$\hat{t}_{di} = \frac{1}{\hat{Z}_d}\gamma_{di}\left[m(x_i;\hat{\beta}) - \hat{R}_{p,d} z_i\right].$$

Cette extension de l'approche de base présentée dans Kim et Rao (2012) permet d'adopter l'estimateur de projection pour de nombreux indicateurs des ODD pertinents pour la FAO, tels que :

- **Indicateur ODD 2.1.1** : Prévalence de la sous-alimentation ;
- **Indicateur ODD 2.1.2** : Prévalence de l'insécurité alimentaire modérée ou grave, évaluée selon l'échelle de mesure de l'insécurité alimentaire vécue ;
- **Indicateur ODD 2.3.1** : Volume de production par unité de travail, en fonction de la taille de l'exploitation agricole, pastorale ou forestière ;
- **Indicateur ODD 2.3.2** : Revenu moyen des petits producteurs alimentaires, selon le sexe et le statut d'autochtone ;
- **Indicateur ODD 5.a.1.a** (Proportion de la population agricole totale ayant des droits de propriété ou des droits garantis sur des terres agricoles (sur la population agricole totale) par sexe) et **5.a.1.b.** (proportion de femmes parmi les titulaires de droits de propriété ou de droits garantis sur des terrains agricoles, par type de droit)

L'approche basée sur l'estimateur de projection permet de produire des tableaux croisés de la variable cible y ainsi que des domaines de désagrégation non inclus initialement dans l'instrument de collecte de données utilisé pour obtenir A_2 (échantillon fournissant des informations relatives à y). Par exemple, supposons que l'on s'intéresse à l'estimation d'un paramètre lié à y, ventilé par statut d'autochtone. Supposons en outre que les informations relatives au statut d'autochtone des répondants ne soient pas disponibles en A_2, mais le sont uniquement en A_1. En projetant les valeurs de y sur A_1, il est possible d'utiliser les informations auxiliaires sur le statut d'autochtone pour estimer le paramètre cible en tenant compte de cette dimension de désagrégation. Enfin, il est important de souligner que l'estimateur de projection est un outil très flexible. Les praticiens des services nationaux de statistiques (SNS) et des organisations internationales peuvent adopter cette approche pour intégrer les données d'enquête aux différentes sources de données telles que les recensements, les dossiers administratifs et/ou les informations géospatiales. De plus, la prédiction d'une variable cible dans l'échantillon d'une enquête plus étendue à partir de laquelle la plupart des statistiques officielles nationales sont produites permet d'améliorer la cohérence des estimations.

3.2. Quelques résultats empiriques

Au sein de la FAO (2021b), l'estimateur de projection a été adopté pour produire des estimations désagrégées relatives à l'indicateur ODD 2.1.2 sur *la prévalence de l'insécurité alimentaire modérée et grave évaluée selon l'échelle de mesure de l'insécurité alimentaire vécue (FIES)* en utilisant les sources de données suivantes :

- La quatrième enquête intégrée sur les ménages du Malawi (IHS4) 2016-17 ;
- Le module d'enquête FIES du Malawi collecté par le Gallup World Poll (GWP) - 2016.

La FAO (2021b) élargit l'exercice pratique présenté dans les Lignes directrices sur la désagrégation des données et fournit un guide étape par étape pour une éventuelle réplication. À chaque étape, les scripts R pour

l'application font l'objet d'un rapport et sont expliqués ; en outre, ce travail enrichit l'application empirique présentée par la Directive en fournissant les outils requis pour apprécier l'exactitude des estimations désagrégées.

Étapes pour l'application de l'estimateur de projection

L'un des résultats les plus pertinents du rapport technique de la FAO portant sur la désagrégation des données (FAO, 2021b) est l'identification claire de toutes les principales étapes opérationnelles dans l'application de l'estimateur de projection. Ces étapes peuvent être résumées comme suit :

1. **Identification et recodage des variables auxiliaires.** L'application de l'estimateur par projection nécessite la disponibilité d'un même ensemble de variables auxiliaires dans les deux enquêtes à intégrer. Ces variables doivent également avoir la même structure et définition.
1. **Définition de la fonction $m()$ et des paramètres de l'estimation de la projection dans l'échantillon de petite taille.** La sélection de la forme fonctionnelle pour la fonction de lien $m()$ pour faire l'estimation des paramètres de projection dépend essentiellement du type de variable y considérée (exemple : échelle, nominale, dichotomique).
2. **Calcul de valeurs synthétiques.** En utilisant le paramètre de projection estimé, les valeurs synthétiques de la variable cible sont calculées dans l'ensemble de données de grande taille. Ceci permet en retour de produire des estimations indirectes désagrégées de l'indicateur cible.
3. **Évaluation de l'exactitude des estimations.** Après avoir produit des estimations synthétiques, leur précision peut être évaluée en estimant leur variance, leur coefficient de variation et leurs intervalles de confiance.

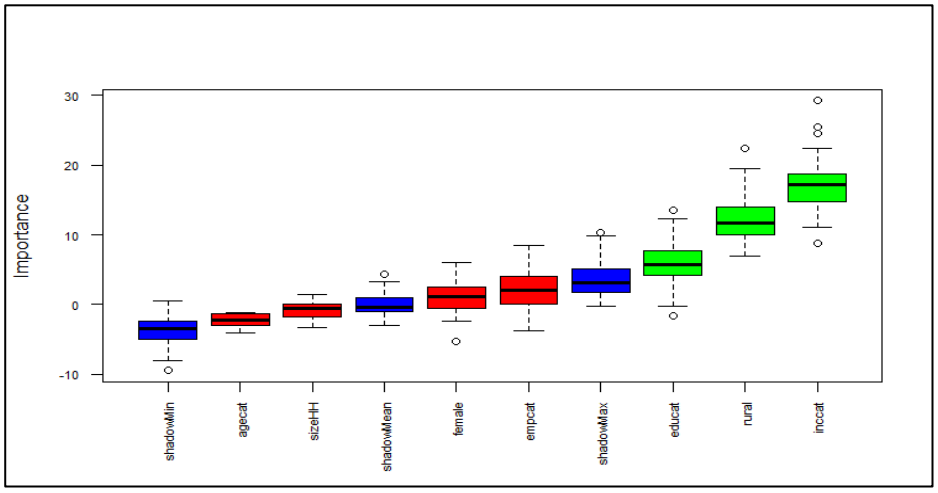
Étape 1 : Identification et recodage des variables auxiliaires

Une identification appropriée des variables auxiliaires x_i au niveau de l'enquête de petite taille constitue une étape clé pour assurer la qualité de l'estimateur de projection. Dans ce contexte, l'utilisation de méthodes de sélection de variables peut s'avérer utile lorsqu'il existe de nombreuses variables auxiliaires potentielles, bien que dans certains cas, il est possible que des problèmes de multi-colinéarité puissent rendre cette tâche plus complexe. Il existe un grand nombre de littératures sur les approches de sélection de variables. Ryan (2008) ou Harrel (2015) nous proposent par exemple une synthèse complète des méthodes courantes utilisées pour la sélection de variables auxiliaires dans les modèles de régression.

Malgré la disponibilité d'un nombre relativement faible de variables auxiliaires communes aux deux ensembles de données, la FAO (2021b) montre comment utiliser la méthode de sélection des caractéristiques de Boruta, proposée dans Kursa et Rudnicki (2010), à travers une approche enveloppée construite autour d'un classificateur de forêt aléatoire. (Breiman, 2001).

La figure 3.2 présente le résultat de Boruta, représentée par une série de diagrammes-boîtes de différentes couleurs : les diagrammes-boîtes rouges, jaunes et verts représentent respectivement les scores des variables auxiliaires rejetées (non importantes), provisoires et confirmées (importants), tandis que les diagrammes-boîtes bleus représentent les caractéristiques alternatives identifiées par l'algorithme. Les variables indicatives sont celles pour lesquelles Boruta n'a pas pu spécifier clairement leur pertinence, car leur niveau d'importance n'était pas significativement différent de leurs meilleures caractéristiques alternatives.

Figure 3.2. Niveau d'importance des variables auxiliaires pour l'insécurité alimentaire modérée et grave



Source : FAO, 2021b

Tous les niveaux de variables auxiliaires identifiés comme provisoires ou importants par Boruta ont été utilisés pour cadrer une régression logistique sur la base de la probabilité d'être en insécurité alimentaire modérée ou grave. En outre, toutes les dimensions pertinentes pour la désagrégation des données (sexe, tranche d'âge, revenu, situation rurale/urbaine) ont été prises en compte par le modèle de régression, afin d'augmenter l'absence de biais de l'échantillon de l'estimateur du domaine de projection.

Il est important de noter que l'une des conditions à remplir par les variables auxiliaires avant l'utilisation de la méthode de projection est qu'elles doivent avoir les mêmes définitions et une structure similaire dans les deux échantillons à intégrer. Par conséquent, avant d'utiliser cette approche d'estimation indirecte, toutes les variables auxiliaires sélectionnées ont été recodées et harmonisées entre les deux enquêtes telles que détaillées par la FAO (2021b).

Étape 2 : Définition de la fonction $m()$ et estimation des paramètres de projection au sein de l'échantillon de petite taille

Dans l'étude de cas, une régression logistique multivariée pondérée a été appliquée à l'échantillon de petite taille pour estimer les paramètres de projection $\hat{\beta}$ devant être utilisés pour la prévision de la valeur de la variable cible dans l'enquête à grade taille. Soit $\hat{p}_{ms,i}$ la probabilité pour l'individu $i-ième$ d'être en insécurité alimentaire modérée ou grave au sein de l'échantillon de petite taille. Cette probabilité a été estimée en utilisant les données GWP collectées grâce au module individuel de la FIES.

Puisque $\hat{p}_{ms,i}$ était concentré autour de quelques valeurs discrètes dans l'intervalle [0,1], alors, il a été recodé en une variable muette $y_{ms,i}$ telle que : $y_{ms,i} = 1$ if $\hat{p}_{ms,i} \geq 0{,}5$, et $y_{ms,i} = 0$.
Ensuite, les valeurs $y_{ms,i}$ ont été modélisées avec une fonction logistique multivariée de l'ensemble des variables auxiliaires catégorielles discrètes $x_i' = (1, x_{i1}, x_{i2}, \dots, x_{ik})$:

$$P(y_{ms,i} = 1|x_i) = m(x_i;\beta) = \frac{\exp(\beta_{ms,0} + \beta_{ms,1}x_{i1} + \beta_{ms,2}x_{i2} + \ \dots + \ \beta_{ms,k}x_{ik})}{1 + \exp(\beta_{ms,0} + \beta_{ms,1}x_{i1} + \beta_{ms,2}x_{i2} + \ \dots + \ \beta_{ms,k}x_{ik})},$$

with $\beta = (\beta_{ms,0}, \beta_{ms,1}, \beta_{ms,2}, \dots, \beta_{ms,k})$.

Étape 3 : Calcul des valeurs synthétiques dans l'échantillon de grande taille

Ayant obtenu les estimations $\hat{\beta} = (\hat{\beta}_{ms,0}, \hat{\beta}_{ms,1}, \hat{\beta}_{ms,2}, \dots, \hat{\beta}_{ms,k})$ des paramètres β en utilisant les outils statistiques normalisés, les probabilités prédites sont obtenues par la formule :

$$\hat{P}(\hat{y}_{ms,i} = 1|x_i) = \frac{\exp(\hat{\beta}_{ms,0} + \hat{\beta}_{ms,1}x_{i1} + \hat{\beta}_{ms,2}x_{i2} + \ \dots + \ \hat{\beta}_{ms,k}x_{ik})}{1 + \exp(\hat{\beta}_{ms,0} + \hat{\beta}_{ms,1}x_{i1} + \hat{\beta}_{ms,2}x_{i2} + \ \dots + \ \hat{\beta}_{ms,k}x_{ik})}$$

En utilisant les valeurs de $\hat{P}(\hat{y}_{ms,i} = 1|x_i)$, nous pouvons obtenir l'estimateur de projection :

$$\hat{Y}_{PR,ms,d} = \sum_{i \in A_1} w_{i1} \hat{P}(\hat{y}_{ms,i} = 1|x_i)\, \gamma_{di}$$

pour le total présent dans la population cible, et

$$\hat{R}_{PR,ms,d} = \frac{\sum_{i \in A_1} w_{i1} P(\hat{y}_{ms,i} = 1|x_i) \gamma_{di}}{\sum_{i \in A_1} w_{i1}\, \gamma_{di}}$$

pour la proportion dans la population cible.

Étape 4 : Estimations désagrégées et évaluation de leur exactitude

Les estimations, les erreurs types et les intervalles de confiance ont été calculés pour les dimensions de désagrégation pertinentes (par exemple, par sexe, tranche d'âge, quintile de revenu et milieu urbain/rural). Le tableau 3.1 ci-dessous présente les principaux résultats empiriques. La comparaison des estimations projetées par rapport aux estimations directes en termes de coefficient de variation (CV) et d'intervalles de confiance montre que la précision des estimations projetées est supérieure (ou au moins égale) à celle des estimations directes dans presque tous les cas.

Tableau 3.1 Estimations projetées vs estimations directes de la probabilité d'être en insécurité alimentaire modérée ou grave (prob.ms)

		Insécurité alimentaire modérée ou grave			
		prob.ms	**CV (%)**	**CI plus faible**	**CI plus élevé**
IHS4*	**Total**	0,91	1,2	0,89	0,93
GWP**		0,91	1,3	0,89	0,93
IHS4	**Femme**	0,91	1,4	0,88	0,93
GWP		0,90	1,5	0,89	0,94
IHS4	**Homme**	0,91	1,9	0,87	0,94
GWP		0,91	2,0	0,87	0,94
IHS4	**Rural**	0,93	1,2	0,90	0,95
GWP		0,92	1,3	0,90	0,94
IHS4	**Urbain**	0,81	5,7	073	0,92
GWP		0,82	5,9	0,74	0,93
IHS4	**15-24**	0,91	2,0	0,87	0,94
GWP		0,89	2,1	0,85	0,93
IHS4	**25-49**	0,91	1,6	0,88	0,93
GWP		0,92	1,6	0,89	0,95
IHS4	**50-64**	0,87	3,6	0,82	0,94
GWP		0,90	3,5	0,84	0,96
IHS4	**65+**	0,97	1,6	0,94	1,0
GWP		0,98	1,7	0,95	1,0
IHS4	**Inc_1**	0,96	1,5	0,94	0,99
GWP		0,97	1,5	0,94	1,0
IHS4	**Inc_2**	0,96	1,5	0,93	0,99
GWP		0,96	1,6	0,93	0,99
IHS4	**Inc_3**	0,97	1,1	0,95	0,99
GWP		0,97	1,1	0,95	0,99

IHS4	**Inc_4**	0,89	3,6	0,82	0,95
GWP		0,88	3,7	0,82	0,94
IHS4	**Inc_5**	0,74	3,8	0,68	0,80
GWP		0,76	3,8	0,71	0,82

* IHS4: Quatrième enquête Intégrée auprès des ménages du Malawi – 2016/17
** GWP: module d'enquête FIES du Malawi collecté via le Gallup World Poll (GWP) – 2016

4. CONCLUSIONS ET RECOMMANDATIONS

La nécessité de produire des estimations désagrégées relatives aux indicateurs pour le Cadre de suivi des ODD pose des défis importants aux SSN. Dans ce cadre, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO) – en tant que membre du groupe de travail sur la désagrégation des données – est bien placée pour accompagner les pays n'ayant pas la capacité de produire le niveau de désagrégation requis pour les indicateurs des ODD. Pour ce faire, le Bureau du statisticien en chef (OCS) de la FAO a élaboré des Lignes directrices sur la désagrégation des données relatives aux indicateurs des ODD en utilisant les données d'enquête (FAO, 2021). Ces lignes directrices proposent des orientations méthodologiques et pratiques pour la production d'estimations directes et indirectes des indicateurs des ODD en s'appuyant sur les enquêtes comme source principale ou préférée de données. En outre, ces lignes directrices fournissent des outils pour évaluer la précision de ces estimations et définissent des stratégies pour améliorer la qualité des résultats grâce à l'estimation indirecte, y compris les méthodes SAE.

Lors de la planification de la désagrégation des données à l'étape du plan d'échantillonnage, les Lignes directrices montrent comment le plan d'échantillonnage devrait assurer une taille d'échantillon planifiée pour tous les domaines de désagrégation. Les techniques d'échantillonnage traditionnelles utilisent le suréchantillonnage, la stratification plus poussée ou encore l'introduction de plans à plusieurs phases avec sélection des répondants afin de réaliser la désagrégation des données. Cependant, pour les petits domaines ou les populations isolées et hors d'atteinte, les techniques traditionnelles sont généralement irréalisables puisqu'elles tendent à augmenter exponentiellement les coûts des enquêtes. D'autres techniques plus complexes permettent d'améliorer les plans d'échantillonnage en répartissant géographiquement les unités d'échantillonnage et en réduisant le niveau de regroupement. Des approches plus récentes – telles que les techniques de stratification marginale, l'échantillonnage indirect, l'échantillonnage équilibré et à sources multiples – permettent de surmonter certaines de ces contraintes. Cependant, le problème majeur est que les bureaux nationaux n'ont pas encore adopté ces techniques et que leur adoption nécessiterait des programmes d'assistance technique et de développement des capacités. Les Lignes directrices (FAO, 2021 ; Chapitre 3) présentent en détail les forces et les faiblesses de toutes les méthodes proposées. En outre, elles sont assorties d'une annexe utile comprenant des ensembles de logiciels devant être utilisés dans des applications empiriques. Enfin, les Lignes directrices abordent les méthodes et les outils pour estimer l'exactitude des estimations directes désagrégées (chapitre 4).

Dans la phase d'analyse, les Lignes directrices utilisent une approche d'estimation indirecte assistée par modèle qui permet de générer des estimations désagrégées relatives aux indicateurs des ODD en s'appuyant sur l'utilisation intégrée de deux enquêtes indépendantes. Particulièrement, l'utilisation de l'estimateur de projection permet de combiner une enquête de grande taille ou un recensement, recueillant un ensemble d'informations auxiliaires, avec une enquête d'une taille plus petite, recueillant des données sur une variable cible avec le même ensemble de variables auxiliaires. L'approche d'estimation indirecte abordée couvre un grand nombre d'applications empiriques intéressantes et pertinentes pour la production de données désagrégées relatives aux indicateurs ODD (et autres). La plupart des pays peuvent généralement s'appuient sur des variables auxiliaires obtenues grâce aux enquêtes de grande taille, aux recensements, aux dossiers administratifs ou à des informations géospatiales. Dans ce contexte, certains des phénomènes cibles du suivi des ODD et la désagrégation des données sont souvent trop coûteux ou trop complexes pour être intégrés dans les campagnes de collecte de données à grande échelle. L'approche présentée permet de mesurer la variable cible en utilisant une enquête à petite échelle se basant sur l'échantillon à partir duquel il est possible de faire des estimations des paramètres d'un modèle statistique de type régression en liant cette variable à un ensemble de variables auxiliaires. Sur la base de ces paramètres, il est possible de faire des prédictions des valeurs de la variable cible dans une source de données à plus grande échelle en collectant les informations auxiliaires utilisées pour ajuster le modèle. Le fait de s'appuyer sur un échantillon plus large permet d'augmenter la précision des estimations

désagrégées et de prendre en compte des domaines de désagrégation qui ne sont pas disponibles au niveau de la petite enquête. Par ailleurs, la prédiction d'une variable cible sur l'échantillon d'une enquête plus étendue à partir de laquelle la plupart des statistiques nationales officielles sont produites permet d'améliorer la cohérence des estimations. En conclusion, il est important de souligner que la stratégie proposée pourrait être facilement étendue à d'autres contextes empiriques où, au lieu d'intégrer deux enquêtes indépendantes, une enquête de petite taille pourrait être intégrée à des informations auxiliaires provenant d'autres types de sources de données, telles que les recensements, les registres administratifs et/ou les données d'observation géospatiale.

5. RÉFÉRENCES

Birnbaum, Z.W. & Sirken, M.G. 1965. Design of sample surveys to estimate the prevalence of rare diseases: three unbiased estimates. *Vital Health Statistics*, 2(11): 1–8.
Breiman, L. 2001. Random forests. Machine Learning, 45(1): 5–32.
Cochran W. G. 1977. *Sampling Techniques*. Wiley. New-York.
Chauvet, G., Tillé, Y. 2006. A fast algorithm for balanced sampling. *Computational Statistics* 21, 53–62 (2006). https://doi.org/10.1007/s00180-006-0250-2.
Deville, J.-C. & Tillé, Y. 2005. Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128: 569–591.
FAO 2014. *The Global Strategy to Improve Agricultural and Rural Statistics. Technical Report on the Integrated Survey Framework*, Technical Report Series GO-02-2014, http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf. Accessed on 1 December 2014.
FAO. 2015. *Integrated Survey Framework*, Guidelines. Rome, FAO. (also available at http://www.gsars.org/wp-content/uploads/2015/05/ISF-Guidelines_12_05_2015-WEB.pdf).
FAO. 2016. *Guidelines for enumeration of nomadic and semi-nomadic livestock*. http://www.fao.org/3/ca6397en/ca6397en.pdf.
FAO. 2021. *Guidelines on data disaggregation for SDG Indicators using survey data*. http://www.fao.org/documents/card/en/c/cb3253en.
FAO, 2021b. *Using the projection estimator for data disaggregation of SDG indicators based on survey data*. Technical Report. To be published.
Falorsi, P.D. & Righi, P. 2015. Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys. *Survey Methodology*, 41(1): 215–236.
Falorsi P. D., Righi P., Lavallée P. 2019. Optimal Sampling for the Integrated Observation of Different Populations. *Survey methodology*, Vol. 45, No. 3, pp. 485–511. Statistics Canada, Catalogue No. 12-001-X.
Grafström, A., Lundström, N.L.P. & Schelin, L. 2012, Spatially Balanced Sampling through the Pivotal Method, *Biometrics*, 68: 514, 520.
Jessen, R, J, 1978. *Statistical Survey Techniques*. New York City, John Wiley & Sons.
Harrell J., **F.E**. 2015. Describing, Resampling, Validating, and Simplifying the Model. In: Harrell Jr., F.E., *Regression Modeling Strategies, Springer Series in Statistics*. Switzerland, Springer International Publishing, pp. 103–126.
Kalton, G. 2009. Methods for oversampling rare subpopulations in social surveys. *Survey methodology*, 35(2): 125–142.
Kursa, M. & Rudnicki, W., 2010. Feature Selection with the Boruta Package. Journal of Statistical Software. September 2010. Volume 36, Issue 11. https://www.jstatsoft.org/article/view/v036i11
Kim, J.K. & Rao, J.N.K. 2012. Combining data from two independent surveys: a model-assisted approach. Biometrika, 99(1): 85–100.
Lu, W. and Sitter, R. R. 2002. Multi-way Stratification by Linear Programming Made Practical, *Survey Methodology*, 2, 199–207.
Rao, J.N.K. 2003). Small Area Estimation. New York City, USA, John Wiley & Sons
Särndal, C.-E., Swensson, B. & Wretman, J. 1992 Model Assisted Survey Sampling. New York City, USA, Springer-Verlag.
Singh, A.C. & Mecatti, F. 2011. Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys. *Journal of Official Statistics*, 27(4): 633–650.
Tillé, Y. (2020). Sampling and estimation from finite populations. *John Wiley & Sons*.
Valliant, R., Dorfmann, A.H & Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York City, USA, John Wiley & Sons.

UNSD 2020. Report on the results of the UNSD survey on 2020 round population and housing censuses. Background document presented at the Fifty-first session of the United Nations Statistical Commission, 3–6 March 2020, New York City, USA (https://unstats.un.org/unsd/statcom/51stsession/documents/BG-Item3j-Survey-E.pdf).

Woodruff, R.S. 1971. A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association. 66(334): 411–414.

Verma, V. 2013. *Sampling for household-based surveys of child labour*. Geneva, Switzerland, International Labour Office. (also available at https://www.ilo.org/ipec/ChildlabourstatisticsSIMPOC/Manuals/WCMS_304559/lang--en/index.htm).