4

UN NOUVEL ECHANTILLON DE LA POPULATION DU RUANDA-URUNDI

PAR V. NEESEN, chercheur

1. Le territoire sous tutelle du Ruanda-Urundi s'étend sur une superficie de 54.700 kmz. Administrativement, il est divisé en deux "résidences", celle du Ruanda et celle de l'Urundi, comptant respectivement 8 et 9 territoires. Ceux-ci se composent de chefferies, à leur tour subdivisées en sous-chefferies, qui forment les unités administratives les plus petites.

La gestion des résidences et des territoires est confiée à des Administrateurs européens, celle des chefferies et sous-chefferies à des notables indigènes.

La population est une des plus denses de l'Afrique au Sud du Sahara, mais elle se caractérise en même temps par un éparpillement extrême : on n'y trouve point de villages, mais des habitations isolées qu'entourent des champs de bananeraies. C'est la colline qui constitue le lien géographique entre ces habitations, dite "ingo" (huttes ou petits groupes de huttes entourées d'un enclos) qu'on rencontre, suivant les particularité du terrain, sur les sommets ou dans les bas-fonds, mais le plus souvent sur les flancs des milliers de collines, qui sillonnent tout le Ruanda-Urundi.

Trois races vivent côte à côte dans ce pays: les Batutsi, d'origine hamitique, peuple de pasteurs, qui dominent les deux autres, c. à d. les Bahutu, agriculteurs et de loin les plus forts en nombre, et les Batwa, des pygmoïdes peu nombreux.

Dans le Rapport sur l'Administration Belge au Ruanda-Urundi pendant l'année 1924, la population totale du Ruanda était évaluée à près de deux millions, tandis que pour l'Urundi on dit qu'elle n'était pas inférieure à 3.000.000 d'habitants. En 1935 le même document estimait ce total à 3.385.883, en 1939 à 3.775.335 et en 1950 à 3.904.779.

Le calcul de ces estimations s'est toujours fait en partant du rapport entre le nombre des mâles adultes valides et la population totale dans une série de groupements recensés chaque année.

D'ailleurs, toutes les statistiques démographiques du Ruanda-Urundi ont été basées jusqu'ici sur le système bien connu des enquêtes démographiques. Il est vrai qu'on a introduit depuis quelques années un recensement complet fait annuellement par les sous-chefs ou des employés indigènes des territoires, mais les expériences nous enseignent que ces chiffres sont fort sujets à caution. Il en est de même pour l'état-civil embryonnaire qu'on instaure au fur et à mesure que les autorités indigènes évoluent, bien que ce soit là un effort très louable.

Bull. Mensuel de Statistique du Congo Belge et du Ruanda-Urundi – juillet 1952, n°21

De l'aperçu qui précède, il résulte que les meilleures statistiques démographiques fournies par l'Administration reposent sur les résultats des enquêtes démographiques.

Celles-ci, dans leur forme présente, ne donnent cependant pas de garanties suffisantes.

S'il importe de souligner que la méthode statistique adoptée, en l'occurence celle des sondages, est bonne en principe, il convient par contre d'en signaler les erreurs majeures commises dans son application, à savoir: l'erreur dans le choix des groupements, le nombre trop limité de groupements, l'erreur dans le calcul des estimations et la technique défectueuses qui préside aux opérations sur le terrain.

2. Le but des statistiques démographiques est de fournir des données numériques permettant :

a) de calculer la population totale du pays ainsi que sa répartition par circonscriptions administratives.

b) de connaître la répartition de la population par sexe, groupe d'âge et race.

c) de calculer certains taux donnant une idée de l'évolution actuelle de la population (taux de natalité, de fécondité, de mortalité) et ouvrant des perspectives démographiques.

En outre, pour le Ruanda-Urundi, il est intéressant de connaître le pourcentage de la population intéressée à l'élevage de bovidés.

3. Dans l'état actuel des choses, il n'y a qu'un recensement par échantillons qui est à même de fournir ces données avec une marge d'erreur assez réduite, permettant de les prendre comme base pour des décisions économiques et sociales

En effet, il est hors de question d'organiser un dénombrement complet. Où trouvera-t-on le personnel qualifié suffisant pour interroger tous les possesseurs de huttes, avec leur famille ? En supposant qu'on y arrive, quelles sommes d'argent et dépenses d'énergie n'exigerait pas une analyse approfondie de cette masse de résultats ?

D'autre part, la tenue des registres de décès et des naissances ainsi que l'enregistrement sur fiches sont à l'état embryonnaire et leur valeur, lorsqu'ils sont organisés, sujette à caution.

Il ne reste en fin de compte que la méthode d'échantillonnage, qui n'est rien d'autre qu'une application des dernières méthodes de la science statistique aux enquêtes démographiques.

L'échantillonnage permet incontestablement de concentrer le travail dans les échantillons choisis, de réduire ainsi le personnel nécessaire pour le "field-work" et les frais de déplacement, tout en augmentant la qualité de l'information. Il est d'ailleurs remarquable que même dans les pays très avancés comme les Etats Unis ou la Grande Bretagne, on recourt de plus en plus à cette méthode.

L'expérience statistique enseigne que les résultats de l'échantillonnage scientifique méritent souvent plus de confiance que les données d'un recensement complet exécuté dans les meilleures conditions; de plus, le recensement coûte beaucoup plus cher et la publication des résultats exige des délais très longs.

4. Les conditions du Nouvel Echantillon.

1. L'échantillon doit fournir avec une limite de confiance raisonnable des estimations de la population totale, de la mortalité, de la natalité, de la composition par sexe et par groupe d'âge pour tout le pays et par territoire.

2. Tenant compte des contingences budgétaires, le Gouvernement du Ruanda-urundi a organisé un réseau d'enquêteurs autochtones à raison d'un pour chaque territoire et de deux pour les 5 territoires les plus peuplés; en outre un élément de réserve est prévu par territoire, qui cependant devra être utilisé pendant la période des enquêtes. Le nouvel échantillon et plus spécialement le nombre d'habitants y inclus doit être déterminé en fonction du nombre des éléments disponibles pour faire le travail sur les collines.

3. L'échantillon doit prévoir un minimum de déplacements pour les enquêteurs, et doit permettre en même temps un maximum de contrôle de la part des autorités administratives.

4. Il serait en outre souhaitable que le nouvel échantillon puisse servir pour des enquêtes éventuelles dans d'autres domaines (agriculture, économie, santé, enseignement, etc...)

5. Les principes d'échantillonnage.

La façon la plus simple de choisir les échantillons aurait été de prendre un certain pourcentage de collines par chefferie. Les chefferies, en effet, englobent généralement des populations qui, dans les limites d'un territoire, se différencient par leur composition. par la race, par l'activité économique ou par d'autres facteurs, dont on peut supposer qu'ils exercent une influence sur les phénomènes démographiques. Un certain pourcentage de collines pris dans chaque chefferie peut donc constituer un échantillon représentatif de tout le pays.

Plusieurs considérations s'opposent à cette méthode : d'abord, on ne dispose pas pour le moment d'une liste exacte et complète des collines peuplées; ensuite, les difficultés de déplacement et de contrôle du travail pratique auraient été insurmontables si on devait atteindre environ 8% de toutes les collines dans chacune des chefferies du pays.

L'échantillonnage en deux temps ou en deux degrés (multi-stage sampling) peut remédier à ces difficultés. Dans ce cas, on prend d'abord un certain nombre d'unités du premier degré et dans ceux-ci on choisit des unités du second degré. Les groupements ainsi choisis constituent les échantillons définitifs.

b) Ni la population du Ruanda-Urundi ni celle d'un territoire ne forme un tout homogène,. Le rapport entre la population totale et les redevables de l'impôt de capitation est influencé par la mortalité masculine; la natalité, la mortalité, la composition de la population par groupe d'âge et par race diffèrent de région en région: le tout est hétérogène.

4

Néanmoins, il y a des couches ou strates relativement homogènes et le choix au hasard peut donner de meilleurs résultats si l'on constitue un échantillon stratifié. Ceci est possible en se basant sur une étude des caractéristiques démographiques de la population et des corrélations existant entre celles-ci et d'autres facteurs, de manière que les strates ne soient pas un découpage purement arbitraire. Un second principe est par conséquent le groupement des unités du premier degré en strates relativement homogènes.

Une fois les grandes lignes du plan déterminées, il nous reste à décider d'une série de questions également très importantes : quelles seront les unités du premier degré et du second degré ? Quels principes de stratification adoptera-t-on ? Comment fera-t-on le choix des unités du premier et du second degré ?

6. Les unités du premier degré.

Si toutes les unités dans chacune des strates étaient parfaitement égales du point de vue démographique, on n'aurait qu'à choisir par strate n'importe laquelle de ces unités pour obtenir des données tout à fait exactes, chaque unité représentant fidèlement toutes les autres. Ceci est évidemment un cas purement hypothétique, mais il permet de comprendre que l'erreur d'échantillonnage sera réduite dans la mesure où l'on a des unités hétérogènes dans ce sens-ci qu'elles comprennent une population où l'on trouve la plus grande partie des caractéristiques démographiques de la strate.

Le choix de la sous-chefferie comme unité du premier degré semble logique dans cet ordre d'idées. En effet, elle comporte toujours plusieurs collines, ce qui garantit une certaine variabilité du point de vue démographique. Si on combine pourtant deux ou trois sous-chefferies en unités élargies, on augmentera les chances que chaque unité élargie représente toute la strate, puisque dans ce cas la natalité et la mortalité, etc.... y seront plus hétérogènes et plus rapprochées de la situation générale de la strate. Par conséquent, on a groupé les sous-chefferies par deux ou trois, formant des unités d'environ 7. 500 habitants. Ce groupement est pourtant limité par les nécessités pratiques énumérées sous 5 (restreindre les déplacements - faciliter le contrôle administratif).

Les unités du second degré seront les collines, parties constitutives des unités du premier degré.

7. La stratification.

La phase suivante du plan est la classification des unités du premier degré en strates de façon

que la population de chaque strate soit aussi homogène que possible du point de vue démographique.

Il va de soi qu'il est impossible de déterminer des strates tout à fait homogènes mais il faut les faire aussi homogènes que possible en tenant compte des informations dont on dispose. Au Ruanda-Urundi, les deux races prépondérantes, les Batutsi, race dirigeante, s'adonnant à l'élevage, et les Bahutu vivant de l'agriculture, se distinguent par plusieurs facteurs, notamment leur milieu géographique, leur nutrition et leur genre de vie.

On a subdivisé la population en strates en partant de l'hypothèse que la différenciation des deux races avec tout ce qu'elle comporte crée aussi une différence dans la natalité, dans la mortalité et dans la composition par âges et sexes. Cette hypothèse a été confirmée par les pilot-tests dont on parle dans l'article ci-après.

On peut donc supposer raisonnablement qu'une région pastorale, une région agricole, une région mixte offrent une certaine homogénéité démographique et la stratification s'est faite en conséquence.

Pour des raisons administratives, on a divisé chaque territoire en strates, mais on comprendra aisément que certaines régions humaines s'étendent sur plusieurs territoires. Rien n'empêche de classer ces strates homologues en trois grands groupes : les régions pastorales, agricoles et mixtes. En outre, il y a plusieurs strates "sui generis", formées par des régions spéciales: une région sous l'influence d'un grand centre européen, une région industrielle, une région marécageuse particulièrement malsaine, etc.

Il est superflu de souligner que la stratification d'un pays dont on possède si peu de données numériques est nécessairement imparfaite. La division exacte en régions humaines ne sera possible qu'au moment où on disposera de statistiques démographiques sérieuses.

8. Le choix des échantillons.

Ce choix s'est fait par strates et au hasard à l'aide d'une table de "random numbers".

Ci-dessus on a expliqué qu'il est possible de grouper les strates en quatre catégories :

1. les régions agricoles
2. les régions pastorales
3. les régions mixtes
4. les régions spéciales.

Or, les régions agricoles sont de loin les plus importantes du point de vue nombre de leur population, et il est donc logique de les étudier plus intensivement. D'autre part, il s'avère important de posséder des estimations raisonnables pour les caractéristiques démographiques des autres régions, p. e. de celles où le Gouvernement envisage des immigrations. Un troisième facteur dont il fallait tenir compte dans la détermination de la fraction des unités du premier degré à choisir est que les enquêteurs , destinés à se mettre en campagne pour

la collection des données dans les échantillons, sont attachés à l'administration de chaque territoire et doivent donc travailler par territoire.

Les unités du premier degré représentant le mieux leur strate, en s'est efforcé d'en atteindre le plus possible, tout en les échantillonnant aussi intensivement que nécessaire pour obtenir une bonne estimation de leurs caractéristiques démographiques.

La situation géographique des territoires et notamment leur étendue permet généralement aux enquêteurs d'atteindre un quart des unités du premier degré.

Les "pilot-tests" de leur côté ont indiqué que la variabilité par strate des différents taux démographiques est de tel ordre qu'en prenant un échantillon d'un quart des unités du premier degré et d'environ un quart des unités du second degré (collines), les estimations de la population totale par territoire auront 95% de chances d'être exactes dans une limite de $\pm$ 5% (S = 2,5%).

Par conséquent, on a choisi au hasard 1/4 des unités du premier degré. Tenant compte des remarques ci-dessus et des résultats des pilot-tests, on a choisi au hasard 2/5 des collines composant les unités du premier degré dans les régions agricoles et 1/4 dans les autres régions (variable Sampling fractions).

9. Les enquêteurs.

Comme il était d'avance exclu de faire appel à des éléments européens pour le travail sur les collines (field-work), on a mis au point un réseau d'enquêteurs démographes autochtones. Monsieur A. d'ARIANOFF, Administrateur Territorial, a été chargé de leur formation (un article à ce sujet est publié dans ce bulletin).

10. Les Pilot-tests.

Une série de pilot-tests ont eu lieu, tant dans l'Urundi qu'au Ruanda, avant les enquêtes définitives qui se sont déroulées entre le 9 juin et le 15 août.

Le but de ces pilot-tests a été triple :

1. - compléter la formation pratique des enquêteurs et étudier le comportement des populations recensées lors des enquêtes.
2. - étudier la variabilité des différentes caractéristiques démographiques.
3. - constituer un moyen de contrôle pour les résultats qui seront fournis par les enquêtes définitives.

En effet, ces tests ont eu lieu en quatre régions typiques et ont porté sur un nombre d'habitants égal à 10% du nombre atteint pendant les enquêtes définitives.

11. Les estimations (1).

a) Population totale. La variabilité du nombre d'habitants par colline est très grande. Il en résulte qu'on devait comprendre dans l'échantillon un nombre extrêmement élevé de collines pour être à même de calculer avec des limites de confiance raisonnables une population moyenne par colline. Pour y remédier on a fait appel à des données supplémentaires et connues, c. à. d. le nombre des hommes adultes valides (HAV), constitué par ceux qui payent l'impôt de capitation et ceux qui sont déclarés exempts de cet impôt pour d'autres raisons que de santé. Il y a indiscutablement une corrélation très étroite entre le nombre des HAV et la population totale.

On calculera donc le rapport entre ces deux facteurs pour la population des échantillons : là en effet on les connaît tous les deux.

Ce coëficient multiplié par le nombre des HAV de la strate, donnera la population totale de celle-ci (Yi).

Formule : $Y_i = \frac{y_i}{x_i} X_i = \bar{r}_i X_i \qquad \left(\bar{r}_i = \frac{y_i}{x_i}\right)$

Où : y_i = la population totale des échantillons dans la strate i.

x_i = le nombre total des HAV des échantillons dans la strate i.

X_i = nombre total des HAV dans la strate i.

En principe, une estimation doit être :

1°- consistante, c. à. d. convergente en probabilité vers le paramètre réel de la population.

2°- non faussée (unbiased), c. à. d. que son espérance mathématique est égale au paramètre réel.

Les estimations basées sur des taux ou rapports sont généralement faussées, mais les résultats des pilot-tests montrent que la variance des différents taux par strate n'est pas considérable. Il semble par conséquent justifié de négliger cette erreur.

La population totale de tout le pays (Y) sera estimée selon la formule :

$$Y = \sum_{i=1}^{k} \bar{r}_i X_i = \sum_{i=1}^{k} Y_i$$

Où : k = nombre total des strates.

$\sum$ = sommation.

(1) - Je tiens à remercier Mr. J. B. DERKSEN, Chef de la Section du Revenu National de l'O. N. U., le Dr. S. H. KHAMIS du Bureau Statistique du même office ainsi que MM. L. A. BODART et H. LEDOUX, de la Section Statistique du Congo Belge pour leur collaboration. -

Dans l'échantillonnage à deux degrés, la variance (1) des estimations des paramètres de la population est composée de deux éléments ; la variance entre les unités du premier degré et celle entre les unités du second degré, formant les unités choisies du premier degré.

Voici une formule approximative pour le calcul de la variance de l'estimation de la population totale de la strate i (Yi) : $V'(Y_i) = X_i^2 V'(\bar{e}_i)$

ou $V'(Y_i) = X_i^2 \frac{1-f_i}{n_i} \bar{e}_i^2 \left[\frac{V(y_{ij})}{\bar{y}_i^2} - 2 \frac{cov(y_{ij}, x_{ij})}{\bar{x}_i \bar{y}_i} + \frac{V(x_{ij})}{\bar{x}_i^2} \right]$

avec $f_i = \frac{n_i}{N_i}$ ou la fraction d'échantillonnage des unités de premier degré.

En vue du calcul des $V(y_{ij})$, $cov(y_{ij}, x_{ij})$, $V(x_{ij})$, il y a lieu de tenir compte successivement des variances dues au premier et au second degré de sondage.

La part revenant à la variance entre les unités du premier degré dans la valeur de $V'(Y_i)$, peut s'exprimer comme suit :

$$\frac{N_i(N_i - n_i)}{n_i(n_i - 1)} \times \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2} \times \sum_{j=1}^{n_i} (y_{ij} - \bar{r}_i x_{ij})^2 \qquad (1)$$

Si l'on néglige, comme ce fut le cas dans de nombreuses enquêtes où les fractions d'échantillonnage pour le second degré étaient élevées, les effets de la variance entre les unités du second degré à l'intérieur des unités du premier degré, on obtient comme valeur approché de $V'(Y_i)$, celle de l'expression (1). -

Il importe cependant de mesurer les effets de la variance des unités du second degré sur la variance $V'(Y_i)$ - L'expression suivante :

$$\left(\frac{N_i}{n_i}\right)^2 \times \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2} \times \sum_{j=1}^{n_i} \sum_{k=1}^{l_{ij}} \frac{(y_{ijk} - \bar{y}_{ij})^2}{l_{ij} - 1} \cdot \frac{n_{ij} - l_{ij}}{n_{ij} - 1} \cdot l_{ij} \qquad (2)$$

peut être considérée comme satisfaisante.

où X_{ij} est la population totale des HAV dans l'unité du premier degré ij

x_{ij} est la population totale des HAV dans l'échantillon prélevé dans l'unité du premier degré ij

(1) Le calcul d'une estimation de la variance nous permet de déterminer la confiance que les estimations des paramètres de la population méritent. Si d'importantes décisions doivent être prises en partant de l'estimation du paramètre, la détermination de son degré de confiance est très importante. Jusqu'ici on ne connaît pas de formule exacte pour le calcul de la variance d'un taux (ratio-estimate), mais les formules approximatives qui suivent donnent satisfaction. -

y_{ij} est la population totale de l'échantillon prelevé dans l'unité du premier degré ij.

y_{ijk} est la population totale de la colline k dans l'unité du premier degré ij

$\bar{y}_{ijk}$ est la population moyenne d'une colline de l'échantillon prélevé dans l'unité du premier degré ij 9

n_{ij} est le nombre de collines dans l'unité du premier degré ij

l_{ij} est le nombre de collines dans l'échantillon prélevé dans l'unité du premier degré ij

La nécessité de chiffrer l'expression (2) est confirmée par une étude récente de Emil. M. JEBE lequel a démontré que la variance due au second degré est généralement importante. (Vide: Journal of the American Statistical Association, March, 1952 p. 49).

L'expression définitive de $V'(Y_i)$ est le résultat de la somme des expressions (1) et (2).

On obtient ainsi :

$$V'(Y_i) = \frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2} \times \left\{ \frac{N_i(N_i - n_i)}{n_i(n_i - 1)} \cdot \sum_{j=1}^{n_i} (y_{ij} - \bar{r}_i x_{ij})^2 + \left(\frac{N_i}{n_i}\right)^2 \cdot \sum_{j=1}^{n_i} \sum_{k=1}^{l_{ij}} \frac{(y_{ijk} - \bar{y}_{ijk})^2}{l_{ij} - 1} \cdot \frac{n_{ij} - l_{ij}}{n_{ij} - 1} \cdot l_{ij} \right\}$$

Une estimation de la variance de l'estimation de la population est donnée approximativement par :

$$V'(Y) = \sum_{i=1}^{k} V'(Y_i).$$

b) Taux démographiques : on calculera, en se basant sur les données des échantillons, par strate des estimations pour le taux de natalité (le rapport entre le nombre des naissances et la population totale), le taux de fécondité (rapport entre le nombre des naissances et le nombre des femmes fécondables), le taux de mortalité générale (nombre de décès par rapport à la population totale), le taux de mortalité infantile, etc... On déterminera des moyennes pondérées de ces taux pour tout le pays, en donnant aux taux de chaque strate un poids en rapport avec sa population totale.

Le calcul des variances des taux se fera selon des formules analogues à celles employées pour les estimations de la population totale.