

Chaîne de traitement classique des données de recensement

I. Le recensement de la population

1. Contexte et enjeux

La réalisation d'un recensement de la population et de l'habitat tous les dix ans est une recommandation du Système des Nations-Unies. Le dernier recensement de 2013 au Sénégal, après les expériences de 1976, de 1988 et de 2002, s'inscrit dans ce cadre. Cette opération, inscrite en activité phare de la stratégie de développement de la statistique au Sénégal, est indispensable à la gouvernance de la nation sénégalaise. Ainsi, à travers cette opération, inscrite dans les priorités du Gouvernement du Sénégal, il sera possible aux pouvoirs publics de mesurer le niveau de développement économique et politique du pays, d'apprécier les progrès accomplis, d'estimer les besoins et de planifier des programmes d'action ciblés, d'accompagner les décisions, etc.

2. Objectifs visés et résultats attendus

Le recensement général de la population et de l'habitat, de l'agriculture et de l'élevage (RGPHAE) a pour objectifs de faire le bilan démographique, de déterminer les caractéristiques de l'habitat, de faire le point sur le potentiel agricole et pastoral et de donner une situation de référence comme base d'appréciation de l'impact des politiques et des programmes socio-économiques mis en œuvre au cours de ces dernières années. Il sert également à donner une situation de référence comme base d'appréciation de l'impact des politiques et des programmes socio-économiques à mettre en œuvre pour les prochaines années.

Les principaux résultats attendus sont de:

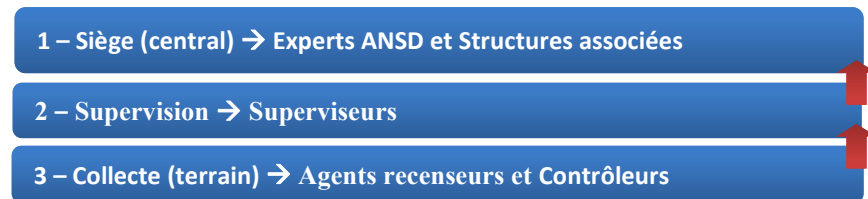
- Fournir l'effectif global de la population sénégalaise répartie par région, département, arrondissement, commune, commune d'arrondissement, communauté rurale, village et quartier, selon la situation de résidence ;
- Déterminer les caractéristiques démographiques, économiques et socioculturelles de la population ;
- Déterminer les différents niveaux de natalité, de fécondité, de stérilité et de mortalité ainsi que l'accroissement naturel et global de la population ;
- Etudier les conditions d'habitation des ménages ;
- Constituer une base de données socio-économique sur les collectivités en vue d'accompagner le processus de mise en œuvre, par le gouvernement, de la politique de bonne gouvernance locale et, plus particulièrement, de la politique de régionalisation ;
- Aider à l'établissement des projections démographiques, agricoles et de l'élevage sur une base scientifique et à la constitution de bases de sondage complètes et mises à jour, avec des cartes de collectivités administratives et locales pour la réalisation des futures enquêtes auprès des ménages, et sur l'agriculture et l'élevage ;
- Renforcer les capacités nationales dans la planification, l'exécution, les opérations de collecte, le traitement, l'analyse, la publication et la dissémination des résultats des recensements et enquêtes.

3. Dispositif organisationnel

Pour mener le recensement de la population, plusieurs phases sont nécessaires : cartographie des zones de dénombrement ou districts de recensement, développement des applications, réception et tests du matériel informatique, opérations de tests du dispositif organisationnel et technique, recrutement des agents recenseurs, contrôleurs et coordonnateurs techniques, gestion des confirmations, gestion du système de paiement, collecte proprement dite, enquête post-censitaire. Deux innovations majeures sont à noter pour le dernier recensement de 2013 : la combinaison de trois recensements en une seule opération et l'utilisation des TIC dans la collecte. En effet, la collecte a été réalisée avec les ordinateurs de poche (PDA, Personal Digital Assistant). Ce choix technologique se justifie pour plusieurs raisons : c'est une recommandation

des Nations-Unies, il accroît fortement la qualité des données car permettant de prendre en compte des mécanismes de contrôle et c'est une tendance internationale.

En termes d'organisation, trois niveaux de responsabilité ont été mis en place comme suit :



1 Organisation de la collecte

En termes de dispositif informatique, les ressources matérielles suivantes ont été utilisées:

- **800** ultraportables
- **47** ordinateurs fixes au niveau des 45 départements
- **300** clés Internet
- **900** téléphones portables
- **20 000** puces téléphoniques pour la communication avec les différents acteurs.

A cela, il faut ajouter un dispositif humain composé de :

- 17 164 agents recenseurs ;
- 2566 contrôleurs ;
- 633 superviseurs ;
- 252 assistants informaticiens;
- 34 coordonnateurs techniques régionaux ;
- 55 coordonnateurs techniques départementaux ;
- et un comité de veille.

II. Problématique

Le recensement de la population a consisté à interviewer l'ensemble des ménages au Sénégal (sénégalais comme étrangers) sur des questions préalablement choisies de manière consensuelle avec les Administrations publiques, les universités, le réseau des journalistes en population, les Organisations non gouvernementales, etc. Les données collectées à partir de ces interviews sont stockées dans une base de données et servent à la production de tableaux de bord. Ces derniers permettent d'évaluer l'impact des politiques socio-économiques en cours de mise en œuvre et par là-même de décider de politiques à établir dans le court, moyen ou long terme. Pour ce faire, les comparaisons entre les données de recensements différents sont nécessaires. Cependant, elles sont effectuées de manière manuelle. En effet, il est noté l'absence d'un entrepôt de données regroupant toutes les données issues des différents recensements en vue d'une aide à la prise de décision.

De plus, les données de recensement ne sont pas suffisamment valorisées du fait de leur structuration et de l'inexistence d'outils techniques adaptés permettant d'en exploiter la richesse, de les synthétiser et de les transformer en savoir.

Entre la collecte de terrain et la production des tableaux de bord, les expériences précédentes lors des recensements de 1976, 1988 et 2002, ont montré que la durée moyenne de disponibilité des données est de cinq (05) ans. Autrement dit, les données d'un recensement ne pouvaient être utilisées dans la prise de décisions qu'après un délai relativement long de cinq (05) ans. Cela constituait un biais supplémentaire sur la structure socio-économique du pays. En effet, le recensement se veut une photographie de la population à un instant donné, ce qui explique la durée très courte de quinze (15) jours pour finaliser la collecte au niveau national afin de minimiser les écarts entraînés par les faits d'état civil (naissance, décès) et les mouvements de population qui surviennent à tout moment. Utiliser les données après un délai relativement long, même en présence d'hypothèses sur l'évolution des facteurs (projections démographiques), peut poser un problème de fiabilité et de précision.

D'autre part, au cours des différentes phases du recensement (cartographie censitaire, recensement pilote, dénombrement général, enquête post censitaire, production du répertoire des localités entre autre), plusieurs bases de données sont produites. En guise d'illustration, nous faisons le focus sur la cartographie censitaire. Lors de cette étape, les agents de relevés de données cartographiques sillonnent le territoire national afin de localiser tous les ménages. Au terme de leur travail, l'ANSD dispose d'une base de données contenant les coordonnées des concessions, le nombre de ménages par concession, l'identité des chefs de ménage et les superficies occupées par les différentes concessions. Ces informations constituent les inputs nécessaires à l'organisation des équipes d'agents recenseurs sur le terrain. Chacun d'entre eux doit couvrir un district de recensement (zone de 1000 à 1200 ménages en milieu urbain, et de 800 à 1000 habitants en milieu rural). Le défi de taille est que les ménages soient bien localisés et que deux agents recenseurs travaillent dans des zones sans possibilité de chevauchement. Lors du dénombrement, les écarts constatés ont porté sur 20% des districts de recensement (DR), soit sur 4000 DR environ. Cela veut dire que l'intégrité et la cohérence nécessaire entre ces bases de données ne sont toujours pas garanties du fait de l'inexistence d'outils d'extraction et de chargement automatiques lors du passage d'une étape à l'autre. Lors des contrôles d'exhaustivité, des différences sont souvent notées dans les chiffres d'individus, de ménages ou de localités d'une base à l'autre à cause de l'inexistence d'un référentiel unique des données.

Par ailleurs, la production des tableaux de bord ou tabulation permettait juste de disposer des tableaux préalablement demandés par les différents organismes ayant contribué à la spécification du questionnaire. Les tableaux spécifiques nécessitaient parfois un retraitement des données.

La tabulation est une activité assez complexe qui nécessite l'intervention d'informaticiens spécialisés en traitement de données d'enquêtes, les statisticiens fonctionnels émettant juste leurs besoins. Pour plus d'efficacité, il serait intéressant que les fonctionnels puissent exécuter directement leurs requêtes sur la base de données et à partir d'interfaces conviviales, ne nécessitant pas la connaissance de SQL.

En synthèse, les principaux problèmes relevés sont les suivants :

- **Absence d'entrepôt de données ;**
- **Absence de souplesse dans la production des tableaux de bord ;**
- **Délai de production des indicateurs très longs (05 ans en moyenne) ;**
- **Problèmes d'intégrité et de cohérence entre les différentes bases de données du recensement ;**
- **Absence d'un système de comparaison automatique entre les données issues de différents recensements.**

Pour toutes ces raisons, il apparaît nécessaire de disposer d'un entrepôt de données permettant d'améliorer la prise de décisions.

III. Objectifs

Face aux problèmes énoncés précédemment, notre travail sera de mettre en place un système d'information décisionnelle basé sur les données du dernier **RGP HAE**.

Ce système permettra entre autres:

A l'Etat

- de disposer des meilleures informations (indicateurs) afin de mettre en place des politiques de développement socio-économique plus adaptées (accès aux infrastructures sociales de base (écoles, maternités, etc.), meilleure connaissance du tissu productif (activités professionnelles, organisation de la formation professionnelle, etc.)) ;
- de mieux répartir les subventions aux communes et communes d'arrondissement selon leurs populations ;

Aux fonctionnels

- de raccourcir les délais de mise à disposition des tableaux de bord ;
- de permettre une comparaison automatique entre les données communes issues des différents recensements de la population ;
- de mieux valoriser les données de recensements et d'en tirer des connaissances ;

- de simplifier la mise à disposition des tableaux de bord et de faciliter l'accès et l'extraction de données aux fonctionnels ;

Aux informaticiens

- d'assurer l'intégrité et la cohérence des bases de données issues des différentes phases du recensement (cartographie censitaire, recensement pilote, dénombrement général, enquête post censitaire, production du répertoire des localités) ;
- de disposer d'un entrepôt de données facile à maintenir ;
- d'assurer la sécurité pour la restitution des données en mettant en place des rôles pour l'accès aux des rapports ;
- de restituer les rapports sous divers formats (Excel, PDF, WEB).

IV. Etude de l'existant

De 1976 à 2002, le traitement des données de recensement se faisait à travers les étapes de collecte sur le terrain, saisie et ressaisie des questionnaires, apurement des données (recodification et regroupement de variables, contrôle des données, traitement des valeurs extrêmes et des données manquantes, recherche de doublons, contrôle de vraisemblance), tabulation et production de rapports. La saisie se justifie par le fait que les questionnaires au format papier étaient utilisés pour la collecte, la ressaisie étant un moyen de contrôle. En effet, les agents de saisie induisaient des erreurs différentes de celles liées à la collecte. Le recensement de 2013 a suivi les mêmes étapes à l'exception de celles liées à la saisie car les PDA ont permis de faire la collecte directement sur un format numérique.

Deux logiciels étaient utilisés dans la chaîne de traitement classique des données de recensement : SPSS (logiciel d'analyse statistique aujourd'hui propriété d'IBM) et CsPRO (logiciel développé par le Bureau du Recensement des Etats-Unis). CsPRO est utilisé dans la confection des programmes de saisie et d'apurement, tandis que SPSS sert à l'analyse statistique.

L'innovation technologique de 2013 avec le choix des PDA comme outil de collecte a permis d'allonger la liste des solutions logicielles. En effet, en plus de CsPRO et SPSS, nous avons noté l'existence de programmes de chargement développés avec le langage C# et de programmes d'exportation de la base de données vers CsPRO. Les programmes de chargement ont permis le stockage des données collectées au niveau de la base de données sous SQL Server 2012.

Afin de retrouver la chaîne classique de traitement des données, la base de données sous SQL Server est exportée au format CsPro au moyen des programmes d'exportation. A partir de cette étape, les corrections et les imputations sont directement appliquées sur la base de données. Cette dernière est ensuite exportée au format SPSS pour la production des syntaxes permettant de disposer des tableaux statistiques ou tableaux de bord.

Toutes ces étapes (SQL Server vers CsPro, puis CsPro vers SPSS, tabulation) sont réalisées par les informaticiens en charge du traitement de données d'enquêtes. La liste des tableaux à générer est reçue des statisticiens.

Sur la durée, la production des tableaux génériques nécessitait quatre à cinq mois. Les tableaux permettant de répondre à des demandes spécifiques étaient produits après la publication des rapports définitifs, donc au-delà des 5 ans, à la fin de l'exploitation.

2. Solutions techniques

L'ANSD est née suite à la mutation institutionnelle de l'ex-Direction de la Prévision et des Statistiques (DPS) en 2006. En 2004, l'ex-DPS avait initié la mise en place d'une solution informatique nommée « REQUETEUR » pour effectuer des comparaisons entre des variables issues d'enquêtes différentes. Le constat de l'époque était qu'il y avait beaucoup de données, mais aucun moyen d'étendre facilement les capacités d'analyse en comparant les données provenant des différentes enquêtes.

La mise en place du « REQUETEUR » avait nécessité la centralisation des données d'enquêtes et l'harmonisation des variables à des fins de comparabilité. Les principales difficultés résidaient sur le fait que ces données étaient difficiles à exploiter car elles étaient sur support papier et pas totalement harmonisées.

Les points forts du « REQUETEUR » sont :

- les variables issues des différentes enquêtes sont liées entre elles par les trois niveaux de thèmes hiérarchisés (supra thèmes, thèmes et sous thèmes). Cette organisation permet de réaliser de manière simple des agrégations par niveaux ;
- la notion de type de variable (qualitative ou quantitative) est prise en compte à travers la nomenclature. Cela permet de mieux gérer les comparaisons pour obtenir des résultats réalistes ;
- l’affichage des valeurs prises par ces variables et le comptage du nombre d’occurrences des unités statistiques observées selon les valeurs prises par ces variables ;
- l’accès aux métadonnées.

Nous notons aussi plusieurs applications web :

- l’**ANADS** (Archive Nationale des Données du Sénégal) permet d’accéder plus facilement aux métadonnées des enquêtes statistiques réalisées au Sénégal. En effet, l’ANADS représente la mémoire des enquêtes : toutes les données et ressources (documents de projet, note méthodologique, rapports techniques, manuels, questionnaires, documents de budget, etc.) sont sauvegardées et accessibles en ligne pour permettre aux usagers de disposer d’un maximum d’informations sur le déroulement des enquêtes, aux statisticiens d’améliorer les prochaines opérations de collecte en tirant profit des difficultés répertoriées ;
- le **Portail de données du Sénégal** est un outil avancé de présentation de données d’enquêtes. Les résultats des requêtes des utilisateurs sont présentés sous formats tabulaire, cartographique et graphique ;
- le **SIG-STAT** est un système d’information géographique sur les données statistiques. Il permet aux usagers de disposer d’informations sur la localisation des principales infrastructures sociales de base (aéroports, structures de santé, fontaines, routes, etc.).

V. Critiques de l'existant (domaines d'amélioration)

Ces outils présentés ci-dessus sont pratiquement les seuls éléments de manipulation et de restitution de données statistiques traitées au sein de l'agence.

Cependant, nous notons quelques limites sur ces outils :

- **Le Requeteur**
 - Il ne donne pas de statistiques sur les données
 - Problème d'harmonisation et de duplication des données
 - Pas de politique de gestion des rôles (qui peut voir et qui ne peut pas voir)
- **L'ANADS, le Portail de données du Sénégal et le SIG-STAT**

Tous ces outils présentés ci-dessus ont une grande importance dans l'exploitation des données statistiques mais ils n'interagissent pas encore avec les données du RGPHAE d'où la nécessité de mettre en place un système d'information. Ce système permettra en premier d'avoir un entrepôt de données fiable puis en deuxième lieu de tirer le plus d'information sur les données du recensement pour ensuite constituer un outil d'aide à la décision.

Présentation de la Structure d'accueil

I. Missions

L'ANSD est une structure administrative dotée de la personnalité juridique et d'une autonomie de gestion et placée sous l'autorité du ministre chargé de la statistique. En particulier, l'Agence est chargée:

- de coordonner les activités statistiques du système statistique national ;
- de produire des données de qualité.

II. Organigramme

L'ANSD est une structure autonome placée sous la tutelle du ministre chargé de la statistique et administrée par un Conseil de surveillance. Elle se compose de services centraux et de

subdivisions au niveau régional. L'ensemble est organisé en Cellules, Directions, Divisions, Bureaux et Services.

La Direction Générale

Outre les Conseillers et chargés de mission, elle comprend la Cellule de passation des marchés (CPM) et la structure d'audit interne.

Les Directions

L'ANSD comprend huit (8) unités principales :

- la Direction des Statistiques Economiques et de la Comptabilité Nationale (**DSECN**) ;
- la Direction des Statistiques Démographiques et Sociales (**DSDS**) ;
- la Direction du Management de l'Information Statistique (**DMIS**) ;
- la Direction de l'Administration Générale et des Ressources Humaines (**DAGRH**) ;
- l'Agence Comptable Particulière (**ACP**) ;
- l'Ecole Nationale de Statistique et de l'Analyse Economique (**ENSAE**) ;
- la Cellule de Programmation, d'harmonisation de Coordination statistique et de Coopération Internationale (**CPCCI**) ;
- les Services Régionaux de la Statistique et de la Démographie (**SRSD**).

La **DSECN** est chargée d'établir des statistiques économiques globales et les comptes nationaux.

La **DSDS** est chargée de la conception, de l'exécution et de l'analyse des enquêtes et recensements démographiques et socio-économiques auprès des ménages.

La **DMIS** est chargée d'assurer la mise à disposition d'un système d'information efficient pour l'ensemble des activités de l'ANSD et de gérer la diffusion des produits de l'ANSD. Elle est structurée en deux divisions:

- la Division de la Documentation, de la Diffusion et des Relations avec les Usagers ;
- la Division des Systèmes d'Informations et des Bases de Données.

La **DAGRH** est chargée de la gestion du personnel et des compétences de l'agence, d'assurer la sécurité de l'agence sur toutes les questions juridiques et réglementaires, d'assurer l'approvisionnement de l'agence et la gestion de la logistique et du matériel.

L'**ACP** assure la gestion des moyens financiers. Elle tient la comptabilité des deniers de l'agence selon le système de comptabilité en vigueur, veille à la disponibilité d'une bonne information financière fiable, à temps et régulière.

La **CPCCI** a pour mission d'appuyer le Directeur Général dans la préparation des dossiers relatifs au Conseil National de la Statistique (CNS) et au Comité Technique des Programmes Statistiques (CTPS).

Les Services Régionaux de la Statistique et de la Démographie (**SRSD**) sont chargés de la coordination des activités statistiques des différentes régions.

Conclusion

Dans cette partie, nous avons présenté le sujet de notre mémoire. Nous avons montré la problématique liée au système existant. Pour remédier à ces problèmes, nous avons jugé nécessaire la mise en place d'un entrepôt de données pour améliorer le système d'aide à la prise de décision.

Dans la partie suivante nous vous présenterons les principales notions de bases à connaître pour pouvoir mettre en place un système d'information décisionnel.