Atelier sur la confidentialité des données et l'anonymisation des microdonnées

Cefil, Libourne, du 4 au 12 juillet 2016

Introduction

Du 4 au 12 juillet 2016, l'Insee et AFRISTAT ont organisé au Cefil, l'atelier sur la confidentialité des données et l'anonymisation des microdonnées à l'attention des participants en provenance des Etats membres d'AFRISTAT (Cameroun, Cap vert, Congo, Côte d'Ivoire, Djibouti, Mali, Sénégal et Togo), de la Tunisie, du Maroc, d'Haïti, des écoles de statistiques d'Abidjan, Dakar et Yaoundé, et de la Commission de l'UEMOA (Voir Liste des participants en annexe).

Cet atelier a connu la participation d'une vingt deux (22) participants, statisticiens et informaticiens. Les travaux ont été animés par les experts de l'Insee, d'AFRISTAT et de Statistique Canada.

Les travaux ont été organisés en 13 sessions théoriques et pratiques. Des exercices ont été réalisés sur micro-ordinateurs par l'ensemble des participants. (Voir Calendrier de l'atelier en annexe).

Contenu des sessions

Traitement de l'information, principes de sécurité et de confidentialité statistique

L'introduction du séminaire a été faite par M. Stéphane GREGOIR, Directeur de l'Ecole d'économie de Toulouse. Elle a porté sur le traitement de l'information, principes de sécurité et de confidentialité statistique. Le présentateur a passé en revue successivement l'environnement de la production des données, les sources de génération des données et les types de données.

Il a montré l'importance des données à caractère individuel pour la mise en œuvre des politiques et des travaux de recherche, des opportunités qu'elles présentent pour la statistique publique en matière de production et de gain en précision. Cependant, il a signalé que toutes les données collectées ne sont pas toutes exploitables.

Il a présenté différentes définitions du caractère des données personnelles dans divers cadres légaux (France, Union européenne, Déclaration universelle des Droits de l'Homme), de l'anonymisation, de la pseudo-anonymisation et de l'identifiabilité.

En matière d'obligations légales et morales, Il a été recommandé aux statisticiens la préservation des conditions de garantie de confidentialité et le respect de la vie privée.

Après la définition et la problématique des données à caractère personnel, une brève revue des méthodes d'anonymisation et des outils a été faite. Parmi celles-ci, la méthode d'échantillonnage et de regroupement semble la mieux indiquée.

La confidentialité des données personnelles est un enjeu pour les INS pour la qualité, la pertinence et la confiance dans les statistiques publiques. Le développement d'un savoir-faire et la maîtrise des techniques constituent une nécessité.

Le contexte réglementaire français et européen en matière de secret statistique

Le contexte réglementaire français et européen en matière de secret statistique a été présenté par M. Michel ISNARD, de l'Unité affaires juridiques et contentieuses de l'Insee.

Sa présentation a essentiellement porté sur la diffusion des données à l'INSEE par rapport au secret statistique en se basant notamment sur l'article 6 de la loi du 07 juin 1951 qui porte sur les délais de

diffusion des renseignements individuels. Ces délais sont de 75 ans pour les données ménages et 25 ans pour les données entreprises.

La deuxième partie de la présentation a concerné les types de données diffusées par l'INSEE. Ce sont des données : (i) sous forme tabulaires ; (ii) des données individuelles ne permettant pas l'identification directe ou indirecte (données individuelles « anonymisées »); (iii) des données individuelles permettant l'identification directe ou indirecte (données individuelles confidentielles). La communication desdites données s'accompagne d'un transfert de la responsabilité pénale des statisticiens aux utilisateurs.

La démarche qualité de l'Insee

La démarche qualité a été présentée par M. Karim Zaari, de l'Unité Qualité de l'Insee comme un élément essentiel pour soutenir les processus mis en œuvre dans la confidentialité des données et l'anonymisation des microdonnées.

La démarche qualité vise la satisfaction des besoins des utilisateurs en prenant en compte la matrice des risques.

L'intervenant a insisté sur l'engagement fort qui doit venir de la haute autorité de l'institution pour que la démarche qualité soit une réussite. Cet engagement qui pourrait émaner du comité de direction, doit être formalisé dans un document à rédiger et à diffuser par le Directeur général.

Par ailleurs, Il a montré que le GSBPM (Generic Statistical Busisness Process Model) permet de décrire les processus statistiques de façon cohérente. Cet outil prend en compte le principe de la confidentialité des données statistiques. Le présentateur a préconisé la mise en place de mécanismes d'évaluation pour une amélioration continue, prenant en compte les différents types d'évaluation, notamment l'évaluation par les pairs.

Gestion de la confidentialité dans les tableaux de données

La « Gestion de la confidentialité dans les tableaux de données agrégées - le secret statistique : pourquoi ? Comment ? » a été présentée par M. Maxime Bergeat de la division Recueil et du traitement de l'information de l'Insee.

L'enjeu de la protection des données individuelles est de conserver la confiance des répondants et de garantir le taux de réponse élevé dans un cadre légal à prendre en compte tout en cherchant à diffuser l'information la plus complète possible. Il s'agit d'empêcher les reconstructions des données individuelles par des personnes qui possèderaient des informations auxiliaires.

Utilisations possibles

Les diverses utilisations possibles des informations constituent les raisons suivantes qui militent pour l'ouverture des données :

- accélérer le rythme et la fréquence de prises de décisions économiques ;
- mesures en temps réel ;
- développer des techniques de prédiction et d'imputation ;
- mesurer de petits domaines et faire des imputations ;
- utiliser le Big data pour répondre à certaines questions de société ;
- améliorer les techniques d'évaluation (ex post et ex ante) ;

- participer à l'économie de l'information ;
- réaliser des partenariats avec le privé pour la production statistique.

Protection des données à caractère personnel

Les données personnelles sont protégées par la loi. Toutefois, des dérogations existent pour les chercheurs. Des règlements internationaux préconisent que les données individuelles ne soient utilisées que pour la statistique.

Les données peuvent être diffusées sous plusieurs formes : fichiers « grand public », Fichiers de production et de recherche, fichiers accessibles uniquement à travers un centre d'accès sécurisé.

La gratuité fait perdre à l'INS une partie de ses ressources. Mais, cette perte devrait être compensée par l'activité de conseils et les travaux à façon sur les données.

Les principales méthodes de protection des données à caractère personnel présentées sont :

- agrégation dans des classes de valeurs ;
- recodification des valeurs extrêmes ;
- suppression des observations appartenant à une cellule très peu peuplée ;
- utilisation des arrondis ;
- ajout de bruits ;
- échantillonnage ;
- échange ou permutation ;

Des méthodes plus sophistiquées (K-anonymat, L-diversité, Post response randomization, confidentialité différente) ont également été présentées.

Les outils élaborés et utilisés à l'heure actuelle par les pays et les instituts nationaux de statistique ont été présentés. Grâce à un consensus trouvé au niveau européen, deux logiciels sont en cours d'utilisation : Tau-Argus dévéloppé par Central Bureau of Statistics (CBS) des Pays-bas et Sdc-Micro Statistics Austria. L'Insee a choisi Tau-Argus.

En ce qui concerne les microdonnées, on distingue trois types de données : (i) les données issues d'enquêtes simples, (ii) les données d'enquêtes enrichies par des données fiscales, (iii) les données administratives.

Les méthodes appliquées ont été présentées : (i) enlèvement des clés d'identification (identifiants directs) ; (ii) évaluation du risque (k-anonymat et l-diversité) ; (iii) utilisation des méthodes perturbatrices, génération de données synthétiques, sous-échantillonnage, puis calage, agrégation d'information, suppressions locales, etc.

Les logiciels suivants ont été indiqués aux participants : µ-Argus dévéloppé par CBS (Pays-bas), Sdc-Micro développé par Statistics Austria, et d'autres outils spécifiques (Arx, Package R simPop).

Anonymisation des données : cadres et aperçu à Statistique Canada

Mme Michelle Simard de Statistique Canada a présenté l'expérience en matière d'anonymisation des données. Le Canada utilisant des méthodes similaires à celles de l'Insee, la présentatrice a insisté sur les spécificités du système canadien. Ainsi, elle a introduit les notions de personnes sûres, de systèmes sûrs, d'emplacements sûrs et de données sûres. Le Canada a démarré la mise en place des centres de données dans les universités depuis 15 ans. Une autre spécificité de Statistique Canada est la gestion du risque et l'établissement de la carte Risque/Utilité (Ducan, 2001).

Enfin, elle a insisté sur l'importance de la composante « informatique » dans les activités d'anonymisation et de confidentialité. A cet effet, elle encourage le suivi des développements des outils informatiques et l'adoption des outils développés par les instituts avancés. Les INS doivent améliorer leurs services de la diffusion afin de valoriser les données. Pour cela, il faut mieux communiquer et sensibiliser les utilisateurs sur les données disponibles. Il s'agit de créer la demande en informant les utilisateurs sur l'existence des données.

Centre d'accès sécurisé aux données (CASD)

Mme Roxane Silberman a présenté les activités du Centre d'accès sécurisé aux données (CASD) créé en 2010 qui a pour vocation de gérer les bases de données de l'Insee à destination des chercheurs sur la base d'une convention qui lie les deux institutions. Les fichiers sont classés en trois types : fichiers grands publics, fichiers scientifiques et fichiers très détaillés. Le CASD est partenaire du réseau Quetelet créé depuis 2000. Ce réseau a la responsabilité de (i) récupérer les données, (ii) archiver sur le long terme, (iii) documenter les données pour faciliter l'utilisation avec les outils NESSTAR, (iv) mettre à la disposition des chercheurs et (v) contribuer à l'amélioration des enquêtes et à leur valorisation. Le réseau a mis pour cela en place des data center dans les universités et gère aussi une base de données de questions et de variables. Les formats d'accès aux données sont divers (SPSS, SAS et STATA).

Travaux pratiques

Deux sessions ont été consacrées aux travaux pratiques dirigés par MM. Maxime Bergeat et Julien Lemasson. L'objectif était de donner aux participants une formation de base pour la gestion de la confidentialité dans les tableaux de données agrégées en utilisant le logiciel Tau-Argus.

Sur la base d'exercices, les participants ont mis en œuvre les méthodes présentées pour les données tabulées avec le logiciel Tau-Argus. Il s'agit de comment traiter le secret d'un tableau de données agrégées en suivant les étapes suivantes :

- La mise en forme des données ;
- L'importation des données dans Tau-Argus et la préparation du fichier des métadonnées ;
- Le traitement du secret primaire ;
- Le Traitement du secret secondaire ;
- L'enregistrement des tableaux anonymisés.

Par contre, les participants n'ont pas eu à réaliser des exercices d'anonymisation sur des microdonnées.

Présentations d'AFRISTAT, des écoles et des pays en matière de diffusion de données.

AFRISTAT a présenté ses activités en matière de documentation, d'archivage et de diffusion des données ainsi que sur l'élaboration de la politique de diffusion des données. Après assisté les pays à disposer d'archives nationaux de données d'enquêtes et recensements, l'Observatoire a mis en place un portail régional, dénommé « Nada régional». A terme, ce portail devrait pointer sur les Nada nationaux.

Le dépouillement des questionnaires adressés aux INS et remplis avant la tenue de l'atelier a montré que :

- la plupart des pays dispose de systèmes de documentation et d'archivage des données d'enquêtes et recensements statistiques (Nada et / ou IMIS-Redatam) ;
- les lois statistiques nationales interdissent la diffusion des microdonnées non anonymisées. Toutefois, la plupart des pays diffusent des données ;
- dans la majorité des cas, les données transmises par les pays ne sont pas anonymisées ;
- des actions de sensibilisation des différents acteurs doivent être entreprises par rapport à l'anonymisation des données (tabulées, microdonnées).

Plusieurs intervenants ont présenté les expériences menées par leur institution :

- la « Protection des données à caractère personnel et dispositions légales en Côte d'Ivoire » ;
- l'anonymisation des données d'une enquête réalisée par téléphone au Togo ;
- l'évaluation des Nada au niveau régional et la mise en place des bases de données au Cameroun ;
- l'organisation mise en place pour la diffusion des microdonnées à l'ANSD du Sénégal et la certification de intervenants dans les opérations de collecte ;
- l'organisation de la statistique à Haïti et son expérience en matière de diffusion ;
- l'expérience du Mali en matière de documentation et archivage des microdonnées.

Principales recommandations

A l'issue des travaux de l'atelier sur la confidentialité des données et l'anonymisation des micro-données, les participants ont formulé les recommandations suivantes :

Aux INS,

1. réaménager les lois statistiques nationales pour les mettre en accord avec les principes fondamentaux de la statistique officielle et la Charte africaine de la statistique afin de prendre en compte la diffusion des micro-données et la confidentialité des données publiées (tableaux, indicateurs, …) ;
2. réfléchir sur les méthodes d'anonymisation des données et mettre en place des systèmes de gestion de micro-données pour faciliter les relations avec les utilisateurs;
3. préserver les conditions qui garantissent la confidentialité et le respect de la vie privée des enquêtés. La communication desdites données s'accompagne d'un transfert de la responsabilité légale des statisticiens aux utilisateurs par la signature de protocole d'accord.
4. Promotion des bases données existantes….

A AFRISTAT et à l'Insee :

5. accompagner les INS dans la formation des cadres du SSN sur les techniques d'anonymisation des données et la révision des lois statistiques pour les pays qui ne l'ont pas encore fait ;
6. participer aux discussions internationales relatives au développement d'outils adaptés et adopter les outils produits par les institutions plus avancées sur la question ;
7. Rendre disponible des tutoriels et manuels de formation en français

Aux écoles,

8. conclure avec les INS et les universités des accords pour la mise en place de centres de données de recherches (CDR) en leur sein afin d'appuyer les chercheurs, universitaires et élèves en formation statistique et sciences sociales ;
9. renforcer les programmes d'enseignements dans les écoles de statistiques des cours sur l'éthique et la déontologie statistique ainsi que des notions sur l'organisation statistique mondiale, les principes fondamentaux de la statistique officielle, les lois statistiques nationales, la charte africaine de la statistique, ...

Conclusion

D'autres sujets intéressant les statisticiens africains ont été identifiés et discutés au cours de cet atelier. Ils pourraient faire l'objet des futurs ateliers :

- les enquêtes en ligne ;
- la mise en place d'un réseau d'enquêteurs « professionnels » ;
- la formation à distance en statistique.

La formation reçue ainsi que les enseignements tirés de cet atelier constituent une base solide pour bâtir des systèmes efficients d'anonymisation et de diffusion des données. Toutefois, elle ne suffit pas à garantir l'acquisition totale du savoir-faire pour l'utilisation des logiciels, notamment en ce qui concerne les microdonnées.

Les participants ont exprimé un grand intérêt pour les techniques présentées et reconnaissent l'importance de leur mise en œuvre au sein de nos instituts pour améliorer la diffusion des données dans le respect des règles déontologiques. Ils s'engagent à poursuivre la veille technologique sur la question et souhaitent parfaire leurs connaissances au cours d'autres formations sur la question dans le futur.