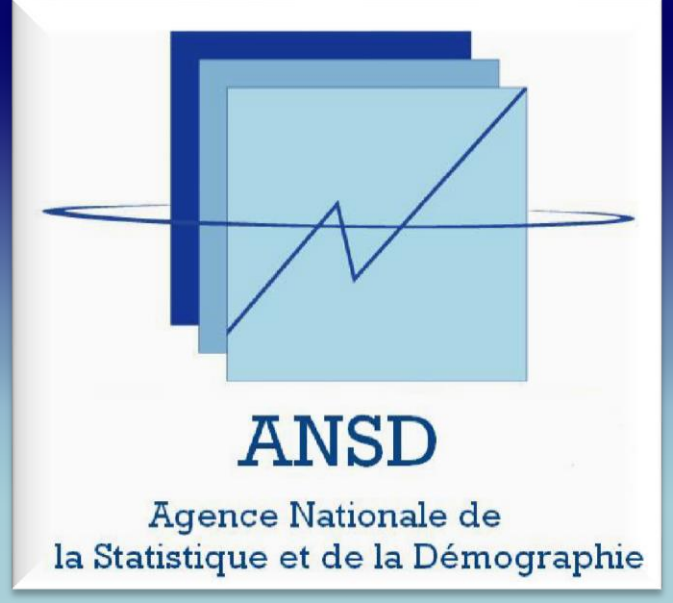


Initiation à l'anonymisation des microdonnées d'enquêtes

Abdoulaye BA, ANSD
El Hadji Malick GUEYE, ANSD
Luc Decker, IRD
Pascal Aventurier, IRD

Gestion des microdonnées et métadonnées statistiques

- Cadre de référence
- Anonymisation
- Cas pratique

ANSD
Agence Nationale de la Statistique et de la Démographie

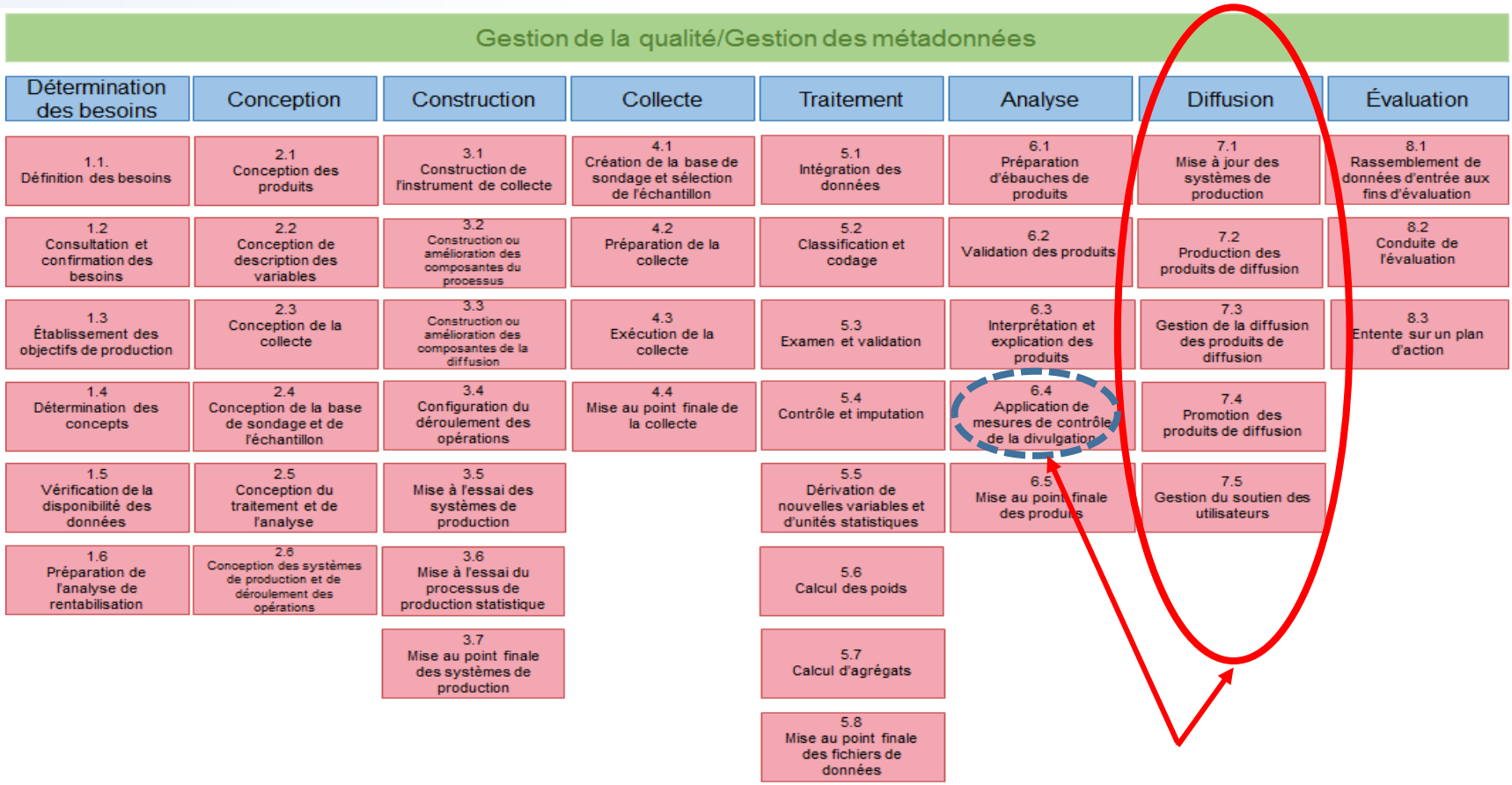
REPUBLIQUE DU SENEGAL

Cadre de référence

Cadre méthodologique

Modèle Générique du Processus de Production Statistique

Generic Statistical Business Process Model (GSBPM)



Outil flexible, conçu pour s'appliquer à toutes les activités réalisées par des producteurs de statistiques, tant à l'échelle nationale qu'internationale. Le GSBPM en est aujourd'hui à sa 5ème version. Largement utilisé par les organismes de statistique nationaux et internationaux, il permet une harmonisation et une comparaison des processus suivant les différentes phases de la chaine de production statistique.

Diffusion des microdonnées – Problématique de l'anonymisation

Diffuser ⟷ Préserver confidentialité

Mettre à la disposition des usagers des statistiques de qualité et préserver la confidentialité des données à caractère personnel.

Comment y arriver???

Initiation à l'anonymisation

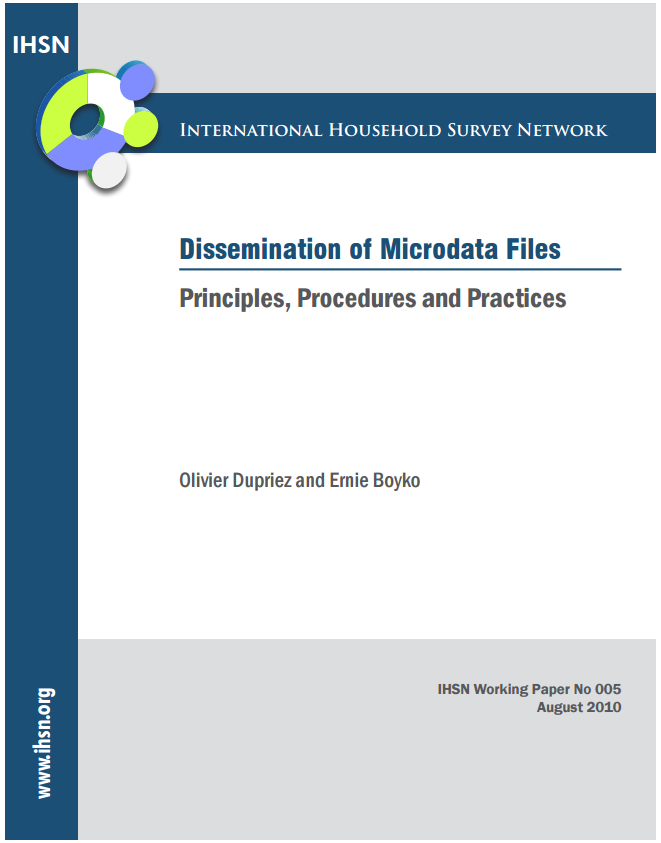
1. Modalités de diffusion des métadonnées
2. Principaux avantages de la diffusion des microdonnées
3. Coûts et risques associés
4. Enabling environment: legal, ethical, and technical considerations
5. Concepts clés : identifiants, quasi-identifiants, variables sensibles, risque de divulgation, utilité
6. Mesure du risque de divulgation
7. Réduction du risque de divulgation
8. Equilibre entre réduction du risque et perte d'information
9. Outil SDC, manuels et bonne pratique
10. Importance des métadonnées

Modalités de la diffusion des microdonnées

- Enclave de données / centre de données de recherche/laboratoire → cher, peu pratique
- Accès à distance
    - Soumission de scripts → lent, rigide, peu pratique pour les utilisateurs, peut être cher à mettre en oeuvre
    - Analyse en ligne → Limitations techniques, coûts associés au logiciels pouvant être élevés, logiciel potentiellement onéreux
- Diffusion des fichiers de microdonnées
    - Fichiers à usage public (*PUF*)
    - Fichiers à usage scientifique (*SUF*)
  - Un jeu de données peut être publié dans différents modes (SUF, PUF)
  - Un seul fichier est produit par mode

Avantages de la diffusion des microdonnées

- Favorise et diversifie la recherche (les tabulations ne répondent qu'à des questions pré-définies)
    - Maximise l'utilisation et donc l'utilité des données
    - Justifie les investissements dans la collecte → favorise le financement de la collecte
- Evite la duplication des opérations de collecte
- Améliore la qualité et la crédibilité des données
    - Pertinence, fiabilité grâce aux commentaires des utilisateurs
    - Transparence et reproductibilité en tant que garanties scientifiques
- Satisfait une obligation légale et / ou contractuelle (avec les sponsors)



Coûts et risques de la diffusion des microdonnées

- Garantir la confidentialité et la protection de la vie privée est essentiel pour maintenir la confiance des répondants
- Exposition à la critique
- Estimations non officielles par rapport aux estimations officielles
- Peut avoir à répondre à des questions techniques

- Coût financier (préparation, anonymisation, diffusion des données)
- Construire et maintenir une expertise technique

Protection de la confidentialité

- Définir les conditions d'utilisation de chaque ensemble de microdonnées en fonction de la sensibilité du contenu, du risque de ré-identification:
    A. A qui peut-on donner accès aux données?
    B. Pour quelle utilisation?
    C. Sous quelles conditions?
- Options: fichiers à usage public, accès sous licence, enclave, autres
- Peut avoir > 1 option pour une même enquête, par exemple, PUF et sous license
- Principe: accès aussi libre que possible dans le respect des contraintes légales et éthiques
- Conditions d'utilisation sont une importante mesure de protection des données

- Besoin de contrôle de la divulgation statistique (SDC) / anonymisation des microdonnées

Qu'est-ce que SDC?

- SDC est un processus de traitement de données visant à réduire le risque de divulgation
- L'objectif étant d'atteindre un "niveau acceptable" de risque de divulgation et de permettre une diffusion appropriée
- Le niveau d'acceptabilité du risque de divulgation est généralement à la discretion du producteur de microdonnées et est guidé par une legislation ainsi que par des préoccupations éthiques

Concepts de base: divulgation et risque

- **Divulgation** a lieu lorsqu'une personne ou une organisation (intrus) reconnaît ou apprend quelque chose qu'elle ne savait pas déjà à propos d'une autre personne ou organisation par le biais de données publiées (*Australian Bureau of Statistics*)
  Hypothèse clé: il existe un intrus doté de capacités et de ressources

- **Le risque de divulgation** survient si une estimation inacceptable des informations confidentielles d'un répondant est possible ou si la divulgation exacte est possible avec un niveau de confiance élevé (OCDE)

  quantifié en tant que probabilité de ré-identification correcte / nombre de personnes à risque

- La définition du risque dépend des microdonnées, du type de diffusion et du scénario de divulgation

Concepts de base: types de divulgation

- **Divulgation d'identité** se produit si l'intrus associe un individu connu à un enregistrement des données publiées

- **Divulgation d'attributs** se produit si l'intrus est capable de déterminer de nouvelles caractéristiques d'un individu sur la base des informations disponibles dans les données publiées

- **Divulgation inférentielle** se produit lorsque l'intrus est capable de déterminer la valeur de certaines caractéristiques d'un individu avec plus de précision avec les données publiées que cela n'aurait été possible autrement

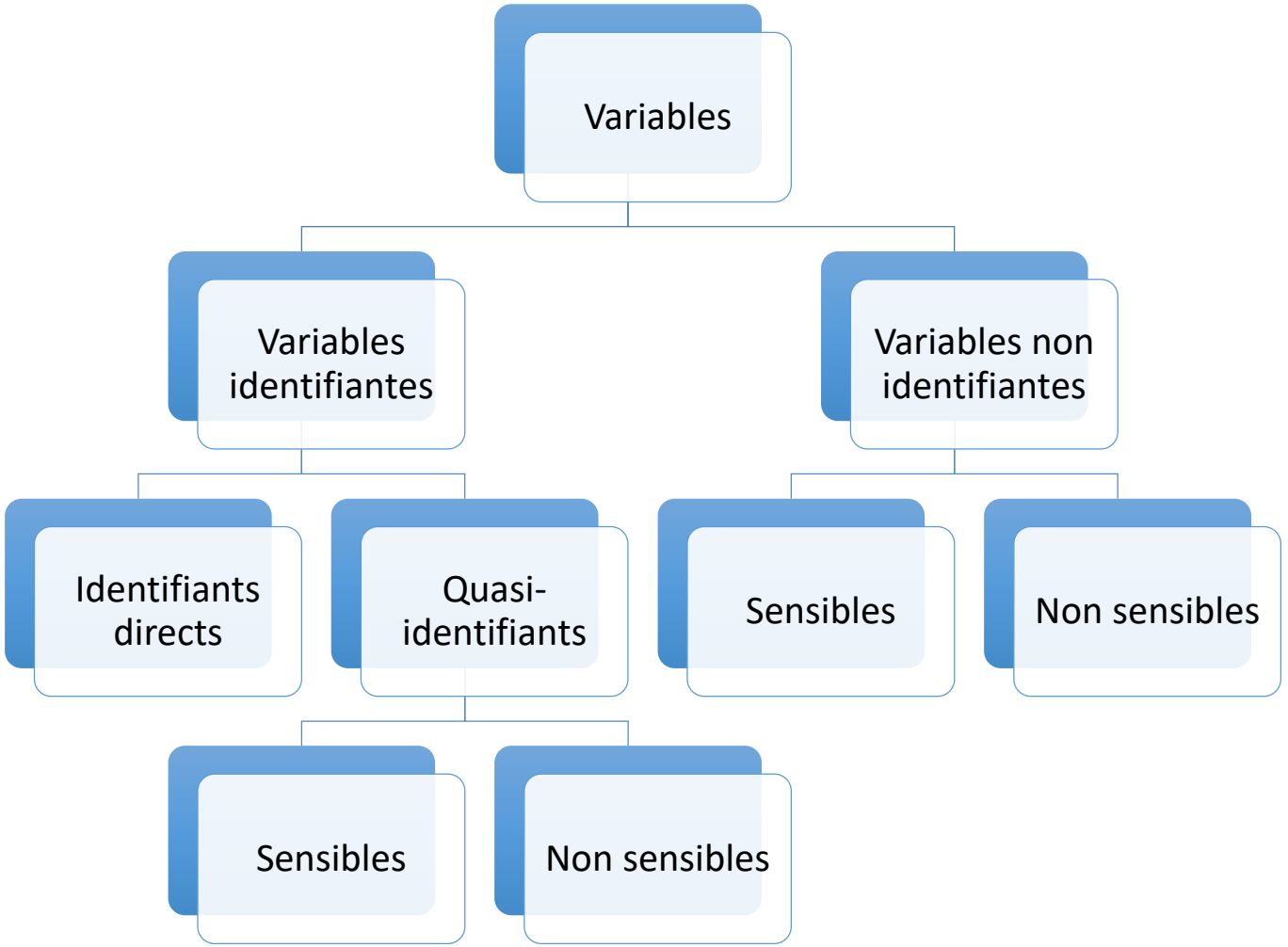
Concepts de base: types de variables

- **Variables identifiantes** contiennent de l'information pouvant contribuer à la réidentification des répondants
  - Identifiants directs → à supprimer
  - Quasi-identifiants (ou indentifiants indirects ou variables clés) → à traiter

- **Quasi-identifiants** peuvent en outre être classés en variables catégorielles (par exemple, sexe, région, nationalité), en variables continues (par exemple, revenus ou dépenses) et en variables quasi-continues (par exemple, âge).

- **Variables sensibles**: variables non-identifiantes, mais qui nécessitent une protection en raison de leur nature

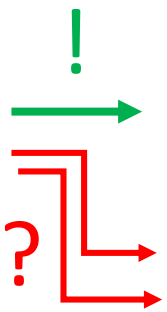
Concepts de base: types de variables



Comment un intrus réidentifie un répondant?

- On suppose que l'intrus possède un fichier de données contenant les identifiants indirects et les noms, permettant de lier les enregistrements

NOM	SEXE	DATE NAISS
ALI	M	12 / 1963
CATHY	F	02 / 1968



SEXE	DATE NAISS	REVENU
M	12 / 1963	80 000
M	07 / 1975	60 000
F	02 / 1968	25 000
F	02 / 1968	150 000

NOM
ALI
??
?
?

- Quels sont ces jeux de données / identifiants?
  - Pas toujours connu. Hypothèse du "scénario du pire" (l'intrus possède un fichier exhaustif, sans erreur, contemporain).
  - Très improbable → on surestime probablement le risque réel

Scénarios de divulgation/ d'intrusion

- Description des données potentiellement disponibles pour les intrus et comment un intrus peut utiliser ces données pour identifier des individus à partir des données
  - Correspondance avec des jeux de données externes (quelles données sont disponibles?)
  - Reconnaissance spontanée (valeurs aberrantes)
- Un intrus est une personne ou organisation qui essaie d'identifier des personnes dans les jeux de données publiés
- Plusieurs scénarios possibles pour un ensemble de données
- Les scénarios d'intrusion dépendent du type de version
- Le résultat final est un ensemble de quasi-identifiants

Mesure du risque – Mesure basée sur la probabilité

- Dans quelle mesure chaque combinaison de variables clés est-elle unique / rare?
  - Mesurer le risque pour chaque observation en fonction de variables clés
  - Échantillon unique ≠ population unique: estimation modélisée du risque pour tenir compte des poids d'échantillonnage
  - Somme des risques individuels → Nombre attendu de réidentifications
  - Peut prendre en compte la structure hiérarchique des données: par exemple, si un membre d'un ménage est réidentifié, tous les autres membres sont réidentifiés → mesure la probabilité qu'au moins un membre soit réidentifié

- Définition d'une cible pour l'anonymisation
  - "Aucune observation avec un risque supérieur à un certain seuil X"
  - "Pas plus de N (or N%) des observations réidentifiées attendues"

Mesurer le risque – Les Fréquences

- Les mesures de risque sont basées sur la fréquence réelle de l'échantillon et sur la fréquence estimée de la population des clés
- Fréquence d'échantillonnage: décompte du nombre d'individus possédant cette clé dans l'échantillon
- Fréquence de la population: estimation du nombre d'individus possédant cette clé dans la population

Individu unique dans l'échantillon et dans la population → le plus grand risque de réidentification

Mesure du risque: k-anonymity

- Risque mesuré par la fréquence de chaque combinaison des varibales clés
- Fréquence basse → Risque de réidentification élevé

SEXE	DATE NAISS	REVENU	fk
M	12 / 1963	80 000	2
M	07 / 1963	60 000	1
F	02 / 1963	110 000	2
F	02 / 1963	150 000	2
M	05 / 1963	75 000	1
M	12 / 1963	100 000	2

SEXE	ANNEE NAISS	REVENU	fk
M	1963	80 000	4
M	1963	60 000	4
F	1963	110 000	2
F	1963	150 000	2
M	1963	75 000	4
M	1963	100 000	4

Moins de détail → Risque réduit

- Définir une cible pour l'anonymisation: imposer une valeur minimale de k (souvent: k=3)

Mesure du risque: l-diversity

- Ajouter un critère de diversité des variables sensibles (“l-diversity”)

NOM	SEXE	DATE NAISS
ALI	M	12 / 1963
ROSE	F	02 / 1968

!

?

SEXE	DATE NAISS	REVENU	NOM
M	12 / 1963	80 000	ALI
M	07 / 1975	60 000	??
F	02 / 1968	150 000	?
F	02 / 1968	150 000	?

- ROSE ne peut pas être liée avec assurance, mais son revenu est certainement 150 000

Réduction du risque de divulgation

- Modifications au jeu de données:
  - Suppression de variables clés (par exemple, identifiant géographique détaillé)
  - Suppression de variables sensibles (ne réduit pas la mesure du risque, mais réduit les conséquences de la divulgation)
  - Suppression de (quelques) observations (valeurs aberrantes)
  - Modification des variables clés (modification/suppression de valeurs spécifiques)
- Plusieurs algorithmes disponibles, optimisés pour minimiser les pertes d'informations
- Pas de solution optimale standard / automatique
- Des "décisions contextuelles " à prendre

Réduction du risque – Quelques leçons apprises

- Beaucoup de recherches universitaires, mais documentation limitée de la pratique
- Risque généralement faible dans les jeux de données d'enquêtes, après suppression des identifiants directs et simple regroupement des variables géographiques
- Les méthodes simples fonctionnent souvent bien; résister à la tentation d'appliquer trop d'algorithmes
- Utilité réduite (trop de transformations sur les données) est également un risque

Réduction du risque: types de méthodes

- **Les méthodes non perturbatrices**: réduisent les détails dans les données ou suppriment certaines valeurs (masquage) sans déformer la structure des données

- **Les méthodes perturbatrices**: perturbent les valeurs pour limiter les risques de divulgation en créant une incertitude autour des vraies valeurs

- Les deux méthodes peuvent être utilisées sur des variables catégorielles et continues

Réduction du risque: quelques méthodes

METHODE	APPLIQUEE AUX VARIABLES DE TYPE	TYPE DE METHODE
Suppression de variables	Continue et catégorielle	Non perturbatrice
Recodage global	Continue et catégorielle	Non perturbatrice
Codage des extrêmes	Continue et catégorielle	Non perturbatrice
Suppression locale	Catégorielle	Non perturbatrice
PRAM	Catégorielle	perturbatrice
Micro-agrégation	Continue	perturbatrice
Ajout de bruit	Continue	perturbatrice
Shuffling	Continue	perturbatrice
Swapping	Continue	perturbatrice

Recodage des variables

- Diminuer le nombre de catégories ou de valeurs distinctes pour une variable, d'où le nombre de combinaisons possibles dans les données publiées
- Peut-être utilisé pour les variables continues et catégorielles
- Le recodage global combine plusieurs catégories d'une variable catégorielle ou construit des intervalles pour des variables continues
  - Exemple: Regrouper les départements en régions; groupes ethniques en groupes ethniques plus larges ;
- Codage des extrêmes: seules les valeurs extrêmes hautes et/ou basses sont remplacées
  - Exemple: Age > 80 → 80+ ; revenu > 100.000 → 100.000+

Suppression locale

- Pour les variables catégorielles ou semi-continues
- Suppression des valeurs individuelles d'une variable (remplacer par la valeur manquante)
- L'algorithme cherche à minimiser le nombre de suppressions
- Deux approches
  - Supprimer les valeurs pour atteindre un certain niveau de k-anonymity
  - Supprimer les valeurs pour les observations avec un risque plus élevé qu'un certain seuil

SEXE	OCCUPATION	AGE	fk
M	FERMIER	25	2
M	FERMIER	26	1
M	DOCTEUR	26	1
M	FERMIER	25	2
F	DOCTEUR	25	1
F	DOCTEUR	26	1

Multiples options → Donner "poids d'importance" à chaque variable

SEXE	OCCUPATION	AGE	fk
M	FERMIER	25	2
M	*	26	2
M	DOCTEUR	26	2
M	FERMIER	25	2
F	DOCTEUR	*	2
F	DOCTEUR	26	2

SEXE	OCCUPATION	AGE	fk
M	FERMIER	25	2
M	FERMIER	26	2
M	*	26	2
M	FERMIER	25	2
F	DOCTEUR	25	2
F	DOCTEUR	*	2

Post-Randomization algorithm (PRAM)

- Echange les catégories (valeurs) d'une variable → ne réduit pas la mesure du risque, mais induit une incertitude
- Une matrice de transition prédéfinie spécifie les probabilités pour chaque catégorie d'être échangée avec d'autres catégories
- Algorithmes permettant d'optimiser la matrice de transition
- Les chercheurs devront être informés de la matrice de transition
- Particulièrement utile lorsque la suppression et le recodage locaux entraîneraient des pertes d’information importantes
- Des combinaisons improbables peuvent être générées si elles ne sont pas prudentes

Post-Randomization algorithm (PRAM)

- Basée sur la matrice de transition :
  - Une observation de l'EST restera toujours à l'EST
  - Une observation de l'OUEST a 10% de chance d'être recodée à l'EST et 5% de chance d'être recodée comme CENTRE.
  - Une observation de CENTRE a 10% de chances d'être recodée comme étant EST et 10% comme OUEST.

VALEUR D'ENTREE \ VALEUR DE SORTIE	EST	OUEST	CENTRE
EST	1.00	0	0
OUEST	0.10	0.85	0.05
CENTRE	0.10	0.10	0.80

Micro-aggregation

- Pour les variables continues
- Crée de petits groupes homogènes d'observations à partir des valeurs d'une ou de plusieurs variables sélectionnées (par exemple, revenu), et remplace ces valeurs par la moyenne ou médiane du groupe
- Les algorithmes diffèrent:
  - Comment sont définis les groupes homogènes
  - L'algorithme pour trouver de tels groupes
  - La valeur de remplacement

Micro-aggregation univariée

- La formation des groupes est simple
  - Trier par valeur de la variable
  - Créer g groupes de taille n_i
  - Ceci maximise l'homogénéité au sein du groupe (basée sur la somme des erreurs au carrés - *SSE*)
- Nous remplaçons ensuite les valeurs par la moyenne ou la médiane du groupe

Exemple avec des groupes de 3, remplacement par la moyenne

IDmen	REVENU
4	12,000
7	12,200
6	13,250
1	14,000
9	15,500
8	16,000
2	17,250
5	18,000
3	45,000

IDmen	REVENU
4	12,483
7	12,483
6	12,483
1	15,167
9	15,167
8	15,167
2	26,750
5	26,750
3	26,750

Micro-agrégation multivariée

- Appliquer la micro-agrégation à chaque variable conduit à une probabilité de réidentification élevée
- La micro-agrégation multivariée crée des groupes homogènes basés sur plusieurs variables. La 1ère étape est la création des groupes homogènes basés sur des distances multivariées entre les individus. Puis remplacer toutes les valeurs d'un groupe par des valeurs moyennes pour le groupe

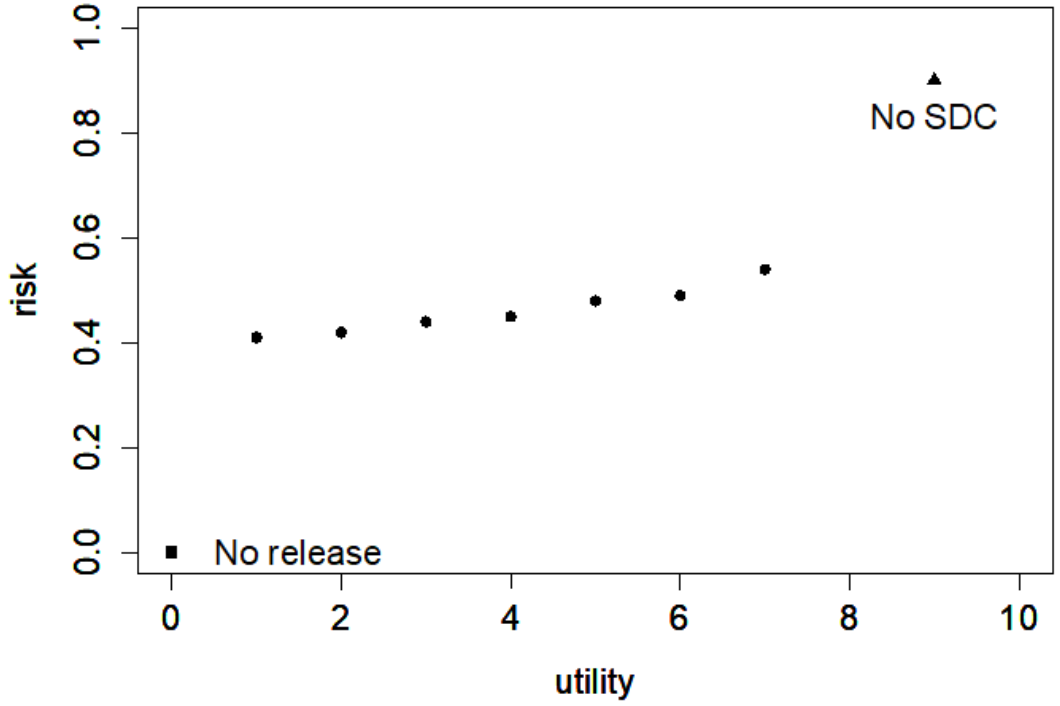
ID	Group	Before microaggregation			After microaggregation		
		Income	*Exp*	*Wealth*	*Income*	*Exp*	*Wealth*
1	1	2,300	1,714	5.3	2,285.7	1,846.3	6.3
2	1	2,434	1,947	7.4	2,285.7	1,846.3	6.3
3	1	2,123	1,878	6.3	2,285.7	1,846.3	6.3
4	2	2,312	1,950	8.0	3,567.3	2,814.0	8.3
5	2	6,045	4,569	9.2	3,567.3	2,814.0	8.3
6	2	2,345	1,923	7.8	3,567.3	2,814.0	8.3

Equilibre entre risque et perte d'information

- La suppression et la modification des données entraînent toujours une perte d'informations
- L'anonymisation peut donc rendre le jeu de données :
  - Moins pertinent (par exemple, revenu de codage extrême (top) → impact sur l'analyse des inégalités
  - Moins fiable
- Trop de transformations → estimations officielles non réplicables
- Défi: trouver le bon équilibre
- Objectif: minimiser les modifications apportées aux données pour atteindre un niveau de risque acceptable (le risque est réduit, pas totalement éliminé)

Mesures de la perte d'information (de l'utilité)

- Nombre de valeurs manquantes
- Nombre d'enregistrements modifiés
- IL1S → compare la distance entre les valeurs d'origine et les valeurs perturbées (pour les variables continues).
- Statistiques sommaires: moyenne, variance ou covariance
- Tabulations
- Comparaison des indicateurs clés ou des coefficients de regression avant et après anonymisation



Outil et pratique de l'anonymisation

- Logiciel spécialisé (libre): sdcMicro
- Packages d'analyse statistique pour certains algorithmes (pas tous)
- Manuels recommandés:


IHSN

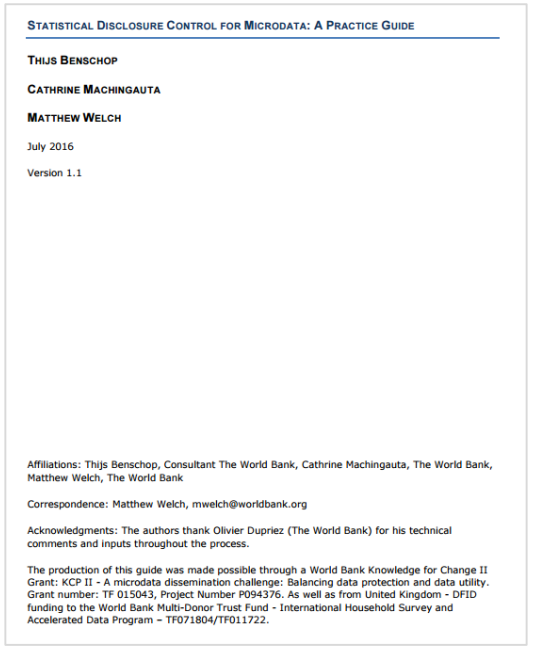
INTERNATIONAL HOUSEHOLD SURVEY NETWORK

Introduction to Statistical Disclosure Control (SDC)

Matthias Templ, Bernhard Meindl, Alexander Kowarik and Shuang Chen

IHSN Working Paper No 007
August 2014

www.ihsn.org


STATISTICAL DISCLOSURE CONTROL FOR MICRODATA: A PRACTICE GUIDE

THIJS BENSCHOP

CATHRINE MACHINGAUTA

MATTHEW WELCH

July 2016

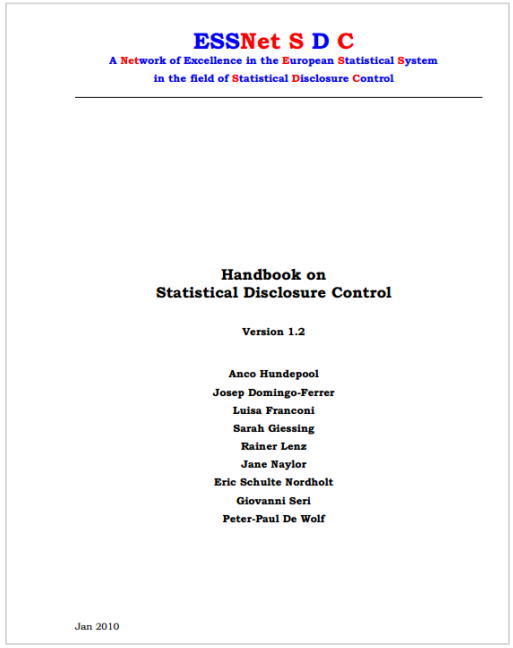
Version 1.1

Affiliations: Thijs Benschop, Consultant The World Bank, Cathrine Machingauta, The World Bank, Matthew Welch, The World Bank

Correspondence: Matthew Welch, mwelch@worldbank.org

Acknowledgments: The authors thank Olivier Dupriez (The World Bank) for his technical comments and inputs throughout the process.

The production of this guide was made possible through a World Bank Knowledge for Change II Grant: KCP II - A microdata dissemination challenge: Balancing data protection and data utility. Grant number: TF 015043, Project Number P094376. As well as from United Kingdom - DFID funding to the World Bank Multi-Donor Trust Fund - International Household Survey and Accelerated Data Program – TF071804/TF011722.


ESSNet S D C

A Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control

Handbook on Statistical Disclosure Control

Version 1.2

Anco Hundepool
Josep Domingo-Ferrer
Luisa Franconi
Sarah Giessing
Rainer Lenz
Jane Naylor
Eric Schulte Nordholt
Giovanni Seri
Peter-Paul De Wolf

Jan 2010

Conclusion, recommandations

- Diffusion des microdonnées
  - Dans un environnement favorable, en suivant des politiques / procédures claires et en conformité avec les meilleures pratiques internationales
- La diffusion de microdonnées nécessite:
  - Documentation (facilité d'utilisation) → Norme de métadonnées DDI
  - Catalogue (visibilité, découvrabilité) → catalogue NADA
  - Protection: conditions d'utilisation → Publier une politique et des protocoles
  - Protection: anonymiser → sdcMicro
  - Transfert → NADA, recommandations IHSN
  - Recueil et exploitation des commentaires des utilisateurs

Expérience pratique

- **Réduction du risque à l'ANSD**

  - Suppression des identifiants directs
  - Recodification des variables (typologie à prendre en compte)
  - Identification et traitement des valeurs extrêmes (suppression locale)
  - Renumérotation District de recensement
  - Traitement des variables géographiques

Diffusion des microdonnées – Limites et perspectives

- Garantir que la divulgation est impossible
- Trouver le juste milieu entre perte d'informations et utilité
- Trouver un moyen de rendre générique le processus
- Gérer de manière efficace les variables « Autres à préciser »

- Promouvoir la diffusion des microdonnées
- Travailler avec les utilisateurs avisés afin d'améliorer la qualité des fichiers de microdonnées

MERCI DE VOTRE ATTENTION