Règles de confidentialité

Version 0.4

TABLE DES MATIERES

1 DESCRIPTION DU DOCUMENT

Lors de vos traitements des données auxquelles vous avez été autorisés à accéder, vous serez amenés à opérer des exports de fichiers depuis votre environnement de travail sécurisé CASD. Ce processus est appelé « demande de sortie ». Vous trouverez dans ce document l'ensemble des règles de confidentialité à appliquer aux fichiers que vous souhaitez exporter. Il s'agit de vous aider à respecter les différents types de secret applicables (statistique, fiscal, etc.) et de s'assurer qu'aucune information permettant d'identifier une personne physique, un ménage ou une entreprise ne sera divulguée.

Si votre sortie est de nature à générer une rupture du secret applicable aux données concernées, l'équipe statistique du CASD vous en informera et vous aidera à masquer des cases concernées ou à agréger des variables différemment pour que le problème soit résolu (zone géographique plus grande, tranche d'âge regroupé au lieu de l'âge fin…).

2 LES TYPES DE SORTIES

Tout d'abord, vous devez décrire de manière précise le contenu de votre sortie, soit dans votre mail de demande de sortie, soit dans un fichier texte inclus dans la sortie. Vous trouverez une description de la procédure pour effectuer une sortie dans le guide utilisateur disponible sur le site web du CASD.

2.1 LES PROGRAMMES

Vous pouvez demander à sortir vos programmes pour les réutiliser en dehors du CASD. Le programme ne doit pas contenir de données.

2.2 LES REGRESSIONS, MODELES ECONOMETRIQUES

Les demandes de sorties peuvent contenir des résultats de modèles économétriques sous SAS, Stata, R, etc. Vous devez garder dans les fichiers de sortie les paramètres et les indicateurs connus liés à ces modèles économétriques (estimateurs, R^2, pseudo R^2, intervalle de confiance, Khi^2, etc.) et le nombre d'observations afin de permettre à l'équipe statistique du CASD de s'assurer qu'il s'agit bien d'une régression ou modèle économétrique.

2.3 LES GRAPHIQUES ET LES CARTES

Pour les graphiques comme les courbes, les histogrammes, les nuages de points, les diagrammes, etc., vous devez fournir les informations qui ont permis leur construction (population et signification des variables utilisées).

D'autres types de graphiques sont plus complexes à traiter car ils peuvent contenir des informations individuelles. Par exemple, les box plot peuvent contenir non seulement le maximum et le minimum mais aussi des points extrêmes (outliers) qui ne doivent pas être identifiés.

Attention aux graphiques d'analyse factorielle représentant les individus s'ils contiennent des individus atypiques qui peuvent être identifiés (par exemple une ACP sur des entreprises où le SIRET est utilisé sur le graphique comme identifiant de chaque point).

Les graphiques Stata en format LIVE ne sont pas autorisés à être exportés car ils contiennent les données qui ont permis la constitution de ces graphiques. Vous pouvez choisir entre les deux options suivantes :

- Enregistrer les graphiques STATA en format AS-IS, qui ne contiennent pas les bases de données, au lieu de les enregistrer en format LIVE (le format par défaut)
- Convertir les graphiques en jpg, bitmap, pdf, etc.

Les cartes peuvent concerner des populations très fines (communales ou infra-communales). Il faut, pour vérifier que les règles de confidentialité sont bien respectées sur la carte, que vous fournissiez les données qui ont permis de construire cette carte.

2.4 LES TABLEAUX DE DONNEES AGREGEES

Vous devez fournir dans la description des sorties toutes les informations permettant de les comprendre et de identifier les types de données utilisés : liste des variables, intitulés en clair, la signification, les effectifs correspondant à chaque case et l'information portant sur la plus forte contribution dans la case.
De manière générale, les règles de confidentialité sont différentes et propres à chaque type de source.

2.5 LES ARTICLES « FINALISES »

Les articles finalisés que vous souhaitez sortir de votre espace projet CASD ne doivent pas contenir de données sur une toute petite population d'individus.
Par ailleurs, votre article doit indiquer que votre travail a été rendu possible grâce au dispositif CASD de la manière suivante :
« Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'avenir portant la référence ANR-10-EQPX-17 (Centre d'accès sécurisé aux données – CASD) » ou en anglais ***« This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-EQPX-17 - Centre d'accès sécurisé aux données – CASD) »***.

3 REGLES GENERALES

3.1 LES DONNEES « MENAGES »

Pour les tableaux fournissant des données agrégées sur les ménages, la seule règle imposée est que l'identification directe ou indirecte des individus soit impossible. Dans la pratique, on considère que le secret statistique est respecté si la connaissance d'une caractéristique pour un individu ne peut pas entraîner la connaissance d'une autre caractéristique avec laquelle elle est croisée dans un tableau.

Exemple :
Le tableau ci-dessous donne la répartition par âge et la situation matrimoniale et indique que les personnes entre 50 et 59 ans ont toutes le même état matrimonial « divorcé ». Le secret statistique n'est plus respecté dans ce tableau, et ce dernier n'est donc pas diffusable. En effet, si l'on sait par ailleurs qu'un individu donné se situe dans la tranche des 50-59 ans, le tableau nous informe que cette personne est divorcée, et ceci même si la case qui croise les modalités « 50 à 59 ans » et « divorcé » comporte plusieurs individus.

Situation matrimoniale et classe d'âge	18-25 ans	26-49 ans	50-59 ans	60 ans et +
Marié	7	27	0	30
Divorcé	0	11	9	22
Autre	21	12	0	4

3.2 Les donnees « Entreprises »

Pour les tableaux fournissant des données agrégées sur les entreprises, la règle est la suivante :

- Aucune case du tableau ne doit concerner moins de trois unités (décision du 13 juin 1980 du directeur général de l'Insee).
- Aucune case du tableau ne doit contenir des données pour lesquelles une entreprise représente plus de 85% du total (règle du CNIS, 7 juillet 1960).

3.3 Les sources mixtes

Les sources mixtes sont des sources provenant de combinaisons (appariements) d'enquêtes statistiques et de données administratives ou bien des sources comportant à la fois des informations d'ordre économique et financier (entreprises) et des informations relatives à des faits et comportements d'ordre privé (ménages).

La démarche à adopter face à de telles sources est, dans son principe, très simple : les règles à prendre en considération s'obtiennent par le cumul des règles applicables d'une part aux enquêtes statistiques, d'autre part aux fichiers administratifs. Par exemple, l'enquête revenus fiscaux et sociaux repose sur la combinaison d'enquêtes statistiques, des résultats de l'enquête emploi, de données fiscales et des données fournies par les Caisses d'allocations familiales, ou bien les enquêtes Esane (Élaboration des Statistiques Annuelles d'Entreprises) et Fusain (FUsion des Statistiques Annuelles dans l'INdustrie).

3.4 Les donnees issues de sources fiscales

- Mêmes règles que pour les entreprises
- Pour les données ménages : seuil de 11 individus

3.5 Secret primaire

Les cases qui ne respectent pas les règles de confidentialité et qui sont donc masquées forment ce que l'on appelle le secret primaire. Il est simple à gérer, il faut appliquer les règles propres à chaque source.

Exemple :

Classe d'âge	Nombre de salariés
20-34 ans	7
35-49 ans	2
50-64 ans	10

Classe d'âge	Nombre de salariés
20-34 ans	7
35-49 ans	s
50-64 ans	10

3.6 Secret secondaire

Le secret secondaire est plus complexe à gérer, il est lié à la présence des marges dans un tableau et empêche la reconstitution, par somme ou par différence, des cases masquées au secret primaire.

Les situations qui peuvent mener au secret secondaire sont les suivantes :

3.6.1 Marges et variables hiérarchisées

En raison des marges diffusées dans les tableaux ou publiées sur internet, il est parfois possible de retrouver les cases masquées du secret primaire. Pour y remédier, il suffit de masquer deux cases.

Exemple :

Classe d'âge	Nombre de salariés
20-34 ans	7
35-49 ans	2
50-64 ans	10
Total	19

Classe d'âge	Nombre de salariés
20-34 ans	s
35-49 ans	s
50-64 ans	10
Total	19

Les variables hiérarchisées concernent souvent des variables géographiques ou des nomenclatures, il faut les traiter comme des marges et les imbriquer dans les tableaux.

Exemple :

Ces deux tableaux ne peuvent pas être diffusés indépendamment car il y a des relations d'additivité. Si un utilisateur souhaite sortir le premier tableau donnant le nombre d'entreprises en Bretagne, il doit, selon les règles de confidentialité appliquées sur les données entreprises, masquer la case mentionnant le nombre d'entreprises dans le Finistère. ATTENTION : dans cet exemple, il est impératif que le second tableau, qui donne le nombre d'entreprises au niveau régional et permet de retrouver la case masquée dans le premier, ne fasse jamais l'objet d'une demande de sortie ultérieure ou soit diffusé de quelque manière que ce soit. Vous pouvez par exemple anticiper tout risque potentiel de brèche de confidentialité en masquant deux cases du premier tableau.

Département	Nombre d'entreprises
Morbihan	8
Finistère	2
Côtes-D'Armor	9
Ille-et-Vilaine	6

Région	Nombre d'entreprises
Basse-Normandie	20
Bretagne	25
Pays de la Loire	28

Département	Nombre d'entreprises
Morbihan	8
Finistère	s
Côtes-D'Armor	9
Ille-et-Vilaine	s

3.6.2 1 seule modalité non nulle

Avoir une seule modalité non nulle dans un tableau permet de connaître la caractéristique d'un individu ou d'une entreprise à partir d'un(e) autre. Pour l'éviter, il faut avoir au moins 2 modalités non nulles.

Exemple :

Si, après traitement, on obtient un tableau indiquant que les personnes ayant entre 20 et 34 ans ont toutes la même situation professionnelle (demandeurs d'emploi), le secret statistique n'est plus respecté dans ce tableau. En effet, si l'on sait par ailleurs qu'un individu donné se situe dans la tranche des 20-34 ans, ce tableau nous informe que cette personne est nécessairement demandeuse d'emploi, et ceci même si la case qui croise les modalités « 20-34 ans » et « demandeur d'emploi » comporte plusieurs individus (7 en l'occurrence).

Situation professionnelle et tranche d'âge	Salarié	Demandeur d'emploi	Non salarié	Autre
20-34 ans	0	7	0	0
35-49 ans	6	5	5	8
50-64 ans	4	5	6	5

Situation professionnelle et tranche d'âge	Salarié	Demandeur d'emploi	Non salarié	Autre
20-34 ans	0	s	s	0
35-49 ans	6	5	5	8
50-64 ans	4	5	6	5

3.7 FICHIER DE CONTROLE

Afin de vérifier les résultats portant sur les variables de montant, vous devez également transmettre un fichier de contrôle contenant toutes les informations à exporter ainsi que des colonnes indiquant le maximum et le pourcentage du maximum des variables de montant.
Ce fichier de contrôle sera retiré de la sortie avant son envoi.

Exemple :

Fichier de contrôle

	Nombre entreprises	Montant du max	Montant du total	% du max
Bretagne	***139***	***1 668***	***27 800***	***6%***
Morbihan	82	1 590	19 882	8%
Finistère	19	590	3 476	17%
Côtes-d'Armor	36	1 103	2 567	43%
Ille-et-Vilaine	2	1 312	1 875	70%
Picardie	***99***	***825***	***20 643***	***4%***
Oise	67	825	13 750	6%
Aisne	5	722	821	88%
Somme	27	790	6 072	13%

Fichier de sortie

	Nombre entreprises	Montant total
Bretagne	***139***	***27 800***
Morbihan	82	19 882
Finistère	S3	S4
Côtes-d'Armor	36	2 567
Ille-et-Vilaine	S1	S2
Picardie	***99***	***20 643***
Oise	67	13 750
Aisne	5	S9
Somme	27	S10

	Valeur cachée parce que :
S1	Nombre d'entreprises inférieur à 3
S2	Informations relatives à S1
S3	Pour ne pas recalculer S1 à l'aide du total
S4	Pour ne pas recalculer S2 à l'aide du total
S9	Pourcentage du maximum supérieur à 85%
S10	Pour ne pas recalculer S9 à l'aide du total

4 REGLES DETAILLEES POUR CHAQUE SOURCE

Pour les sources produites par l'Insee, les règles du secret sont décrites dans le guide du secret statistique. Ce guide est disponible à l'adresse :

http://www.cnis.fr/files/content/sites/Cnis/files/Fichiers/comite_du_secret/COMITE_DU_SECRET_guide.PDF

4.1 DADS (DECLARATION ANNUELLE DE DONNEES SOCIALES)

Tout tableau diffusé ne doit en aucun cas permettre l'identification directe ou indirecte d'un salarié ou d'un établissement.

- Tableaux au lieu de résidence (logique « ménages »)
  - au moins 5 salariés par case
  - aucune case avec 1 salarié représentant + 80 % de la masse salariale

- Tableaux au lieu de travail (logique « entreprises » en plus)
  - au moins 5 salariés par case
  - aucune case avec 1 salarié représentant + 85 % de la masse salariale
  - on ajoute les critères classiques liés aux entreprises

4.2 CLAP (CONNAISSANCE LOCALE DE L'APPAREIL PRODUCTIF)

CLAP fait partie des données « entreprises » et les règles sont les suivantes :

- au moins 3 unités par case
- aucune case avec 1 unité représentant + de 85 % du total
- au moins 5 salariés par case
- unité = établissement (ou entreprise)
- indicateurs soumis au secret = effectifs et rémunérations.

4.3 FARE (FICHIER APPROCHE DES RESULTATS D'ESANE) / FICUS (FICHIER DE COMPTABILITE UNIFIE DANS SUSE)

Les données FARE et FICUS sont des sources mixtes « fiscale et statistique ». Les règles de confidentialité applicables sont donc le cumul des règles applicables d'une part aux enquêtes statistiques, d'autre part aux données fiscales.

- au moins 3 unités par case
- aucune case avec 1 unité représentant + de 85 % du total

- pour les données ménages : seuil de 11 individus

4.4 SINE (SYSTEME D'INFORMATION SUR LES NOUVELLES ENTREPRISES)

SINE est également une source « entreprises », les règles de confidentialité à appliquer sont les suivantes :

- aucun résultat qui concerne moins de trois entreprises
- ni aucune donnée pour laquelle une seule entreprise représente 85 % ou plus de la valeur obtenue.

Par ailleurs, les taux de survie ne doivent pas être calculés pour des populations de moins de 20 entreprises. Ce seuil minimum de 20 entreprises est également requis pour les zonages et regroupements particuliers.

4.5 RECENSEMENT DE LA POPULATION

4.5.1 RP 1999 et avant

Les principes de diffusion des données du RP à partir de 1999, reposent sur l'arrêté du 22 mai 1998, modifié le 8 avril 2002, relatif à la diffusion des résultats du recensement de la population.
Il porte principalement sur les variables sensibles, telle que l'article 10 les définit :

- Au moins 4 unités observées par case (avant pondération)
    - au moins 10 pour un sondage à 40 %
    - au moins 16 pour un sondage au ¼
- Respecter les seuils de diffusion des variables dites « sensibles » : nationalité actuelle (ou à la naissance), lieu de naissance, lieu de résidence antérieur, année (ou ancienneté) d'arrivée en France :
    - communes de + 5 000 habitants,
    - zones infra communales résultant du regroupement de 3 quartiers
    - seuils de 10 000 habitants pour les arrondissements, zones d'emploi, aires urbaines, unités urbaines et zones de la politique de la ville
    - département pour l'année (ou ancienneté) d'arrivée

4.5.2 RP annuel à partir de 2006

Les principes de diffusion des données du RP à partir de 2006, reposent sur l'arrêté du 19 juillet 2007 relatif à la diffusion des résultats du recensement de la population.
Il porte principalement sur les variables sensibles, telle que l'article 8 les définit :

- Au moins 4 unités observées par case (avant pondération)
    - au moins 10 pour un sondage à 40 %
    - au moins 16 pour un sondage au ¼
- Respecter les seuils de diffusion des variables dites « sensibles » : nationalité actuelle (ou à la naissance), lieu de naissance, lieu de résidence antérieur, année (ou ancienneté) d'arrivée
    - communes de + 5 000 habitants
    - seuil de 5 000 habitants pour les arrondissements, ZE, AU, UU et zones de la politique de la ville
    - département pour l'année (ou ancienneté) d'arrivée

4.6 SIASP (SYSTEME D'INFORMATION SUR LES AGENTS DES SERVICES PUBLICS)

La diffusion de résultats statistiques tirés des données issue de SIASP doit être conforme aux dispositions prévues par les textes relatifs au secret en matière de statistique et à la protection des données individuelles. En particulier, aucun tableau destiné à la diffusion ne doit permettre l'identification directe ou indirecte d'un salarié ou d'un établissement :

- aucune case ne doit comporter moins de 5 salariés
- aucun salarié ne doit représenter plus de 80 % de la masse salariale de la case
- aucune case ne doit se rapporter à moins de 3 établissements
- aucun établissement ne doit représenter plus de 85 % de la grandeur étudiée dans la case.

Sur le champ des fonctions publiques territoriale et hospitalière et des établissements publics, un établissement étant bien identifié par son SIRET, cette règle peut être mise en œuvre sans difficulté.

Sur le champ des ministères, le « contour » d'un établissement n'existe pas dans tous les cas, ces derniers n'étant pas toujours sirétisés.

Aussi retient-on les règles suivantes :

- au niveau national, régional et départemental : sont diffusables les effectifs et masse salariale (en distinguant indiciaire et primes total) par caractéristiques du salarié et/ou de son contrat de travail (sexe, âge, statut, catégorie…) par ministère à condition d'avoir au moins 5 agents par case, aucun salarié ne représentant plus de 80 % de la case. Certaines de ces données sont d'ailleurs déjà proposées dans les « chiffres-clés régionaux » (TCRD et RED)
- au niveau infra-départemental, ce sont les « règles classiques » du secret statistique qui sont appliquées.

4.7 DONNEES DEPP

Les règles pour les données de la DEPP sont les suivantes :

- si les données sont individuelles : aucune case avec moins de 10 individus
- si les données sont agrégées au niveau établissement (collèges, lycée, etc…), les cases ne doivent pas contenir moins de 10 établissements.

4.8 DONNEES SSP

Les exploitations agricoles sont assimilées aux entreprises et soumises aux mêmes règles de confidentialité que les données entreprises :

- aucune case du tableau ne doit concerner moins de trois entreprises/exploitations agricoles (le secret s'applique sur les exploitations et non sur les parcelles)
- aucune case du tableau ne doit contenir des données pour lesquelles une exploitation agricole représente plus de 85% du total.

Il faudra fournir un fichier de contrôle non secrétisé permettant de vérifier le respect des règles.

4.9 DONNEES SDSE

A minima un effectif de 5 observations dans chaque case du tableau.