*9ème colloque francophone sur les sondages*
*Gatineau, Canada*
*Atelier de formation – 11 octobre 2016*

# LA VÉRIFICATION DES DONNÉES PRINCIPES, MISE EN OEUVRE

Philippe BRION
INSEE

# 1. Introduction

# 1. Introduction (1)

- Le contrôle des données : un sujet a priori connu de tous les responsables de la production de statistiques : les données brutes sont rarement utilisables « directement » …

… en même temps, peu d'ouvrages de référence sur le sujet (en particulier en français) – *imputation des non réponses mise à part* -, alors qu'il mobilise une partie importante des ressources des INS (instituts nationaux de statistique)

# 1. Introduction (2)

En anglais, *statistical data editing* (SDE) : recouvre la localisation des erreurs puis leur correction

*Remarque : parfois, référence dans la formation aux termes utilisés en anglais*

# 1. Introduction (3)

Granquist (1995) donne quatre objectifs principaux pour le *statistical data editing* :

- Identifier les sources d'erreur pour avoir ensuite un retour sur l'ensemble du processus d'enquête
- Fournir de l'information sur la qualité des données (individuelles et résultats statistiques)
- Identifier et traiter les erreurs les plus importantes
- Si nécessaire, fournir un ensemble complet et cohérent de données individuelles

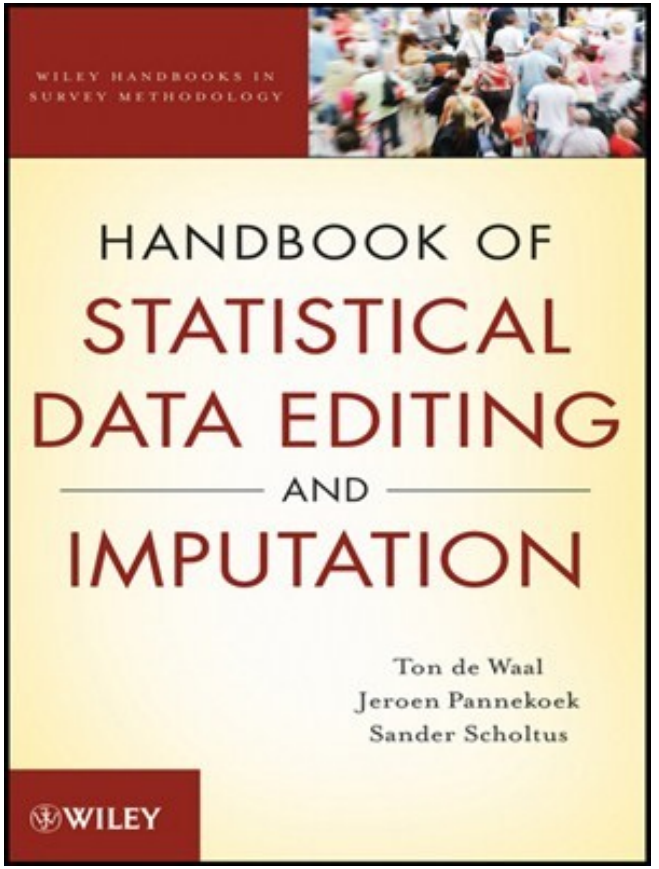
# 1. Introduction (4)

Une évolution de la manière de mener le travail de vérification des données qui a conduit à se poser la question de « l'acharnement »

- voir en particulier le papier de Granquist et Kovar (1997) : *Editing of survey data : how much is enough ?*

# 1. Introduction (5)

La formation s'appuie en grande partie sur le manuel de De Waal, Pannekoek et Scholtus (publié en 2011 chez John Wiley)



# 1. Introduction (6)
# Différents types d'erreurs

- systématiques / aléatoires

- influentes / non influentes

- *outliers* : valeurs aberrantes (qui ne s'ajustent pas sur un modèle), mais pas forcément erreurs (l'*outlier* peut dans certains cas être l'enregistrement complet)

- données manquantes

# 1. Introduction (7)
# Différents types de contrôles

*Hard edits* (vérifications avec rejet)

*Soft edits* (vérifications avec avertissement)

Contrôles relatifs à une variable, contrôles faisant intervenir plusieurs variables

# 1. Introduction (8)
# Différents types de contrôles (suite)

Égalité : CA=VA+CI

Inégalité : CA ≥ 0

La valeur d'une variable catégorielle appartient à une liste de modalités prédéfinie

Si … alors … : par exemple si âge ≤ 18 permis de conduire = non

Contrôles statistiques :

valeur comprise dans ($m-C_1S$, $m+C_2S$) où m est un indicateur de valeur moyenne et S un indicateur de dispersion

Ratio (comme par exemple le Chiffre d'affaires rapporté au nombre de salariés) compris dans une fourchette prédéfinie

# 1. Introduction (9)
# Différents types de méthodes

- Méthodes manuelles / méthodes automatiques

- Les méthodes manuelles font souvent appel à de l'information « externe » (voire au recontact de l'unité enquêtée)

- Elles sont en général ciblées sur les erreurs influentes

- Les méthodes automatiques ont l'avantage d'être « reproductibles », ce qui n'est pas nécessairement le cas des méthodes manuelles

- Le choix de la (ou des) méthode(s) dépend du contexte (enquêtes sociales / enquêtes économiques), et donc du type de statistiques produites

# 1. Introduction (10)
# Plan de la formation

- Méthodes automatiques
- Méthodes manuelles
- Éléments sur les méthodes d'imputation adaptées au contexte de la correction des données
- Mise en place d'une organisation globale

*Remarque : la formation n'examine pas, de manière détaillée, chacune des méthodes présentées, mais a pour objectif de dresser un panorama de l'ensemble de celles-ci, et de montrer la manière dont elles s'articulent*

# 2. Méthodes de contrôle automatique

2.1. Méthodes « déductives »
2.2. Méthodes basées sur le paradigme de Fellegi-Holt
2.3. Autres méthodes

# 2.1. Méthodes « déductives »

# 2.1. Méthodes déductives

Méthodes utilisées pour la détection d'erreurs systématiques (et pour lesquelles le « mécanisme d'erreur » est connu), et utilisées en général au début du processus de correction

On s'appuie sur la valeur fournie pour proposer une valeur corrigée, à partir de mécanismes de correction prédéfinis

# 2.1. Méthodes déductives (suite)

Erreurs classiques :

- Erreurs de signe
- Erreurs d'arrondi
- Erreurs de saisie
- Erreurs d'unités (en particulier en termes de milliers)

# 2.1. Méthodes déductives (suite)

- Algorithmes de détection - correction de ces erreurs proposés dans deux documents de travail de Scholtus (2008, 2009)

- Parfois utilisation de règles simples : valeur d'une variable déduite d'une autre (*ce qui nécessite de mettre cette dernière en « prioritaire »*)

- Idée : avoir un maximum de ces erreurs corrigées avant de traiter les erreurs qui ont un comportement plus aléatoire

# 2.1. Méthodes déductives (suite)

Exemple 1 : erreur d'unité

Idée : utiliser une valeur anticipée (donnée passée pour le même enregistrement, ou valeur médiane, par exemple), et observer le rapport entre la donnée fournie et cette valeur anticipée

Si rapport > seuil (par exemple 300) diviser la valeur fournie par 1000

Autre méthode : partir du nombre de caractères de la valeur fournie (comparé au nombre de caractères de la valeur anticipée)

# 2.1. Méthodes déductives (suite)

Exemple 2 : erreur de signe

Idée : utiliser des variables auxiliaires en (-1,1) et trouver la solution « optimale »

# 2.1. Méthodes déductives (suite)

Exemple 3 : erreur d'arrondi

On a par exemple des contraintes du type :

*Ax=b*

On tolère une erreur d'arrondi si

$$0 < \left| a_k^T x - b_k \right| \leq \delta$$

Où $a_k^T$ est la kième ligne de la matrice *A*

# 2.1. Méthodes déductives (suite)

Exemple 3 : erreur d'arrondi (suite)

Si on se limite aux erreurs d'arrondi, on élimine les lignes telles que

$$\left|a_k^T x - b_k\right| > \delta$$

Le problème se ramène à la résolution de

$A_1 x' = b_1$

# 2.2. Méthodes basées sur le paradigme de Fellegi-Holt

# 2.2. Les méthodes basées sur le paradigme de Fellegi-Holt (1)

Le paradigme de Fellegi-Holt (voir article de 1976) :

- on part d'un ensemble de contrôles définis *a priori* (et internes à un enregistrement) …

... et on cherche à changer le moins possible de valeurs d'origine de façon à rendre l'enregistrement cohérent, en s'appuyant sur la fonction objectif suivante :

$$\sum_{j=1,n} w_j \delta\left(x_j, \hat{x}_j\right)$$

- Où $\left(x_{1,}\ldots,x_n\right)$ est l'enregistrement de départ, et $\left(\hat{x_{1,}}\ldots,\hat{x_n}\right)$ un enregistrement qui satisfait les contrôles, et *wj* est un poids indiquant la fiabilité qu'on accorde à chaque variable *j*

Remarque : ceci ne garantit pas de « trouver » l'erreur ...

# 2.2. Les méthodes basées sur le paradigme de Fellegi-Holt (2)

- L'article de Fellegi-Holt (FH) indique qu'il faut dériver un ensemble de contrôles implicites, à partir des contrôles de départ (explicites) …

… pour se ramener à un problème de couverture d'ensemble

- De nombreux algorithmes ont été proposés pour localiser l'erreur à partir du paradigme de FH ; une des difficultés est liée à la génération des contrôles, qui prend beaucoup de temps

- Illustration sur deux exemples

# 2.2. Les méthodes basées sur le paradigme de FH (3) - exemple avec variables numériques

Quatre variables :

CA (chiffre d'affaires), P (profit), C (charges), N (nombre de salariés)

On accorde un « poids de confiance » de 1 à CA, P et T, et de 2 à N

Contrôles définis a priori :

(1) CA = P+C

(2) P ≤ 0.5 CA

(3) -0.1CA ≤ P

(4) CA ≥ 0

(5) CA ≤ 550 N

*Un enregistrement avec les valeurs CA=100, P=40000, C=60000, N=5*

*Les contrôles 1 et 2 posent problème, on pourrait penser qu'en modifiant la valeur de CA on règle la question, mais ...*

# 2.2. Les méthodes basées sur le paradigme de FH (4) - exemple avec variables numériques

- Méthode proposée par de Waal Quéré (2003) : algorithme *branch-and-bound* (méthode arborescente) s'appuyant sur la construction d'un arbre « binaire »

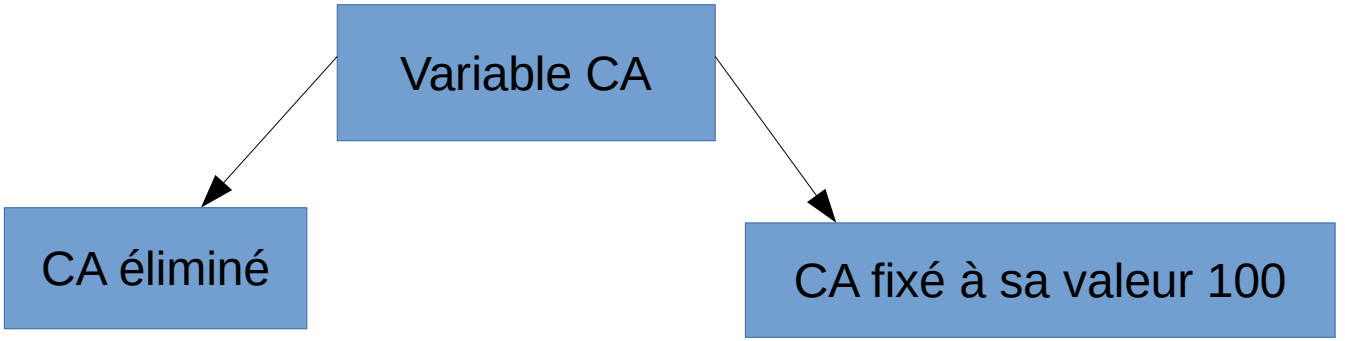
Méthode implantée dans le logiciel SLICE / Cherry Pie de *Statistics Netherlands*

- L'idée est de trouver l'ensemble minimum de variables couvrant l'ensemble des contrôles à partir de l'arbre, où, après sélection d'une variable, on envisage deux « chemins » : l'un où la valeur de la variable est fixée à sa valeur d'origine, l'autre où la variable est « éliminée » (et devra donc être imputée)

Dans les deux cas, l'ensemble des contrôles (qu'ils posent problème ou pas) doit être actualisé (on utilise la méthode de Fourier - Motzkin)

# 2.2. Les méthodes basées sur le paradigme de FH (5) - retour sur l'exemple

*On choisit d'abord la variable CA - dans les deux cas on actualise les contrôles*




| | |
|---|---|
| $P \leq 0.5\ (P+C)$ | OK |
| $-0.1\ (P+C) \leq P$ | OK |
| $P + C \geq 0$ | OK |
| $P + C \leq 550$ N | pb |

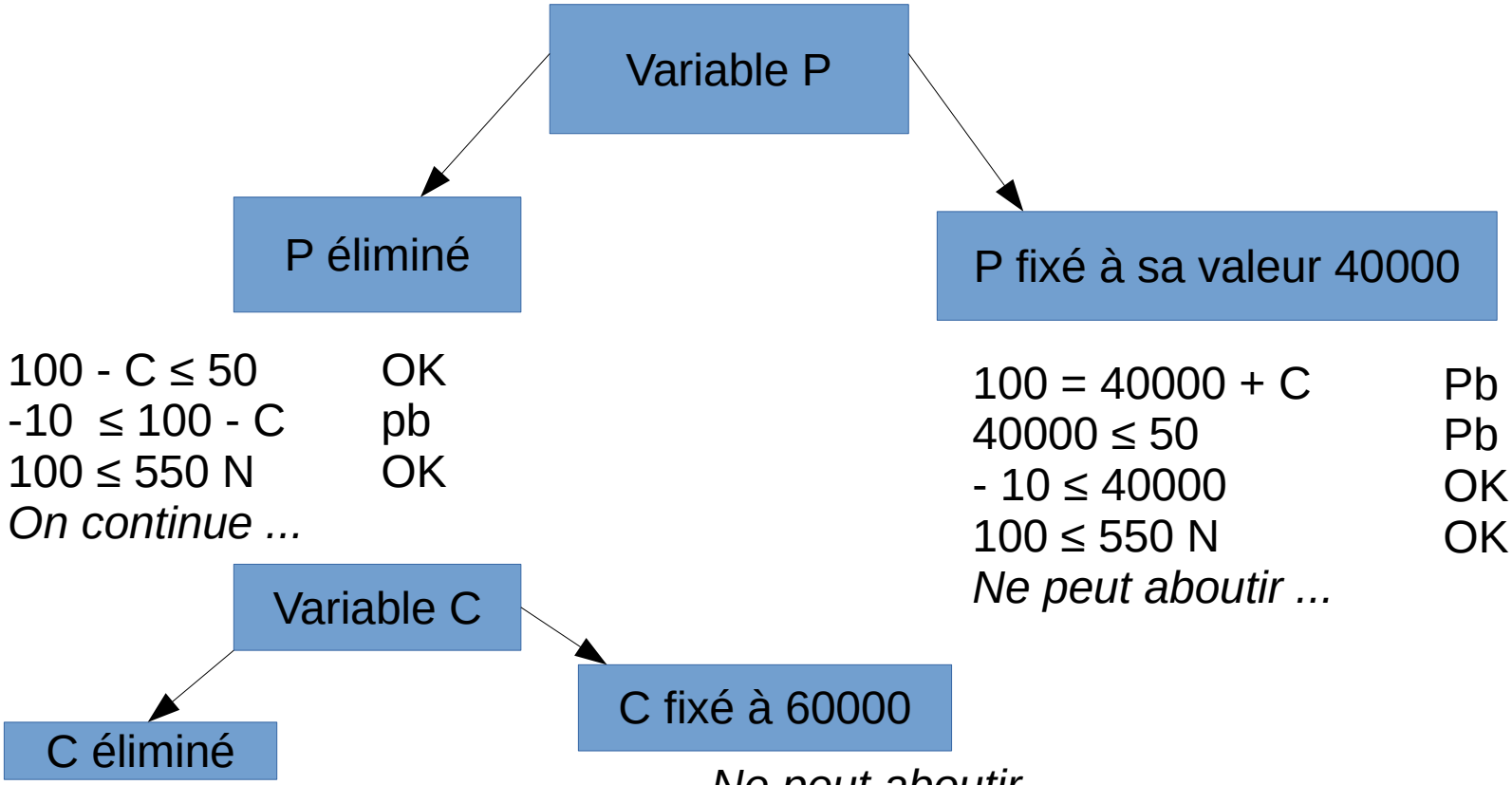
*On devra donc continuer*
*On arrivera à une « fonction objectif » au moins égale à 3*

| | |
|---|---|
| $100 = P + C$ | Pb |
| $P \leq 50$ | Pb |
| $-10 \leq P$ | OK |
| $100 \leq 550$ N | OK |

*On continue sur cette branche*

# 2.2. Les méthodes basées sur le paradigme de FH (6) - retour sur l'exemple

*On continue avec CA à sa valeur d'origine, et on choisit la variable P*




Ne reste que 100 ≤ 550 N OK

***Fonction objectif = 2 ; changer P et C !!!***

# 2.2. Les méthodes basées sur le paradigme de FH (7) - cas des variables qualitatives

L'article de Fellegi-Holt propose une formalisation des contrôles, qui permet ensuite de générer les contrôles implicites

Un contrôle peut se formaliser comme :

$$\left(F_{1,}\ldots,F_{n}\right)$$

Où $F_i$ est un sous-domaine de l'ensemble des valeurs possibles pour la variable *i ; remarque : le contrôle n'est pas satisfait si les modalités de l'enregistrement respectent les conditions définies par la formalisation …*

# 2.2. Les méthodes basées sur le paradigme de FH (8) - cas des variables qualitatives

Exemple : trois variables

Âge : modalités 1 (0 à 14 ans), 2 (15 - 80), 3 (81 et plus)

Statut matrimonial : modalités 1 (marié), 2 (non marié)

Lien avec le chef de ménage : modalités 1 (épouse), 2 (enfant), 3 (autre)

Contrôle 1 : si âge < 15 ans alors statut matrimonial = non marié

s'exprime comme

$$\left(\{1\},\{1\},\{1,2,3\}\right)$$

# 2.2. Les méthodes basées sur le paradigme de FH (9) - cas des variables qualitatives

Génération de contrôles implicites : pour un ensemble S de contrôles *j* : $\left(F_{1,}^{j},\ldots,F_{n}^{j}\right)$

Tel que $\coprod_{j\in S} F_{r}^{j}=D_{r}$ pour une variable *r*

On peut définir un contrôle implicite

$$\left(\prod_{j\in S} F_{1,}^{j},\ldots,\prod_{j\in S} F_{(r-1)}^{j},D_{r},\prod_{j\in S} F_{(r+1)}^{j},\ldots,\prod_{j\in S} F_{n}^{j}\right)$$

si pour chaque *i* différent de *r* , $\prod_{j\in S} F_{i}^{j}\neq\emptyset$

# 2.2. Les méthodes basées sur le paradigme de FH (10) - cas des variables qualitatives

Retour sur l'exemple :

On ajoute un deuxième contrôle explicite : un individu non marié ne peut être l'épouse du chef de ménage

Le contrôle C2 : $\left(\{1,2,3\},\{2\},\{1\}\right)$

s'ajoute à C1 : $\left(\{1\},\{1\},\{1,2,3\}\right)$

On génère le contrôle C3, implicite, à partir de la variable statut matrimonial (qui, *in fine*, est éliminée du contrôle) :

$$\left(\{1\},\{1,2\},\{1\}\right)$$

*Qui indique qu'un individu de moins de 15 ans ne peut être l'épouse du chef de ménage*

# 2.2. Les méthodes basées sur le paradigme de FH (11) - cas des variables qualitatives

- Pour la localisation des erreurs, on peut utiliser le même type de méthode que pour les variables numériques (*méthode arborescente permettant de tester l'ensemble des combinaisons de l'arbre*), et la génération des contrôles, lorsqu'on élimine une variable, se base sur le principe précédent (issu de l'article de Fellegi-Holt)

- Pour un mélange de variables qualitatives et numériques, méthode d'élimination des variables plus complexe (*cf. chapitre 4.4 du manuel de De Waal, Pannekoek et Scholtus*)

# 2.2. Les méthodes basées sur le paradigme de FH (12) - conclusions

- Le paradigme de FH repose sur l'idée que les erreurs sont aléatoires

- Les valeurs manquantes sont considérées « éliminées »

- Il peut y avoir plusieurs solutions optimales : dans ce cas, tirage au sort, ou, mieux, sélection de celle qui s'ajuste le mieux à un modèle (après utilisation de méthodes d'imputation)

# 2.2. Les méthodes basées sur le paradigme de FH (13) - conclusions (suite)

- Le paradigme de FH considère que les contrôles sont tous des contrôles stricts (*hard*) (cf. exemple précédent : contrôle P ≤ 0.5 CA)
  - *mais article de Scholtus (2013a) proposant une méthode pour prendre en compte les contrôles avec avertissement (soft)*

- En pratique : il est recommandé de limiter l'usage de la méthode proposée aux cas où le nombre de variables posant problème est limité
  - *Statistics Netherlands, par exemple, met systématiquement en expertise manuelle un enregistrement pour lequel l'algorithme dépasse plusieurs minutes*

# 2.3. Autres méthodes

# 2.3. La méthode NIM (1)

- Cette méthode (*Nearest-Neighbor Imputation Methodology*) constitue une alternative aux méthodes basées sur le paradigme de FH

- Elle réalise de manière simultanée la localisation des erreurs et leur correction, en s'appuyant sur la méthode du hotdeck (donneur), et peut traiter des données catégorielles ou numériques

*Référence : Bankier & al (2000)*

*Logiciel CANCEIS (Statistique Canada)*

# 2.3. La méthode NIM (2)

- Principe de la méthode : avoir un enregistrement imputé proche de l'enregistrement original …

… ainsi que du donneur

- définir une distance entre enregistrements :

$$D(x_{1,}x_2)=\sum_{j=1,p} w_j D_j(x_{1j},x_{2j})$$

Où *j* indice les variables, et $w_j$ est un poids affecté à chaque variable

# 2.3. La méthode NIM (3)

- Pour chaque enregistrement $x_f$ qui « déclenche » au moins un contrôle, on sélectionne les *Nd* donneurs potentiels (par exemple *Nd*=40) qui sont les plus proches au sens de la distance définie

- On génère ensuite des enregistrements « adaptés » $x_a$ avec la méthode d'imputation suivante, en utilisant les différents donneurs :

$$x_{aj} = \delta_j x_{dj} + \left(1 - \delta_j\right) x_{fj}$$

Avec $\delta_j = 0,1$ , $x_f$ données de départ et $x_d$ données du donneur, *j* indiçant les variables

- on peut avoir plusieurs solutions possibles respectant les contrôles

# 2.3. La méthode NIM (4)

- On souhaite rester le plus proche possible, pour les $x_a$, de $x_f$, mais également de $x_d$ (à savoir être proche d'un enregistrement correct), ce qui assure une certaine plausibilité à $x_a$

- Pour ce faire, on calcule la distance

$$\mu(I) = \alpha D(x_f, x_a) + (1-\alpha) D(x_a, x_d)$$

Avec $\alpha$ compris entre 1/2 et 1

- Remarque : le même donneur peut fournir plusieurs imputations possibles

# 2.3. La méthode NIM (5)

- On conserve ensuite les $x_a$ tels que

$$\mu(I) \leq \gamma \mu_{min}$$

Et on en choisit un de manière aléatoire avec une probabilité proportionnelle à $\left(\frac{\mu_{min}}{\mu(I)}\right)^t$

# 2.3. NIM versus Fellegi-Holt

- NIM plus rapide

- NIM utilise les éléments tirés de la distribution des donneurs pour identifier des imputations « plausibles »

- NIM demande beaucoup de donneurs potentiels

- si les contrôles sont constitués d'égalités comptables nombreuses, il est difficile de trouver des donneurs qui respectent ces égalités

- NIM est plus difficile à utiliser sur des données obtenues sur un échantillon

# 2.3. Une autre méthode : la méthode des CQR

Méthode utilisée pour la production des statistiques structurelles d'entreprise à l'Insee (Rivière, 1996)

Idées de base :

- Chaque micro-contrôle produit une note
- Ensuite, pour une variable donnée, l'ensemble des notes obtenues pour les contrôles où la variable est impliquée est « combiné » en une note permettant de qualifier, ou non, la variable
- Pour une valeur non qualifiée, on essaie un ensemble d'imputations ; si on arrive à une valeur « correcte », on obtient une valeur redressée, sinon, l'enregistrement est mis en erreur
- Nécessité de définir une hiérarchie dans les variables (on travaille par groupes de variables)

# 3. Méthodes manuelles

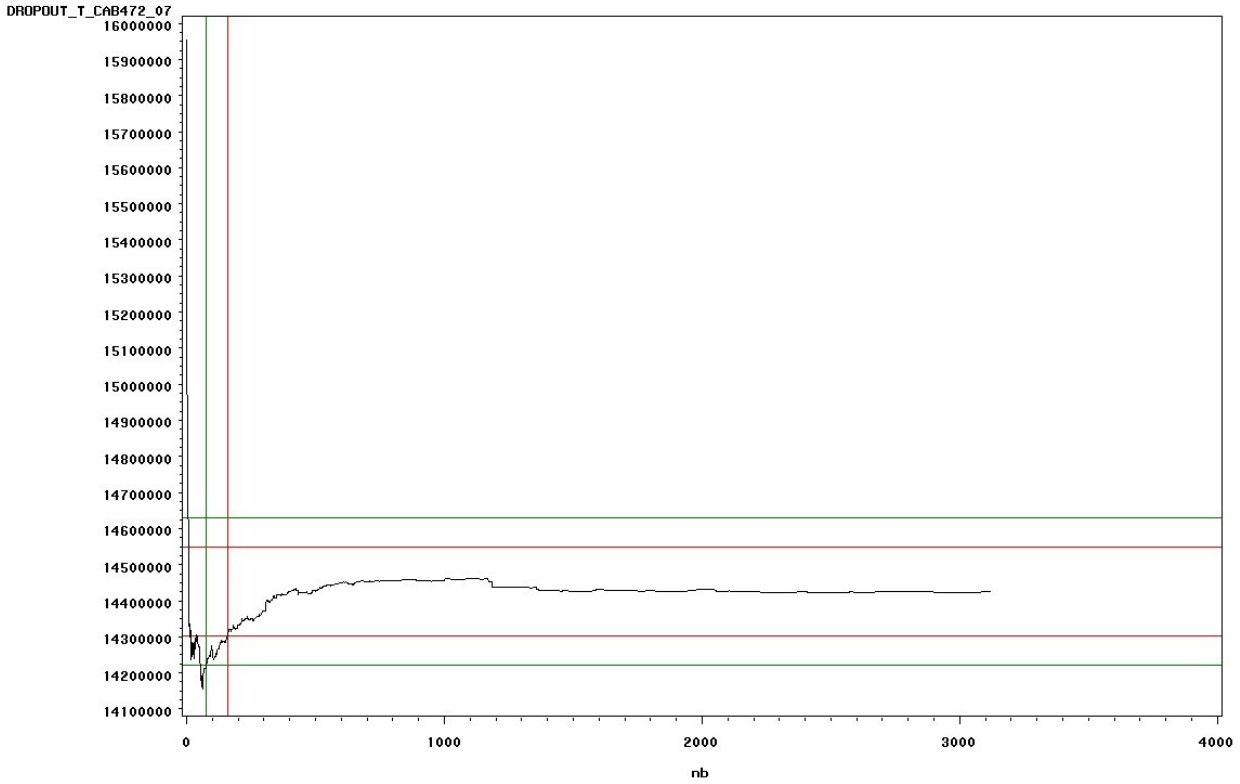
3.1. La vérification sélective des données
3.2. Les méthodes d'*output editing*
3.3. La vérification interactive

# 3. Les méthodes manuelles, introduction

- La vérification manuelle est rarement appliquée sur l'ensemble des données, et l'idée est alors de séparer les enregistrements en deux paquets :
  - un qui est examiné de manière manuelle par des gestionnaires
  - un pour lequel on utilise les données brutes fournies par les répondants, ou pour lequel on utilise une méthode de redressement automatique
- L'idée est d'éviter « l'acharnement », dans le travail de vérification

# 3. Les méthodes manuelles, introduction (suite)

Illustration sur l'enquête annuelle d'entreprises (Insee, 2007) : estimation du chiffre d'affaires de la branche « commerce de détail alimentaire », en « intégrant » progressivement les corrections faites sur les données



# 3.1. La vérification sélective des données

# 3.1. La vérification sélective des données (1)

En anglais : *selective editing*

Idée : utiliser un score pour déterminer les enregistrements à vérifier de façon manuelle, en fonction de leur impact sur des statistiques définies a priori

On commence donc, pour une statistique « cible », par définir un score pour chaque enregistrement, dit score local, comme l'impact attendu de la vérification de la valeur d'une variable sur la statistique cible (fonction DIFF introduite par Latouche Berthelot (1992))

# 3.1. La vérification sélective des données (2)

Par exemple, pour l'estimation du total d'une variable *X*, le score local vaut la différence entre l'estimation avec tous les enregistrements vérifiés, et l'estimation avec tous les enregistrements vérifiés sauf un :

$$w_i\left(x_i - \hat{x}_i\right)$$

Où : $x_i$ est la valeur « brute » de la variable *X* pour l'unité *i*

et $\hat{x}_i$ est la valeur corrigée de *X* pour *i*

# 3.1. La vérification sélective des données (3)

Le problème est qu'on ne dispose pas de la donnée corrigée ; on fait appel à un prédicteur, $\tilde{x}_i$ , qui peut être :

- la valeur de la même variable en (t-1) (éventuellement multipliée par une évolution moyenne)
- une valeur obtenue par un ratio appliqué à une variable auxiliaire
- une valeur obtenue dans une source externe à l'enquête
- la valeur médiane de la variable (ou la moyenne), pour la catégorie à laquelle appartient l'unité

# 3.1. La vérification sélective des données (4)

Le score local peut s'écrire comme :

$$score = w_i \left| x_i - \tilde{x}_i \right| = w_i \tilde{x}_i \frac{\left| x_i - \tilde{x}_i \right|}{\tilde{x}_i}$$

Et se réécrire en deux composantes

$w_i \tilde{x}_i$ est l'influence de l'unité i

$\frac{\left| x_i - \tilde{x}_i \right|}{\tilde{x}_i}$ est le risque associé à la valeur observée

# 3.1. La vérification sélective des données (5)

Pour définir un classement des unités à contrôler de façon manuelle, il est nécessaire de calculer un score global « agrégeant » les scores locaux :

Pour cela il faut d'abord « normer » les scores locaux en les rapportant :

- au total de la variable concernée

- à l'erreur d'échantillonnage relative à l'estimation du total de cette variable

# 3.1. La vérification sélective des données (6)

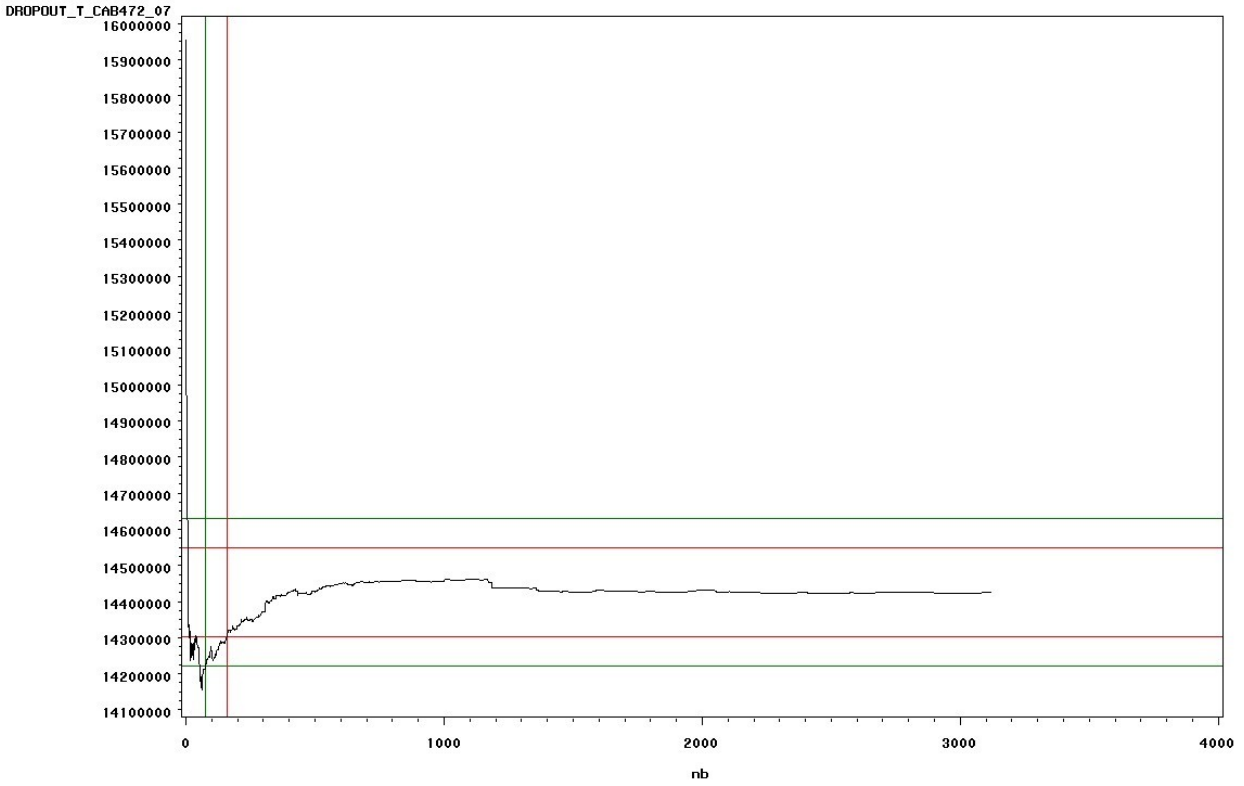
Différentes méthodes de calcul du score global (Hedlin, 2008):

- Somme des scores locaux
- Maximum des scores locaux
- Distance euclidienne (Farwell, 2005) $\sqrt{\sum \left(S_j^2\right)}$
- Distance de Minkowski (Hedlin, 2008)

$$\left( \sum_{j=sclocaux} S_j^a \right)^{\left(\frac{1}{a}\right)}$$

# 3.1. La vérification sélective des données (7)

La détermination du seuil au-dessus duquel on contrôle de façon manuelle peut se faire de manière manuelle, à partir par exemple d'une erreur acceptable par rapport à l'erreur d'échantillonnage (dans le graphique ci-dessous, on utilise 30 % et 50 % de l'erreur d'échantillonnage) ; pour un exemple de mise au point industrielle des seuils, voir par exemple Gros (2009)



# 3.1. La vérification sélective des données (8)

- Dans la pratique, la méthode fonctionne bien pour les statistiques économiques (donc sur des variables quantitatives)
  - Exemple du dispositif de statistiques structurelles d'entreprises en France
- … mais un certain nombre de réglages souvent nécessaires (Gros, 2012): mise au point de totaux « stables » pour la normalisation des scores locaux, méthode d'agrégation pour déterminer le score global, par exemple
- Parfois, introduction d'une composante suspiscion d'erreur en fonction du résultat des contrôles (Nordberg & al, 2010)

# 3.1. La vérification sélective des données (9)

- Peu de papiers donnant une « théorisation » de la méthode, à l'exception de Hesse (2005), Di Zio Guarnera (2013), et Arbues & al (2013)
- Des variables pour lesquelles la méthode est efficace, d'autres pas :
  - Problème pour les variables avec beaucoup de valeurs à zéro
  - Problème de la qualité du prédicteur (Brion, 2016)

# 3.1. La vérification sélective des données (10)

- Statistiques d'intérêt autres que des totaux (par exemple ratio) : utiliser une variable linéarisée (Lawrence McKenzie, 2000)
- On a parfois une situation différente, selon les unités de l'échantillon, pour la disponibilité de valeurs « prédites »
- Dans la pratique, on garde souvent les très grandes unités pour une expertise manuelle
- Logiciel SELEKT (Nordberg & al, 2010)

# 3.1. La vérification sélective des données (11)

<u>La méthode de Hidiroglou Berthelot (1986), pour les enquêtes entreprises périodiques</u>

Idée : utiliser un intervalle pour définir des évolutions plausibles, en utilisant des intervalles interquartiles

$$(r_M - kDRQ_1,\ r_M + kDRQ_3)$$

Où $r_M$ médiane des rapports $x_{t+1}/x_t$

$DQR_1$ distance entre premier quartile et médiane

$DQR_3$ distance entre troisième quartile et médiane

# 3.1. La vérification sélective des données (12)

La méthode de Hidiroglou Berthelot (suite)

Problème : la variabilité des rapports est plus forte pour les petites unités, d'où l'idée de symétriser :

$s_i = 1 - r_M / r_i$ si $0 < r_i < r_M$

$s_i = r_i / r_M - 1$ si $r_i \geq r_M$

Intégrer également un indicateur de taille pour définir un nouvel indicateur

$E_i = s_i \{max(x_i(t), x_i(t+1))\}^U$

# 3.1. La vérification sélective des données (13)

La méthode de Hidiroglou Berthelot (suite)

C'est pour cet indicateur $E_i$ qu'on va définir un intervalle à l'extérieur duquel les données sont jugées aberrantes :

$$(E_M - Cd_{Q1}, E_M + Cd_{Q3})$$

Avec $E_M$ médiane des $E_i$

et $d_{Q1} = Max(E_M - E_{Q1}, |AE_M|)$

# 3.2. Les méthodes d'«output editing »

# 3.2. Les méthodes d'*output editing*

Ces méthodes sont appliquées à partir d'un fichier complet de données (et par conséquent sont utilisables pour le traitement des données administratives)

Deux grands types de méthodes :

- Méthode des agrégats
- Méthodes basées sur l'étude de la distribution des variables

# 3.2. Les méthodes d'*output editing :* la méthode des agrégats

Approche descendante : on repère les agrégats « bizarres » (au sens où ils sont comparés à un prédicteur, par exemple la valeur n-1) à un niveau de diffusion « élevé »

… puis on passe à un niveau plus fin d'agrégation

… afin de cibler, *in fine*, les enregistrements qu'on va traiter de manière interactive

# 3.2. Les méthodes d'*output editing :* la méthode des agrégats (suite)

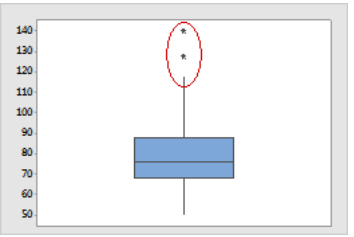
L'« agrégat » peut être un ratio, permettant de contrôler une variable (par exemple chiffre d'affaires) en utilisant une variable auxiliaire (par exemple nombre de salariés)

Le principe de la méthode est proche de celui de la vérification sélective, mais :

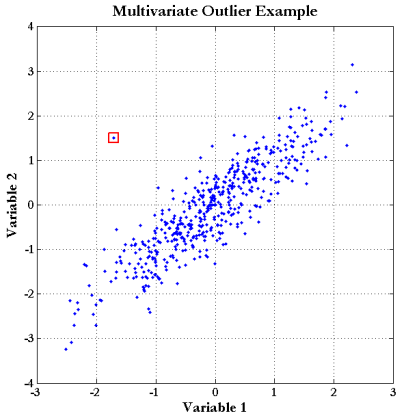
- Les données utilisées sont disponibles pour l'ensemble des unités (on n'est donc pas obligé d'utiliser une donnée « fabriquée » pour certaines unités)
- Les poids de sondage définitifs sont également disponibles
- On n'a pas à calculer un seuil pour la détermination des unités à expertiser de manière manuelle

# 3.2. Les méthodes d'*output editing :* l'étude de la distribution de variables

Utilisation de méthodes d'analyse exploratoire des données, pour identifier les valeurs aberrantes :




- « boites à moustache »
- graphiques avec deux variables, pour identifier des valeurs aberrantes relativement à une distribution bivariée (par exemple pour la même variable en t et t-1)



# 3.3. La vérification interactive des données

# 3.3. La vérification interactive (1)

- Dans ce cas on procède à une expertise manuelle, avec rappel éventuel des unités posant problème
- S'appuyer sur des contrôles assistés par ordinateur
  - Messages d'erreur
  - Possibilité de vérifier immédiatement le résultat des contrôles quand on corrige une donnée
  - Matériau disponible : scan du questionnaire papier, données anciennes, etc.

# 3.3. La vérification interactive (2)

- Le contrôle peut être fait au moment de la saisie (CAPI, CATI, collecte internet, mais aussi saisie de masse …)
  - Pour la saisie de masse deux options :
    - saisie brute
    - saisie incorporant des contrôles
- Dans la pratique :
  - avoir des zones commentaires
  - importance des réunions de mise en commun
  - garder trace des changements

# 4. Éléments sur les méthodes d'imputation adaptées au contexte de la correction des données

# 4.1. Les méthodes d'imputation (1)

- Les méthodes d'imputation : un sujet connu, développé essentiellement pour traiter les valeurs manquantes
- Différents objectifs :
  - Estimation d'agrégats
  - Estimation de distributions
  - Production de fichiers de données individuelles

# 4.1. Les méthodes d'imputation (2)

- Différentes méthodes :
  - Déductives
  - Basées sur un modèle : estimation par la régression, par le ratio, utilisation d'une valeur moyenne, etc.
  - Modèles d'imputation multi-variables
  - Non paramétriques : hotdeck, éventuellement utilisant une fonction de distance
  - Cas particulier des variables à « fréquence de réponse rare » : procéder à une imputation en deux temps

# 4.2. Les méthodes d'imputation adaptées au contrôle des données

- Les contrôles ajoutent des contraintes supplémentaires, concernant les données imputées
- Deux grandes familles de méthodes :
  - Méthodes incorporant les contraintes liées aux contrôles dès l'imputation
  - Méthodes pour lesquelles on impute d'abord sans tenir compte des contrôles, et où on « ajuste » ensuite les données

# 4.3. Méthodes d'imputation incorporant les contrôles (1)

- Méthodes déductives

  Si … alors …

  Égalité comptable

  $Ax=b$

  devient $A_{manq}x_{manq}=b - A_{obs}x_{obs}$

- La méthode du « hotdeck ratio » : si dans une égalité comptable plusieurs valeurs sont manquantes, on distribue « ce qui manque » à partir d'un donneur et du ratio entre ce qui manque chez le receveur et ce qui « aurait manqué » chez le donneur
- Pour les variables catégorielles, possibilité d'utiliser les principes présentés dans la procédure d'élimination de Fellegi-Holt

# 4.3. Méthodes d'imputation incorporant les contrôles (2)

- *Exemple d'imputation déductive utilisant le « principe » FH*

  4 variables V1 V2 V3 V4

  de domaines D1={1,2,3,4}

  D2=D3={1,2,3}

  D4={1,2}

  4 contrôles (1) D1x{3}x{1,2}x{1}

  (2) D1x{2,3}xD3x{2}

  (3) {1,2,4}x{1,3}x{2,3}xD4

  (4) {3}xD2x{2,3}x{1}

# 4.3. Méthodes d'imputation incorporant les contrôles (3)

- *Exemple d'imputation déductive utilisant le « principe » FH (suite)*

  On a un enregistrement (3,2,-,-) : on doit donc imputer V3 et V4

  Si on renseigne les valeurs de V1 et V2 dans les contrôles, restent les contrôles (2) et (4) qui deviennent :

  D3x{2}

  {2,3}x{1}

  On choisit d'imputer V3 : si on élimine V4 des contrôles avec la « procédure FH », reste le contrôle {2,3} pour V3, variable qu'on impute donc à 1 ; la valeur imputée pour V4 est également de 1

# 4.3. Méthodes d'imputation incorporant les contrôles (4)

- Méthodes paramétriques
- Un exemple : l'utilisation d'une loi normale tronquée, par exemple pour respecter le caractère positif d'une variable
  - On tire dans la loi normale, et on itère si le résultat est négatif
- Utilisation de loi « multivariables »
- Le problème est que les variables, souvent, s'ajustent difficilement à des tels modèles

# 4.3. Méthodes d'imputation incorporant les contrôles (5)

- Méthode basée sur la procédure d'élimination de Fourier Motzkin

  Cette méthode utilise une approche séquentielle, ce qui est plus simple du point de vue de la spécification des modèles d'imputation

- *Un exemple (tiré de Scholtus (2013b)) :*

  $\begin{pmatrix} x \\ y \end{pmatrix}$ suit une loi N($\begin{pmatrix} 60 \\ 55 \end{pmatrix}$, $\begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}$)

# 4.3. Méthodes d'imputation incorporant les contrôles (6)

- *Exemple, suite : méthode basée sur la procédure d'élimination de Fourier Motzkin*

  Les contrôles à respecter sont $x \geq 50$

  $$100 \geq y$$

  $$y \geq x$$

  On impute y à partir de N(55,100), avec une loi normale tronquée : y = 70

  On actualise les contrôles $x \geq 50$

  $$70 \geq x$$

  On tire x à partir de N(60,100) dans l'intervalle (50,70) : 52

# 4.4. L'ajustement des données imputées

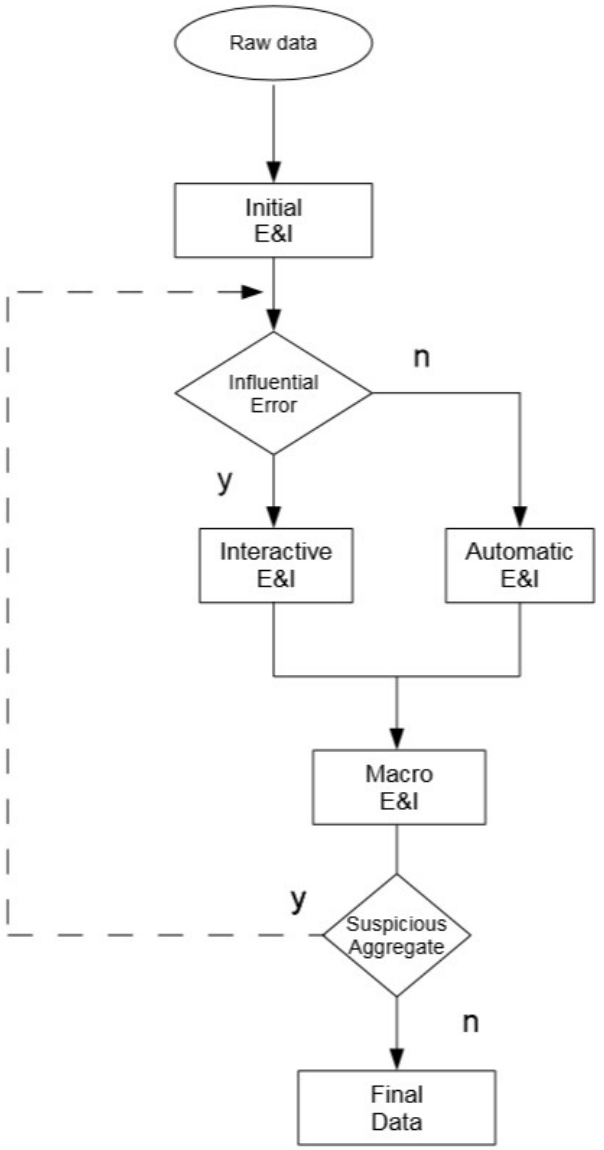
- Cette méthode suppose que l'imputation est d'abord réalisée sans tenir compte des contrôles …
- Ensuite, on cherche à minimiser une fonction de distance entre les valeurs imputées « initiales » et des valeurs proches satisfaisant les contrôles (problème d'optimisation)
- Ce type de méthodes est utilisé dans l'approche générale de réconciliation des données

# 5. Mise en place d'une organisation globale

# 5. 1. Mise en place d'une organisation globale

- Les parties précédentes ont présenté un ensemble de méthodes qui constitue une boîte à outils …
- … la question est de savoir comment les combiner, en tenant compte de différentes composantes :
  - Calendrier de production des statistiques
  - Appropriation des méthodes par les gestionnaires chargés du travail de contrôle
  - Attentes des utilisateurs
  - Aspects budgétaires (et donc arbitrages entre méthodes automatiques et manuelles)

# 5.2. Une proposition d'articulation des différents méthodes



# 5.3. La prise en compte du calendrier

- Dans le cas d'une enquête par courrier ou par internet, on ne maîtrise pas toujours le calendrier de retour des questionnaires
- Il va falloir adapter la charge de travail à la ressource disponible, d'où l'intérêt d'avoir des indicateurs de priorité pour les données à traiter
  - Voir par exemple Merad Wagstaff (2005)

# 5.4. Le travail des équipes de gestionnaires

- La mise en place de méthodes automatisées, ou de vérification sélective, va parfois à l'encontre du souhait des gestionnaires de mener un travail de vérification « jusqu'au bout »
- L'appropriation des méthodes passe donc par de la pédagogie
- Par ailleurs, importance des réunions de mise en commun relativement au travail d'expertise manuelle, afin d'éviter des pratiques divergentes

# 5.5. Les attentes des utilisateurs

- Définition de statistiques cibles, ainsi que des niveaux de diffusion
- Attentes en matière de calendrier
- Données cohérentes ou pas ?
- Pour certains utilisateurs (internes) : accès aux zones commentaires

# 5.6. Un processus d'amélioration en continu

- L'amélioration d'un processus de production passe par la mise en place de métadonnées
- Mais, par ailleurs, il peut être dangereux de ne plus disposer de données « avant / après » sur les unités qui sont traitées de façon automatique → garder un échantillon de « petites » (au sens de l'impact de l'erreur) unités sur lesquelles, de façon plus ou moins rapprochée, on procède à une expertise manuelle ?

# 5.7. Les métadonnées relatives à la vérification des données

- Comme vu précédemment, il est nécessaire de garder des métadonnées, à la fois pour évaluer la qualité des statistiques produites, mais aussi pour améliorer les processus à venir
- Trois types d'indicateurs :
  - Indicateurs « budgétaires »
  - Indicateurs relatifs aux erreurs
  - Indicateurs relatifs au processus de vérification des données

# 5.7. Les métadonnées relatives à la vérification des données (suite)

- Exemples d'indicateurs « budgétaires » :
  - Nombre de « personnes jours » consacrés au travail de vérification manuelle
  - Durée, répartition dans le temps
  - Nombre de rappels d'unités posant problème

# 5.7. Les métadonnées relatives à la vérification des données (suite)

- Exemples d'indicateurs relatifs aux erreurs :
  - Taux de non réponses
  - Par variable : proportion d'unités pour lesquelles la valeur finale est différente de la valeur brute
    - Pour les variables numériques : distribution de cette différence (ou distribution avant et après imputation), part d'un agrégat provenant de données imputées
    - Pour les variables qualitatives : matrice de passage avant / après

# 5.7. Les métadonnées relatives à la vérification des données (suite)

- Exemples d'indicateurs relatifs au processus de vérification des données :
  - Pourcentage d'unités avec au moins un contrôle déclenché
  - Par variable : % d'unités traitées de façon automatique et de façon manuelle
  - Par contrôle : taux de déclenchement, pourcentage d'erreurs détectées (rapportées au nombre de données contrôlées)

# 5.7. Les métadonnées relatives à la vérification des données (suite)

- Pour plus de détails, voir par exemple le manuel EDIMBUS (2007)
- Nécessité de conserver des métadonnées, et en particulier de marquer les données imputées et de ne pas écraser les données brutes, pour pouvoir mener des études méthodologiques

# 5.8. La *work session* des Nations unies consacrée au *data editing*

- Organisée tous les 18 mois par l'UNECE
- Le site Kbase (voir bibliographie) rassemble l'ensemble des présentations
  - En particulier sur les logiciels

http://www1.unece.org/stat/platform/display/kbase/Software

# Bibliographie

# Bibliographie générale


- De Waal, T., Pannekoek, J., Scholtus, S. (2011). *Handbook of statistical data editing and imputation*, J. Wiley
- De Waal, T. (2008). *An overview of statistical data editing*, Discussion paper 08018, Statistics Netherlands
- Granquist, L. (1995), *Improving the traditional editing process,* in Business Survey Methods , John Wiley
- Granquist, L, Kovar, J. (1997), *Editing of survey data : how much is enough ?* in Survey Measurement and Process Quality, John Wiley

# Bibliographie générale


- Hoogland, J., van der Loo, M., Pannekoek, J., Scholtus, S. (2011). *Data editing, detection and correction of errors*, Statistical methods n°201110, Statistics Netherlands
- ISTAT, CBS and SFSO (2007). *Recommended practices for editing and imputation in cross-sectional business surveys,* manuel « EDIMBUS » mis au point dans le cadre d'un projet d'Eurostat
- *Documentation relative aux « work sessions » des Nations unies consacrées au data editing*, disponible à l'adresse :

http://www1.unece.org/stat/platform/display/kbase/UNECE+Work+Sessions+on+Statistical+Data+Editing

# Bibliographie (suite)


- Arbués I., Revilla, P., Salgado, D. (2013), *An optimization approach to selective editing*, Journal of Official Statistics Vol. 29, n°4
- Bankier, M., Lachance, M., Poirier, P. (2000), *2001 Canadian Census Minimum Change Donor Imputation Methodology*, working paper 17, UNECE work session on statistical data editing, Cardiff
- Brion, Ph. (2016), *Vérification sélective des données et qualité du prédicteur utilisé*, Communication au 9ème colloque francophone sur les sondages, Gatineau, Canada
- De Waal, T., Quéré, R. (2003), *A Fast and Simple Algorithm for Automatic Editing of Mixed Data*, Journal of Official Statistics Vol. 19, n°4
- Di Zio, M., Guarnera, U. (2013), *A contamination model for selective editing*, Journal of Official Statistics Vol. 29, n°4

# Bibliographie (suite)


- Fellegi, I.P., Holt, D. (1976), *A systematic approach to automatic edit and imputation*, Journal of the American Statistical Association, Vol. 71,  Number 253
- Farwell, K. (2005), *The general application of significance editing to economic collections*, Research paper, Methodological Advisory Committee, Australian Bureau of Statistics
- Gros, E. (2009), *Setting cut-off scores for selective editing in structural business statistics : an automatic procedure using simulation study,* UNECE work session on statistical data editing, Neuchâtel
- Gros, E. (2012), *Assessment and improvement of the selective editing process in Esane (French SBS)*, UNECE work session on statistical data editing, Oslo

# Bibliographie (suite)


- Hedlin, D. (2008), *Local and global score functions in selective editing*, working paper 31, UNECE work session on statistical data editing, Vienna
- Hesse, C. (2005), *Vérification sélective de données quantitatives,* document de travail de la direction des statistiques d'entreprises E2005/04, Insee
- Hidiroglou, M., Berthelot, J.M. (1986), *Contrôle statistique et imputation dans les enquêtes-entreprises périodiques*, Techniques d'enquête Vol. 12, n°1
- Latouche, M., Berthelot, J.M. (1992), *Use of a score function to prioritize and limit recontacts in editing business surveys*, Journal of Official Statistics Vol. 8, n°3

# Bibliographie (suite)


- Lawrence, D., McKenzie, R. (2000), *The general application of significance editing*, Journal of Official Statistics Vol. 16, n°3
- Merad, S., Wagstaff, H. (2005), *A management information system for controlling editing quality in a survey with multiple requirements*, UNECE Work Session on Statistical Data Editing, Ottawa
- Nordberg, A., & al. (2010), *A general methodology for selective editing*, Statistics Sweden
- Rivière, P. (1996), *Enquêtes annuelles d'entreprise : à la rencontre du 4ème type*, Courrier des statistiques n°78, Insee

# Bibliographie (suite)


- Scholtus, S. (2008), *Algorithms for correcting some obvious inconsistencies and rounding errors in business survey data*, Discussion paper 08015, Statistics Netherlands
- Scholtus, S. (2009) *Automatic correction of simple typing errors in numerical data with balance edits*, Discussion paper 09046, Statistics Netherlands
- Scholtus, S. (2013a), *Vérification automatique en présence de vérifications avec rejet et de vérifications avec avertissement*, Techniques d'enquête Vol. 39, N°1, Statistique Canada
- Scholtus, S. (2013b), *Imputation under edit constraints*, Part of the Memobust Handbook on methodology of business statistics, Eurostat