



Les valeurs influentes dans l'enquête TIC

Traitements et expérimentations en cours

Elvire Demoly
Responsable des enquêtes TIC entreprises

Insee
Mesurer pour comprendre

L’enquête annuelle TIC

Enquête sur les technologies de l’information et de la communication et le commerce électronique

Européenne, annuelle et obligatoire

Champ

Entreprises de 10 personnes occupées ou plus (UL, profilées)

Secteurs marchands,

hors unités agricoles, financières et d’assurance

L’enquête annuelle TIC

Objectif

Mieux connaître l’informatisation et la diffusion des TIC dans les entreprises

- Place des outils nouveaux dans les relations externes de l’entreprise (internet, commerce électronique)
- Usage de ces outils dans leur fonctionnement interne (réseaux, systèmes intégrés de gestion).

L’enquête annuelle TIC

Questionnaire Eurostat adapté pour la France

Traduction, questions optionnelles, questions nationales

Type de questions

- Majoritairement qualitatives

Votre entreprise a-t-elle un accès à internet ?

Votre entreprise envoie-t-elle des factures électroniques ?

Votre entreprise utilise-t-elle un progiciel de gestion intégré (PGI/ERP) ?

- Quelques quantitatives

Parmi les personnes employées dans l’entreprise, combien utilisent au moins une fois par semaine un ordinateur avec accès à internet ?

Quel a été en 2012 le montant du chiffre d’affaires (CA) hors taxes généré par des commandes reçues via un site web ?

Collecte

Collecte internet depuis 2012

Gérée au PEE (DR Midi-Pyrénées)

- réponse aux entreprises
- relances par tél
- apurements

Traitements post-collecte au Pise (DR Pays de la Loire)

- apurements
- correction de la non-réponse (totale et partielle)
- calage
- traitement des valeurs influentes

Ça gratte dans TIC : les valeurs influentes

Repérage jusqu'à présent de valeurs influentes :

Sur les variables de montants (ventes électroniques, achats électroniques, CA, achats totaux)

Pour les unités avec poids (après calage) > 2

Dont contribution au total > 1 %

Plan d'action

1. Prévention à l'étape du plan de sondage
   - Strates exhaustives
2. Prévention après tirage
   - Définition d'unités « non-substituables » avant collecte
3. Traitement quand le mal est fait
   - Quelques années de troncature de poids
   - Des essais de winsorisation « à la main »
   - Un traitement automatisé à l'avenir ?

Plan de sondage

Population cible : environ 190 000 unités

Échantillon : environ 12 500 unités

Plan de sondage stratifié

activité x effectif x chiffre d'affaires (CA) (CA à partir de 2012)

Allocation mixte

moyenne de deux allocations proportionnelles
(nb d'unités/nb de personnes occupées)
avec contraintes de précisions locales

Strates exhaustives

- 500 personnes occupées et plus
- Au-delà d'un seuil de CA

1er traitement

Plan de sondage

Secteurs d'activité

regroupements selon les agrégats demandés par Eurostat

Effectif en 5 tranches

10 à 19, 20 à 49, 50 à 249, 250 à 499, 500 et plus

Chiffre d'affaires

seuils déterminés selon la taille (à partir de TIC 2012)

Renouvellement par moitié

conservation de la moitié de l'échantillon aléatoire de l'année précédente

Pas tout à fait du tirage aléatoire simple sans remise…

Seuils exhaustivité

Les seuils de chiffre d'affaires (CA)

- 10 à 19 personnes occupées : CA >= 50 millions d'euros
- 20 à 49 personnes occupées : CA >= 200 millions d'euros
- 50 à 249 personnes occupées : CA >= 800 millions d'euros
- 250 à 499 personnes occupées : CA >= 1 500 millions d'euros
- 500 personnes occupées et plus : pas de seuil CA (strates exhaustives).

Principe : chiffre d'affaires à partir duquel une unité pondérée par le poids de sondage moyen de sa tranche d'effectif représente plus de 1 % du CA global de sa tranche d'effectif (estimation enquête TIC 2011).

Unités non-substituables

Repérage d'unités atypiques *a priori* dans l'échantillon

Critères : poids=1 et

importance de l'unité dans sa strate (CA, effectif)
ou commerce électronique important (données n-1)

Objectifs :

- empêcher ces unités de devenir influentes
- ne pas propager leur comportement à d'autres

→ exclure les non-substituables de certains traitements de la NR

2ème traitement

Collecte

Pour limiter la non-réponse

- Deux courriers de rappel
- Relances (tél, mail) ciblées sur des unités importantes
- En priorité, relance des non-substituables

Traitements post-enquête

Taux de réponse 2012 : 77 %

Pas tout à fait 100 %…

Apurement

- Suppression des réponses incohérentes
- Imputations de réponses selon règles déterministes

Redressements

- Correction de la non-réponse partielle par imputation
- Correction de la non-réponse totale par repondération
- Calage sur marges
- Traitements des valeurs influentes

3ème traitement

Correction de la non-réponse partielle

Non-réponse partielle aux questions qualitatives

- Corrigée par la méthode du donneur (hot deck)
- Unité donneuse ayant mêmes caractéristiques que la receveuse
- Non-substituables jamais donneuses

Non-réponse partielle aux questions quantitatives

- Source externe si disponible
- Imputation par le ratio par classe sinon
- Non-substituables exclues de ce traitement si données n-1 ou autres sources disponibles

Correction de la non-réponse totale

Non-réponse totale (NRT)

Repondération au sein de groupes de réponse homogène (GRH)

GRH constitué par modèle explicatif de la non-réponse

Choix des variables selon leur pouvoir explicatif parmi :

- la tranche de taille
- la tranche de chiffre d'affaires
- le secteur d'activité
- la localisation du siège (Paris/province/Dom)
- la catégorie juridique
- ...

Non-substituables et non-réponse totale

Traitement de la NRT $\Rightarrow$ hausse des poids

$\Rightarrow$ potentielles valeurs influentes

$\Rightarrow$ non-substituables **exclues** des traitements NRT

Objectif : éviter d'affecter à des unités importantes et atypiques un poids > 1

Pour elles, valeurs imputées pour chaque variable

(assimilation à de la non-réponse partielle)

Calage

Calage sur marges

Calage sur les strates de tirage

Exclusion des non-substituables

$\Rightarrow$ les unités non-substituables conservent un poids de 1

Non-substituables après corrections des NR

Après les traitements post-enquête, les non-substituables :

- ne représentent aucune autre unité
- ne sont représentées par aucune autre
- conservent un poids de 1 (répondantes, non-répondantes)

$\Rightarrow$ les unités non-substituables ne sont ni influentes ni génératrices d'influence

Et les valeurs influentes restantes ?

Après calage, repérage d'unités influentes (diapo 6).

Troncature des poids (jusqu'à l'édition 2011)

Traitement

poids de l'unité *i* ramené à 1

reste du poids ($w_i - 1$) réparti parmi les autres unités de la strate

Hypothèse

l'unité influente ne représente qu'elle-même

Problème

génère biais et erreur quadratique moyenne

Et les valeurs influentes restantes ?

Winsorisation de type 2 (édition 2012)

Ex : y = montant des ventes web

Pour une strate donnée, détermination d'une valeur seuil K jugée valeur maximale plausible… à dire d'expert.

Hum hum…

Si le montant y_i déclaré par l'unité i dépasse ce seuil ($y_i > K$) alors y_i est remplacé par :

$$y_i^{win} = \left[y_i + \left(w_i - 1 \right) K \right] / w_i$$

où w_i est le poids final de l'individu i, après calage

Et les valeurs influentes restantes ?

Winsorisation de type 2 (édition 2012) ***suite***

La valeur winsorisée pondérée pour l'unité *i* devient alors :

$$w_i . y_i^{win} = 1 \times y_i + \left(w_i - 1\right) \times K$$

Interprétation : la valeur déclarée par l'unité *i* est conservée pour un poids de 1, la valeur seuil *K* est comptée pour $\left(w_i - 1\right)$ individus.

Et les valeurs influentes restantes ?

Winsorisation de type 2 (édition 2012) ***suite***

A priori, mieux que la troncature des poids à 1, mais…

il reste des faiblesses :

- La détermination des valeurs seuils à dire d'expert
- Le repérage des valeurs influentes
- L'approche au cas par cas chronophage

Les estimateurs robustes à l'essai : le BHR

Estimation du chiffre d'affaires total (TIC 2011)

Calcul de l'estimateur robuste BHR

$$\hat{t}_R(K_{opt}) = \hat{t}_{HT} - \frac{1}{2}(\hat{B}_{\min} + \hat{B}_{\max})$$

sur l'ensemble de la population et sur chaque strate

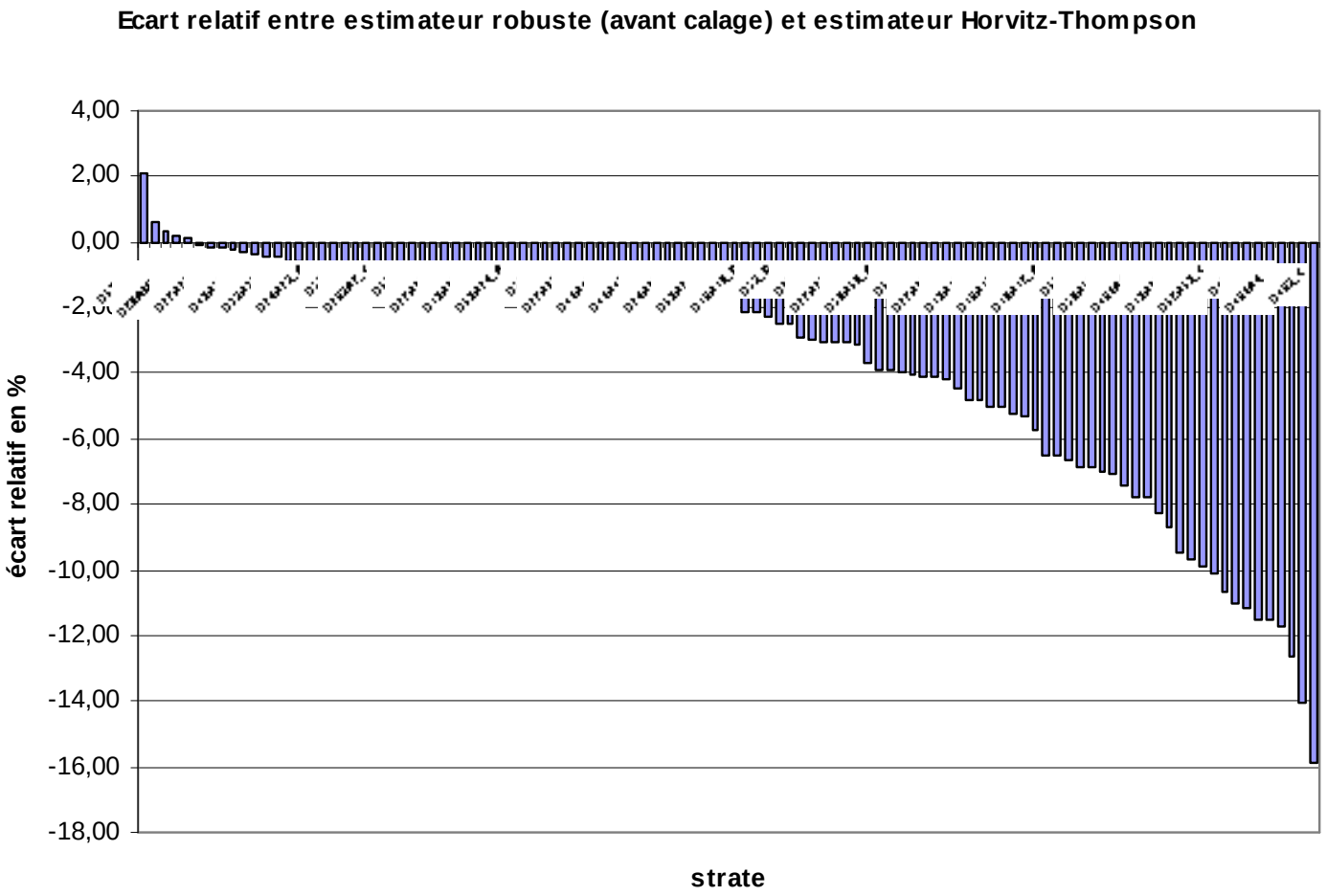
Conditions et hypothèses retenues

Hypothèse : absence de non-réponse (NR)

D'où :

- Utilisation des poids traités de la NR comme poids de sondage
- Utilisation des valeurs redressées de la non-réponse partielle comme valeurs déclarées

Estimateurs robustes BHR par strates (avant calage)



$$\text{écart relatif}_d(\hat{t}_{R,d}, \hat{t}_{HT,d}) = 100 \times (\hat{t}_{R,d} - \hat{t}_{HT,d}) \big/ \hat{t}_{HT,d} \ .$$

L’estimateur BHR global

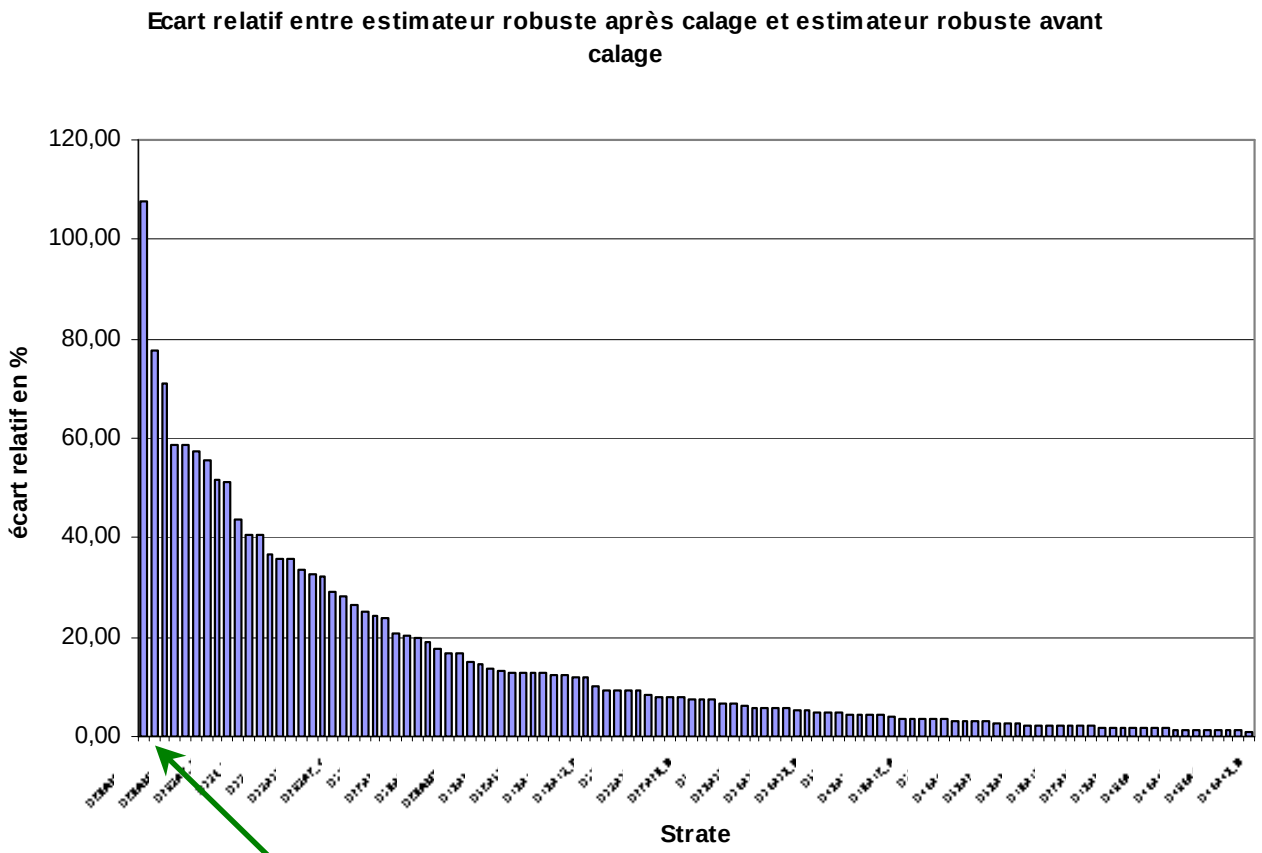
Résultats hors strates exhaustives

	Estimateur CA (Md €)
Estimateur type Horvitz-Thompson	1 466
Estimateur robuste global	1 444
Somme estimateurs robustes par strates	1 385

⇒ Utilisation du calage pour modifier les estimateurs par strates et imposer que leur somme soit égale à l’estimateur global

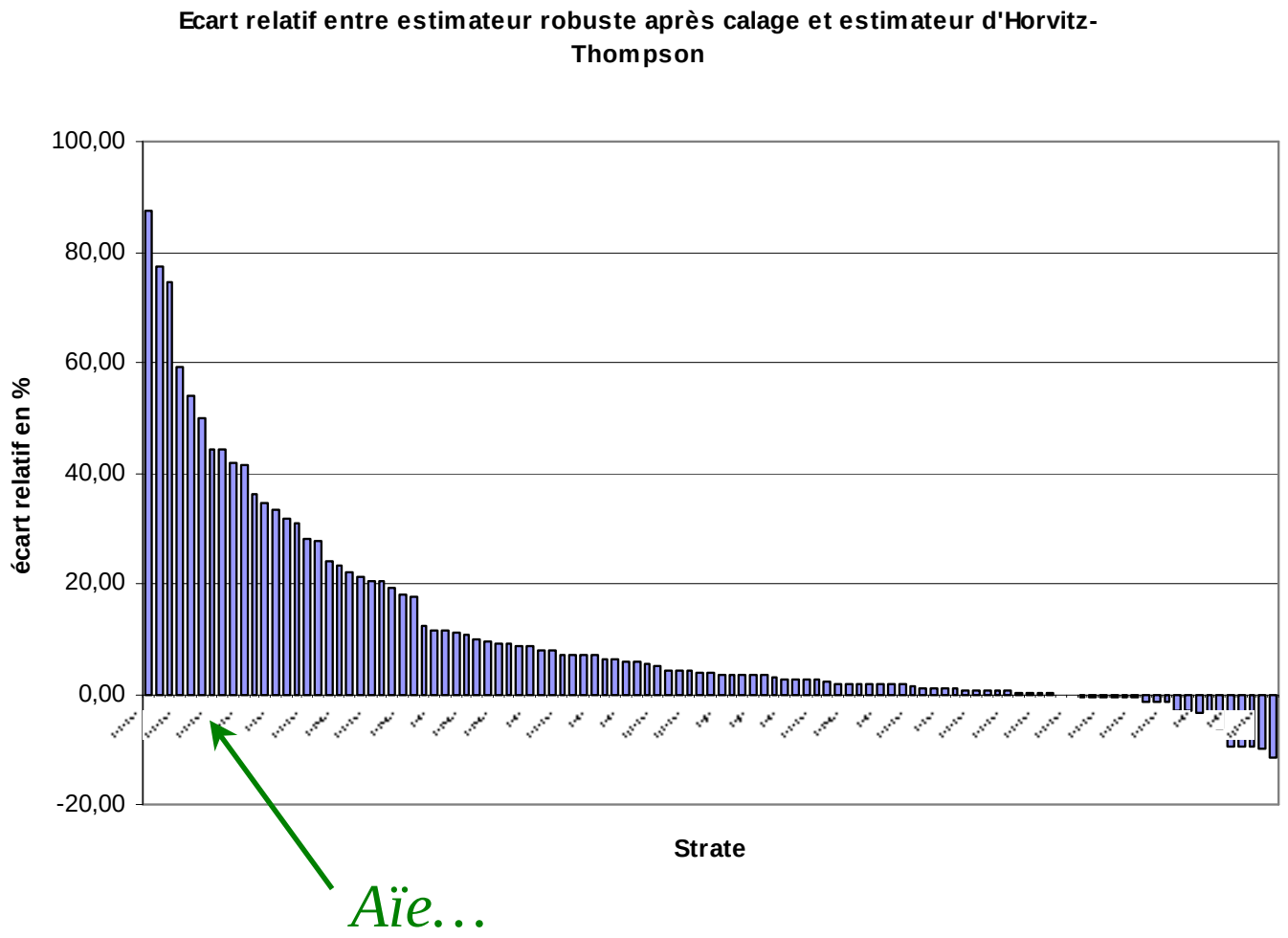
Premiers essais : calage par la méthode linéaire

Estimateurs BHR avant - après calage



Ouille…

Estimateurs BHR après calage vs HT



Autres méthodes de calage

D’autres méthodes de calage sont testées

Par exemple, méthode linéaire tronquée

Permet de limiter le rapport estimateur initial / calé

Les estimateurs BHR en pratique ?

Encore du travail… quelques pistes ?

Pour les estimateurs sur domaine, affiner le calage

Quel impact des hypothèses retenues ici ?
(non-réponse ignorée, assimilation à tirage aléatoire simple)

Le reste à faire…

… pour que la théorie se rapproche de la vraie vie

En pratique

- Nécessité d'un fichier résultats à *n* lignes (individus) et *p* colonnes (variables) avec une variable de pondération qui permet de calculer les estimateurs
- Tirages pas toujours aléatoires simples
- Présence de non-réponse

Les résultats des enquêtes TIC

Envois de résultats agrégés à Eurostat

Dernières publications Insee (insee.fr)

Insee Résultats n°64 Economie - mars 2013

http://www.insee.fr/fr/publications-et-services/irweb.asp?id=tic12

Insee Première N°1413 - septembre 2012

Remplir des formulaires administratifs en ligne, une pratique courante pour les sociétés

http://www.insee.fr/fr/themes/document.asp?ref_id=ip1413

Publications à venir : TIC-TPE, Insee Première et Insee Résultats (fin 2013 – début 2014)

Merci de votre attention

Contacts

Elvire DEMOLY - DSE
Tél. : 01 41 17 56 36
Courriel : elvire.demoly@insee.fr

Nicolas SIGLER - PISE
Tél. : 02 40 41 78 23
Courriel : nicolas.sigler@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr
Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00