Empile les face-à-face : une pièce à plusieurs actes

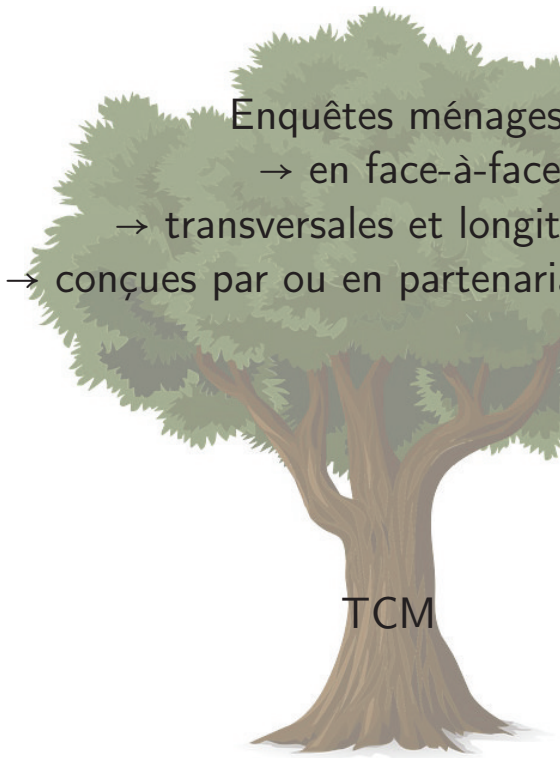
Céline Leroy

Séminaire interne du DMS

Mardi 3 juillet 2018

Insee Mesurer pour comprendre

Qu'est-ce que le Tronc Commun des Ménages (TCM) ?



Pourquoi empiler les bases de données du TCM ? (1/2)

- Le TCM permet de capter des configurations familiales rares : familles monoparentales, enfants en garde alternée, individus multi-résidents, etc.
- Intérêt de certains chercheurs de l'Ined et de la division Enquêtes et Études Démographiques de l'Insee pour ce champ
- Mais configurations familiales trop rares pour être étudiées à l'aide d'une seule enquête

Pourquoi empiler les bases de données du TCM ? (2/2)

Solution : empiler les bases de données du TCM de plusieurs enquêtes pour augmenter la taille de l'échantillon et obtenir des estimations fiables sur ces situations familiales rares

⇒ Création de la base des TCM empilés à partir des tables TCM issues de 29 enquêtes entre 2006 et 2015 : convention signée entre l'Insee et l'Ined pour la création et la mise à disposition de cette nouvelle source en 2018

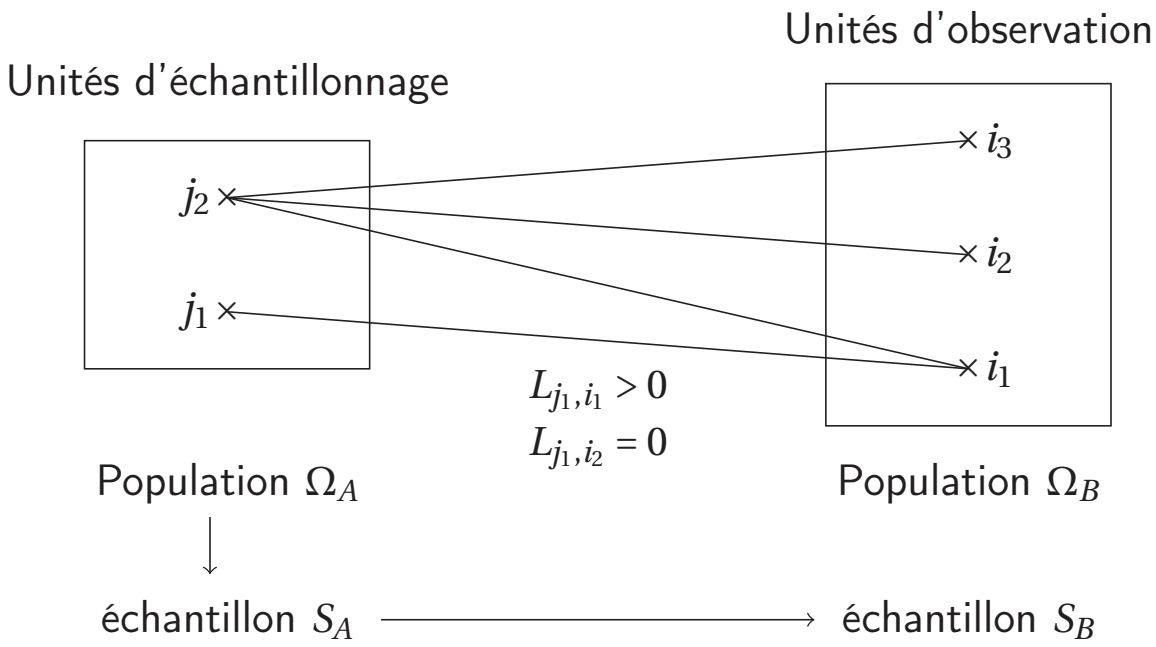
De quelles enquêtes sont issues les tables TCM à empiler ?

- CVS 2006 à 2015 (Cadre de Vie et Sécurité)
- SRCV 2006 à 2015 (Statistiques sur les Ressources et Conditions de Vie)
- Logement 2006 et 2013
- Transports et Déplacements 2007
- Handicap-Santé Volet Ménages 2008 (HSM)
- Patrimoine 2009 et 2014
- Budget de Famille 2011 (BDF)
- Conditions de Travail 2012 (CdT)
- AES 2012 (Adult Education Survey)

Sommaire

Sommaire

Liens entre unités d'échantillonnage et d'observation



Un nouvel estimateur sans biais issu du partage des poids (1/2)

- Total à estimer : $Y = \sum_{i \in \Omega_B} Y_i$
- Problème avec l'estimateur de Horvitz-Thompson car difficile d'obtenir les probabilités d'inclusion des unités i en pratique
- Méthode du partage des poids pour contourner ce problème

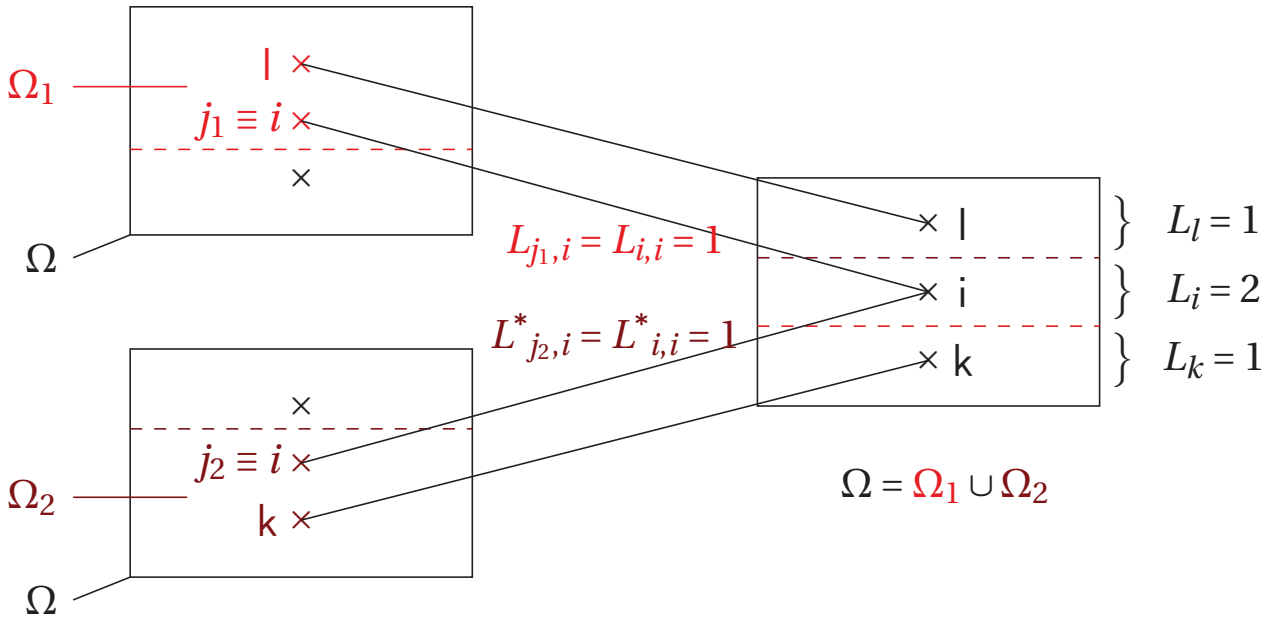
Un nouvel estimateur sans biais issu du partage des poids (2/2)

- Y estimé sans biais par $\hat{Y} = \sum_{i \in S_B} W_i . Y_i$

$$W_i = \sum_{j \in S_A} \theta_j . \frac{L_{j,i}}{L_i} = \text{poids issu du partage des poids}$$

$\rightarrow \theta_j$ poids de sondage de j lié au tirage de S_A dans Ω_A
$\rightarrow L_i = \sum_{j \in \Omega_A} L_{j,i}$ = nombre total de liens que l'unité i peut avoir avec les unités j de Ω_A

Des unités d'échantillonnage tirées dans plusieurs bases



Estimateur issu du partage des poids

- Poids issu du partage des poids

$$W_i = \frac{1}{L_i}\left(\sum_{j_1 \in S_1} \theta_{j_1}.L_{j_1,i} + \sum_{j_2 \in S_2} \theta^*_{j_2}.L^*_{j_2,i}\right)$$

$\rightarrow L_i = \sum_{j_1 \in \Omega_1} L_{j_1,i} + \sum_{j_2 \in \Omega_2} L^*_{j_2,i}$ = nombre de liens qu'une unité d'échantillonnage i possède avec l'ensemble des bases
$\rightarrow \theta_{j_1}$ et $\theta^*_{j_2}$ poids de tirage des unités j_1 et j_2 dans les échantillons S_1 et S_2

Estimateur issu du partage des poids

- Poids issu du partage des poids

$$W_i = \frac{1}{L_i}\left(\sum_{j_1 \in S_1} \theta_{j_1}.L_{j_1,i} + \sum_{j_2 \in S_2} \theta^*_{j_2}.L^*_{j_2,i}\right)$$

$\rightarrow L_i = \sum_{j_1 \in \Omega_1} L_{j_1,i} + \sum_{j_2 \in \Omega_2} L^*_{j_2,i}$ = nombre de liens qu'une unité d'échantillonnage i possède avec l'ensemble des bases
$\rightarrow \theta_{j_1}$ et $\theta^*_{j_2}$ poids de tirage des unités j_1 et j_2 dans les échantillons S_1 et S_2

- Mais unités d'échantillonnage = unités d'observation

$$\Rightarrow W_i = \frac{1}{L_i}\left(\theta_i.L_{i,i}.\mathbb{1}_{i \in S_1} + \theta^*_i.L^*_{i,i}.\mathbb{1}_{i \in S_2}\right) \qquad (1)$$

Détermination du système de liens optimal (1/2)

- On considère $\Omega_1 = \Omega_2 = \Omega$
- Soient S_1 de taille n_1 et S_2 de taille n_2, deux échantillons de Ω et $S = S_1 \cup S_2$
- Soient $\hat{Y}_1$ et $\hat{Y}_2$, deux estimateurs de Horvitz-Thompson de Y sur S_1 et S_2

Détermination du système de liens optimal (1/2)

- On considère $\Omega_1 = \Omega_2 = \Omega$
- Soient S_1 de taille n_1 et S_2 de taille n_2, deux échantillons de Ω et $S = S_1 \cup S_2$
- Soient $\hat{Y}_1$ et $\hat{Y}_2$, deux estimateurs de Horvitz-Thompson de Y sur S_1 et S_2
- $\hat{Y}_{opti} = \alpha.\hat{Y}_1 + \beta.\hat{Y}_2$ (avec $\alpha + \beta = 1$) $= \sum_{i \in S} W_i^{opti}.Y_i$

$$\Rightarrow W_i^{opti} = \alpha.\theta_i.\mathbb{1}_{i \in S_1} + \beta.\theta^*{}_i.\mathbb{1}_{i \in S_2} \qquad (2)$$

Détermination du système de liens optimal (2/2)

- Résolution du programme d'optimisation sous contrainte :

$$\min_{(\alpha,\beta)} \mathbb{V}\left[\hat{Y}_{opti}\right] \text{ sous la contrainte } \alpha + \beta = 1$$

$$\Rightarrow \alpha = \frac{n_1}{n_1+n_2} \text{ et } \beta = \frac{n_2}{n_1+n_2}$$

Détermination du système de liens optimal (2/2)

- Résolution du programme d'optimisation sous contrainte :

$$\min_{(\alpha,\beta)} \mathbb{V}\left[\hat{Y}_{opti}\right] \text{ sous la contrainte } \alpha+\beta=1$$

$$\Rightarrow \alpha=\frac{n_1}{n_1+n_2} \text{ et } \beta=\frac{n_2}{n_1+n_2}$$

- $W_i = W_i^{opti}$

$$\Leftrightarrow \frac{1}{L_i}\left(\theta_i.L_{i,i}.\mathbb{1}_{i\in S_1}+\theta^*{}_i.L^*{}_{i,i}.\mathbb{1}_{i\in S_2}\right)=\alpha.\theta_i.\mathbb{1}_{i\in S_1}+\beta.\theta^*{}_i.\mathbb{1}_{i\in S_2}$$

$$\Leftrightarrow \frac{L_{i,i}}{L_i}=\frac{n_1}{n_1+n_2} \text{ et } \frac{L^*{}_{i,i}}{L_i}=\frac{n_2}{n_1+n_2}$$

$\Rightarrow$ Système de liens optimal défini par :

$$L_{i,i}=n_1, L^*{}_{i,i}=n_2 \text{ et } L_i=n_1+n_2=n$$

Sommaire

Application de l'approche optimale du partage des poids (1/2)

- Empilement annuel = empilement des tables TCM issues des enquêtes d'une même année
- Jeu de poids initial pour les tables TCM de chaque enquête : poids niveau logement, corrigés de la non-réponse et calés, fournis par chaque responsable d'enquête
- Objectif : déterminer un nouveau jeu de poids adapté à l'empilement des différents échantillons

Application de l'approche optimale du partage des poids (2/2)

- Bases de sondage utilisées pour les enquêtes à empiler :
  → Échantillon-maître du recensement
  → Taxe d'Habitation
- Une année donnée, les échantillons à empiler peuvent donc provenir de bases de sondage différentes mais ils extrapolent sur la même population : celle des logements ordinaires en résidence principale.

  ⇒ L'approche optimale du partage des poids peut être utilisée pour l'empilement des tables TCM de niveau logement

Expression finale du poids de niveau logement (1/2)

- Soit S^{lgt}_{annee} l'échantillon annuel empilant les échantillons des différentes enquêtes d'une même année : $S^{lgt}_{annee} = \cup S^{lgt}_{enq}$.
- Système de liens défini de cette façon :
  - Pour un logement k tiré dans l'enquête *enq*, $L^{enq}_{k,k} = n_{enq}$ = taille de l'échantillon de répondants à l'enquête *enq*
  - $L_k = n$ = taille totale de l'échantillon de répondants S^{lgt}_{annee} issu de l'empilement = $\sum_{enq} n_{enq}$
- En appliquant la méthode optimale du partage des poids, nous obtenons :

$$\forall k \in S^{lgt}_{annee} : \quad w^{pp}_k = \sum_{enq} \left(\frac{n_{enq}}{n} \times w\_init^{enq}_k \right)$$

Expression finale du poids de niveau logement (2/2)

Pour les tirages d'échantillons des enquêtes ménages de l'Insee, si un logement k est tiré pour une enquête une année donnée, il ne peut pas être tiré pour une autre enquête cette même année, on obtient donc l'expression finale :

$$\forall k \in S^{lgt,enq}_{annee} : \quad w^{pp}_k = \frac{n_{enq}}{n} \times w\_init^{enq}_k$$

Exemple d'application sur la base TCM 2014 (1/2)

- Trois enquêtes ont embarqué le TCM en 2014 donc trois échantillons de logements :
  - Cadre de Vie et Sécurité : $n_{CVS} = 16\ 372$
  - Statistiques sur les Ressources et Conditions de Vie (SRCV) : $n_{SRCV} = 11\ 370$
  - Patrimoine : $n_{Pat} = 11\ 628$
- Empilement des trois tables TCM $\Rightarrow$ un seul échantillon de taille $n = 39\ 370$

Exemple d'application sur la base TCM 2014 (2/2)

À l'issue du partage des poids :

$$w_k^{CVS} = \frac{n_{CVS}}{n} \times w\_init_k^{CVS} = \frac{16\ 372}{39\ 370} \times w\_init_k^{CVS} \simeq 0.42 w\_init_k^{CVS}$$

$$w_k^{SRCV} = \frac{n_{SRCV}}{n} \times w\_init_k^{SRCV} = \frac{11\ 370}{39\ 370} \times w\_init_k^{CVS} \simeq 0.29 w\_init_k^{CVS}$$

$$w_k^{Pat} = \frac{n_{Pat}}{n} \times w\_init_k^{Pat} = \frac{11\ 628}{39\ 370} \times w\_init_k^{CVS} \simeq 0.29 w\_init_k^{CVS}$$

Calage niveau logement à partir des marges EEC

Variables auxiliaires utilisées pour le calage niveau logement :

- Zone géographique
- Type de ménage
- Tranche d'âge de la personne de référence du ménage
- Catégorie socio-professionnelle de la personne de référence du ménage
- Niveau de diplôme de la personne de référence du ménage

⇒ On obtient les poids logements finaux après partage des poids et calage

Remarque : les poids finaux au niveau ménage correspondent aux poids logements finaux : s'il y a plusieurs ménages dans un même logement, chaque ménage aura le poids de son logement.

Application du partage des poids dans le cas classique du sondage indirect (1/2)

- Tous les habitants du logement sont recensés et interrogés dans le TCM
  ⇒ Tirage par grappe : cas particulier d'un sondage indirect où toutes les unités secondaires de la même unité primaire sont interrogées
- Dans un premier temps, chaque habitant a le poids final de son logement
- Mais pour chaque enquête à empiler sur la période 2006-2015, 4 à 7 % des répondants sont multi-résidents
  ⇒ Certains habitants sont donc susceptibles d'avoir plusieurs logements dans le champ de l'enquête

Application du partage des poids dans le cas classique du sondage indirect (2/2)

La multi-résidence peut être prise en compte en appliquant le principe général de la méthode du partage des poids

- Unités d'échantillonnage = logements
- Unités d'observation = habitants
- Un multi-résident peut être atteint par le biais de plusieurs unités d'échantillonnage

⇒ Un système de liens peut donc être défini entre les logements j et les individus i

Expression finale du poids de niveau individu (1/3)

- Système de liens défini de cette façon :
  - $L_{j,i} = 1$ si l'individu i peut être interrogé par le biais du logement j et 0 sinon
  - L_i = nombre total de liens qu'un individu i a avec la population des logements = nombre de logements ordinaires dans lesquels l'individu i vit habituellement
- Soit S^{lgt}_{annee} l'échantillon de logements de la table TCM_LOG_annee
- Soit S^{ind}_{annee} l'échantillon d'individus de la table TCM_IND_annee

Expression finale du poids de niveau individu (2/3)

- En appliquant la méthode générale du partage des poids, nous obtenons :

$$\forall\, i \in S_{annee}^{ind} : \quad w_i^{ind,pp} = \sum_{k \in S_{annee}^{lgt}} L_{k,i} \times \frac{w_k^{lgt}}{Nb\_lgt_i}$$

- Mais en pratique, si un individu i est présent plusieurs fois dans la table TCM_IND_annee par le biais de ses différents logements, il n'est pas identifiable en tant que doublon car il aura des identifiants différents
- Les poids de ses autres logements qui pourraient être dans l'échantillon S_{annee}^{lgt} ne sont donc pas connus à travers une seule observation

Expression finale du poids de niveau individu (3/3)

- On redéfinit le poids niveau individu de cette façon :

$$\forall\, i \in \text{logement j de } S_{annee}^{lgt} : \quad w_{i,j}^{ind,pp} = \frac{w_j^{lgt}}{Nb\_lgt_i}$$

Remarque : le cadre de la méthode du partage des poids est quand même encore respecté car si un individu i a deux logements j et k tirés dans l'échantillon, il aura deux poids associés à ses deux observations. Et en sommant ces deux poids, on retrouve bien le poids issu du partage des poids.

Expression finale du poids de niveau individu (3/3)

- On redéfinit le poids niveau individu de cette façon :

$$\forall\, i \in \text{logement j de } S^{lgt}_{annee} : \quad w^{ind,pp}_{i,j} = \frac{w^{lgt}_{j}}{Nb\_lgt_i}$$

Remarque : le cadre de la méthode du partage des poids est quand même encore respecté car si un individu i a deux logements j et k tirés dans l'échantillon, il aura deux poids associés à ses deux observations. Et en sommant ces deux poids, on retrouve bien le poids issu du partage des poids.

- Après le partage des poids lié à la multi-résidence, calage de niveau individu à partir des marges EEC, sur le croisement sexe × tranche d'âge

Estimation périodique : moyenne des estimations annuelles (1/2)

- Période 2006-2015 : empilement des 10 bases annuelles des TCM empilés pour chacun des trois niveaux (logement, ménage, individu)
- Exploitations prévues à partir de la base périodique des TCM empilés : études des phénomènes en moyenne sur l'ensemble de la période

  ⇒ Une estimation sur la période est donc équivalente à la moyenne des dix estimations annuelles

Estimation périodique : moyenne des estimations annuelles (2/2)

- Soit Y un total niveau logement ou individu

$$\hat{Y}_{06-15} = \sum_{k \in S_{06-15}} w_k^{periode} \times y_k = \frac{1}{10} \sum_{j=06}^{15} \hat{Y}_j$$

$$= \frac{1}{10} \sum_{j=06}^{15} \sum_{k \in S_j} w_k^{annee} \times y_k = \sum_{k \in S_{06-15}} \frac{1}{10} w_k^{annee} \times y_k$$

$$\Rightarrow \forall k \in S_{06-15} : \quad w_k^{periode} = \frac{w_k^{annee}}{nombre\,d'années\,empilées} = \frac{w_k^{annee}}{10}$$

Les effectifs des tables TCM périodiques finales

- Table TCM niveau logement : 408 772 logements pondérés
- Table TCM niveau ménage : 411 645 ménages pondérés
- Table TCM niveau individu : 987 163 individus pondérés

Sommaire

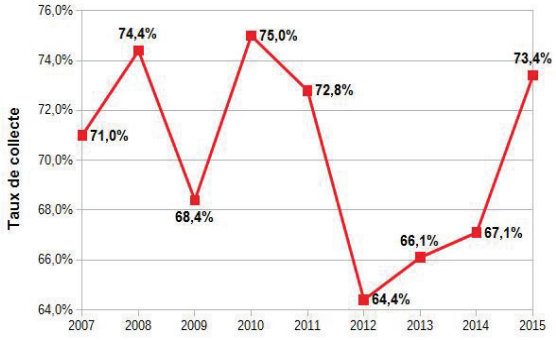
Définition des différents indicateurs (1/2)

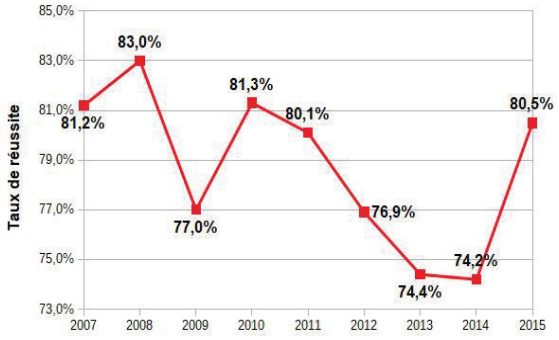
- Taux de collecte = $\frac{\text{Total FA réussies}}{\text{Total FA}}$
- Taux de réussite = $\frac{\text{Total FA réussies}}{\text{Total FA - Total hors champ - THV - THP - SAC}}$
- Taux de déchets = $\frac{\text{Total déchets}}{\text{Total FA}}$
- Taux de hors champ = $\frac{\text{Total hors champ}}{\text{Total FA}}$

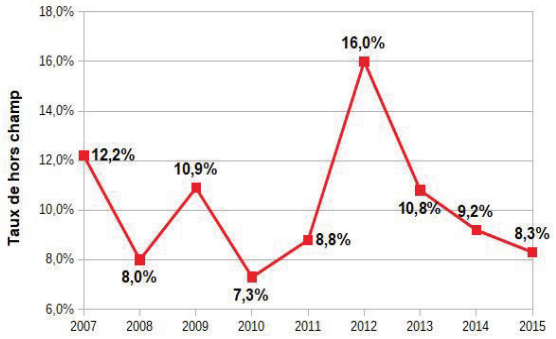
Définition des différents indicateurs (2/2)

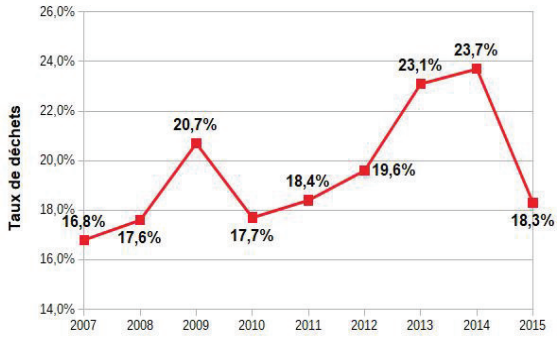
- Déchets = impossible à joindre, absence de longue durée, logement inconnu ou détruit, évitement des enquêtés, refus des enquêtés, entretien impossible car la personne n'est pas habilitée à répondre, fiche-adresse qui n'a pas pu être traitée avant la fin de la collecte
- THV = Tableau des Habitants du Logement (THL) complètement rempli mais l'entretien s'arrête car le ménage n'est pas dans le champ de l'enquête
- THP = abandon de l'enquête en cours de THL
- SAC = pour les enquêtes longitudinales, ménage parti sans laisser d'adresse (Sans Adresse Connue)

Évolution annuelle des différents taux à partir des bases annuelles des TCM empilés, de 2007 à 2015

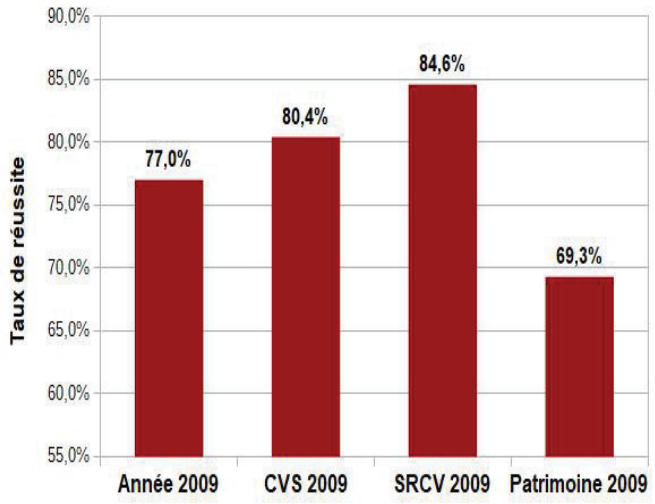


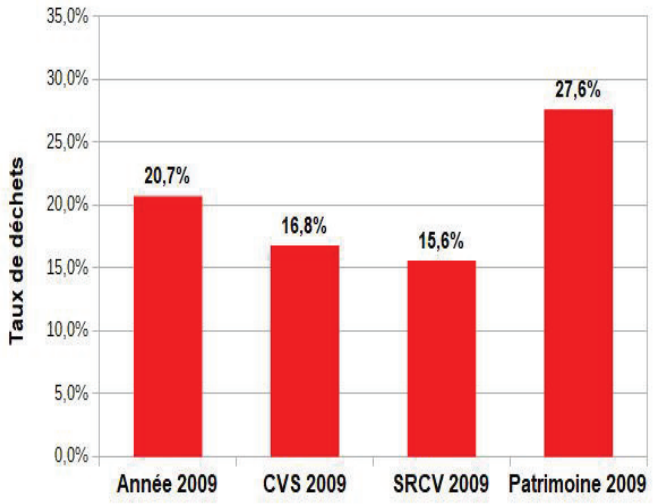






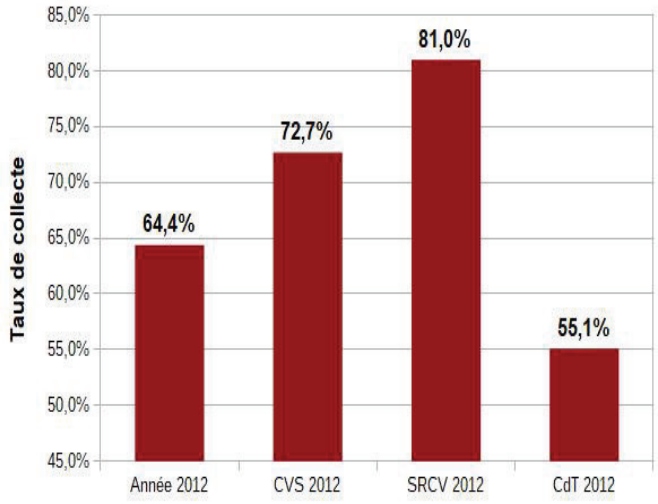
Effet enquête Patrimoine 2009 : baisse du taux de réussite et hausse des déchets

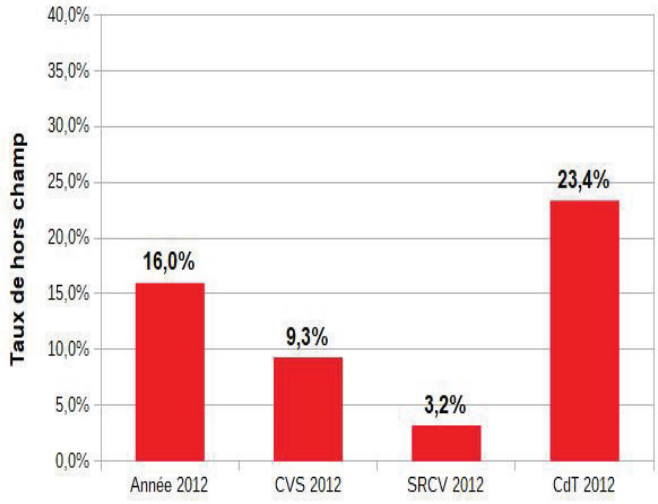




Année 2009 : 63 322 FA ; CVS 2009 : 25 701 FA ; SRCV 2009 : 13 332 FA ; Patrimoine 2009 : 24 289 FA

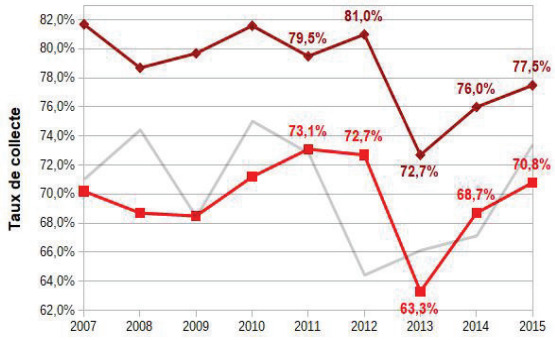
Effet enquête CdT 2012 : pic du taux de hors champ et baisse du taux de collecte

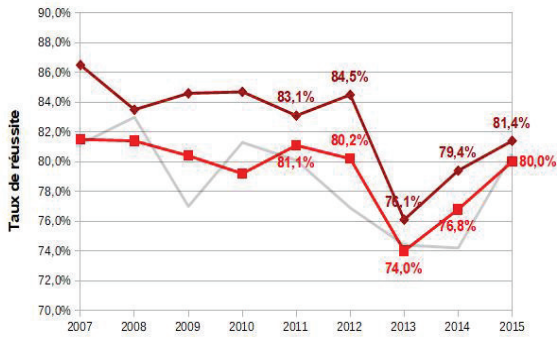


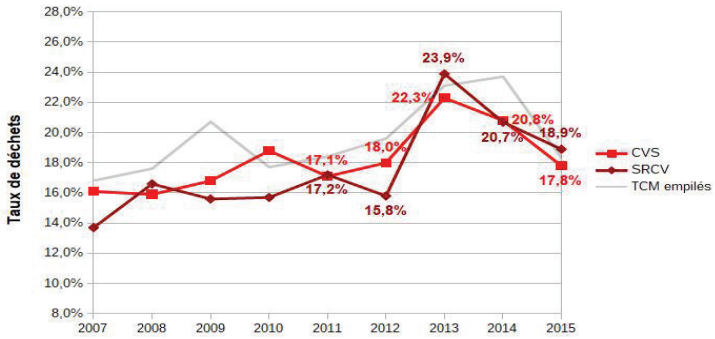


Année 2012 : 86 789 FA ; CVS 2012 : 24 094 FA ; SRCV 2012 : 14 869 FA ; CdT 2012 : 47 826 FA

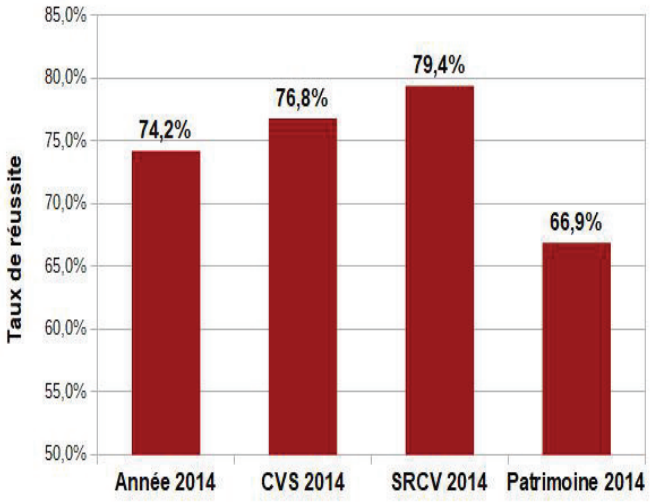
Effet NCEE 2013 : une rupture encore plus nette sur le champ des enquêtes CVS et SRCV

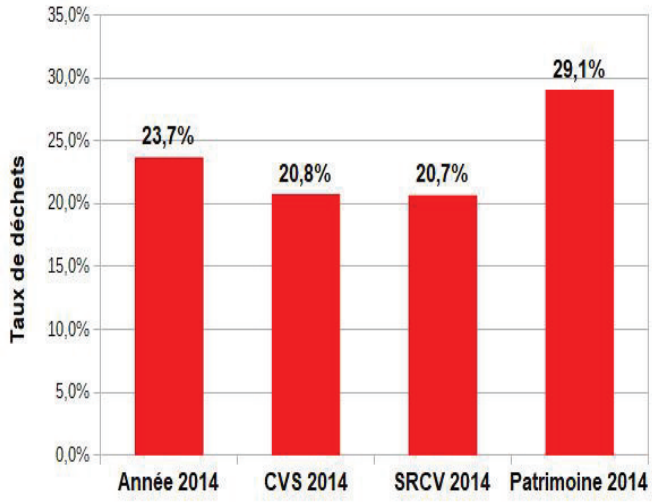






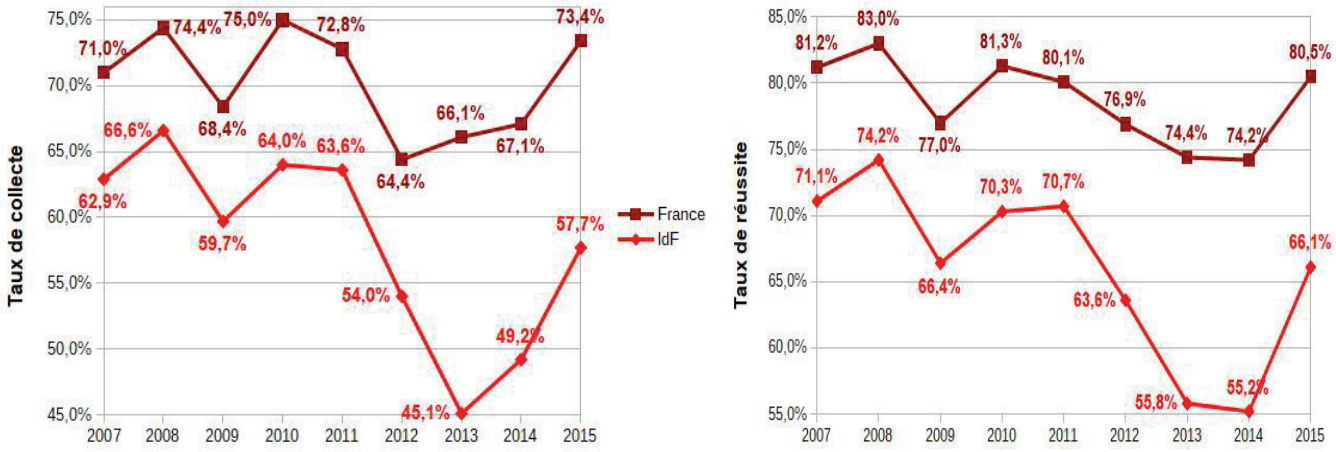
Effet enquête Patrimoine 2014 : stabilité du taux de réussite et des déchets après la rupture liée aux NCEE





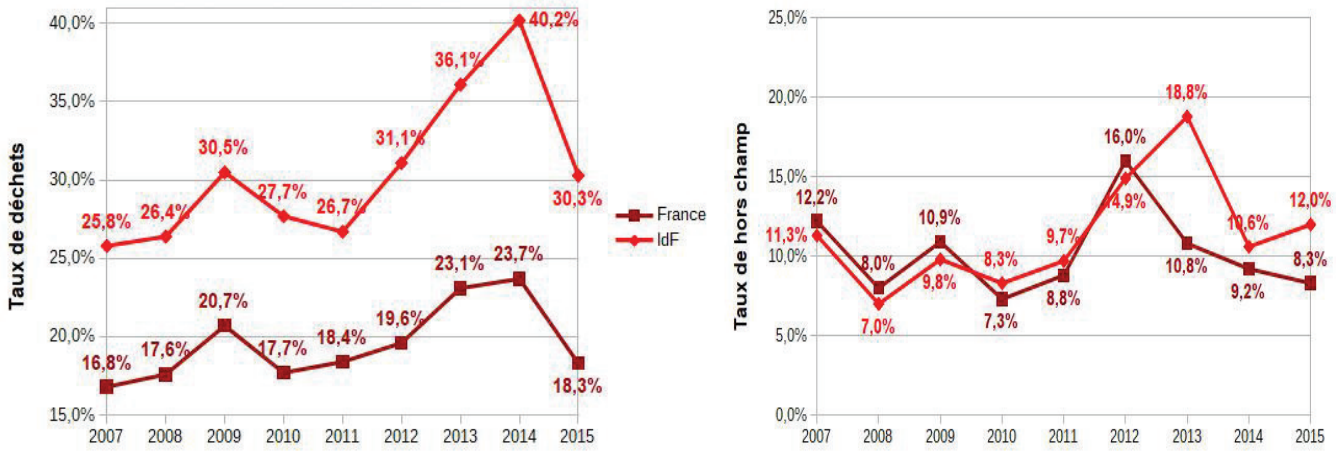
Année 2014 : 59 610 FA ; CVS 2014 : 23 969 FA ; SRCV 2014 : 15 037 FA ; Patrimoine 2014 : 20 604 FA

L'effet NCEE 2013 très marqué en Ile-de-France (1/2)



Champ : Bases annuelles des TCM empilés, de 2007 à 2015, France entière vs. Ile-de-France

L'effet NCEE 2013 très marqué en Ile-de-France (2/2)



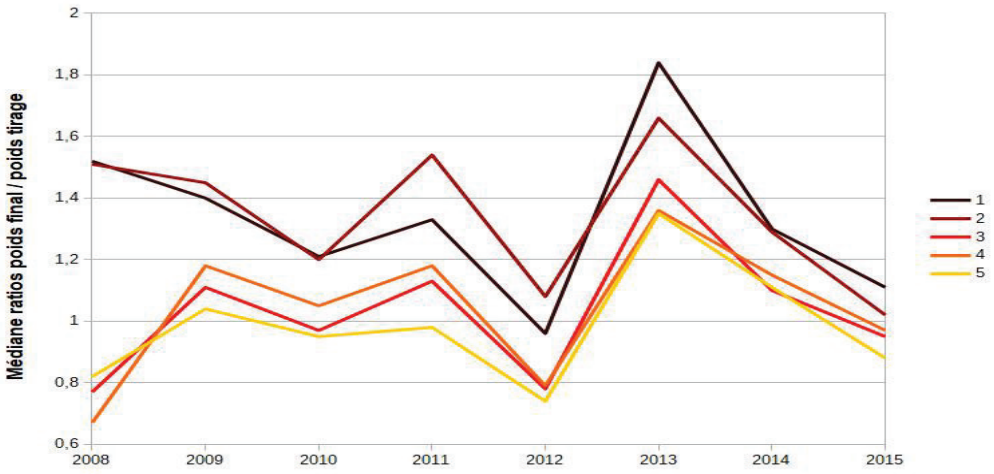
Champ : Bases annuelles des TCM empilés, de 2007 à 2015, France entière vs. Ile-de-France

Méthode pour caractériser les non-répondants

- Champ : bases ménages annuelles des TCM empilés, de 2008 à 2015, restreintes aux observations pondérées, donc aux ménages répondants
- Ratios $\frac{\text{poids final}}{\text{poids de tirage}}$ pour caractériser les ménages non-répondants : les profils non-répondants sont ceux dont le ratio est le plus élevé
- Caractérisation des non-répondants par type de ménage :
  1. Personne seule active
  2. Personne seule inactive
  3. Ménage de plusieurs personnes avec un seul actif
  4. Ménage de plusieurs personnes avec au moins deux actifs
  5. Ménage de plusieurs personnes sans actif

Caractérisation des non-répondants par type de ménage

Médiane des ratios par type de ménage et par année



→ Ménages de plusieurs personnes qui sont davantage susceptibles de répondre

→ Des écarts entre type de ménage qui semblent s'amenuiser depuis 2014

Les perspectives à venir sur le projet des TCM empilés

- Approfondir l'étude sur la caractérisation des non-répondants :
  - Affiner la caractérisation en croisant le type de ménage avec l'âge de la personne de référence et/ou sa situation principale vis-à-vis du travail (en emploi, étudiant, chômeur, etc.)
  - Faire du pseudo-panel sur la période 2006-2015
- Livraison des bases annuelles des TCM empilés de 2006 à 2015 et de la base périodique, sur le réseau Quételet avant fin 2018
- Exploitation de la source par des chercheurs de l'Ined, dans le cadre du projet Big Stat

Merci pour votre attention !

Insee Mesurer pour comprendre