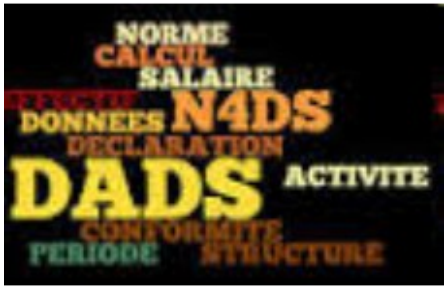


L'échantillon démographique permanent a 50 ans : retours sur un dispositif statistique original.

Sébastien Durier

Insee – Division Enquêtes et études démographiques

Motivations

- L'échantillon démographique permanent (EDP) a 50 ans cette année.
- L'EDP n'a jamais eu de présentation aux JMS !
- Objectif : rassembler dans un document pour les utilisateurs les aspects méthodologiques essentiels de l'EDP.

L'EDP en bref (1)

L'EDP est un panel d'individus ...

... sélectionnés (échantillonnés) par leur jour de naissance dans l'année.

=> appartiennent potentiellement à l'EDP tous les individus nés un des 16 « jours EDP » :

- du 1er au 4 octobre (depuis le lancement de l'EDP en 1968)

- du 2 au 5 janvier et du 1er au 4 des mois d'avril et juillet (depuis les années 2000)

=> quels que soient le lieu et l'année de naissance

L'EDP en bref (2)

... et qui compile pour ces individus les informations issues de cinq sources de données statistiques préexistantes (à l'Insee).

=> les recensements généraux de la population (RP) de 1968 à 1999 et depuis 2004 les Enquêtes Annuelles de Recensement (EAR)

=> les bulletins d'état-civil depuis 1968

=> le fichier électoral (depuis 1990)

=> le panel « tous salariés » depuis 1967

=> les données socio-fiscales (Fidéli, FiLoSoFi) depuis 2011

Plan de la présentation

- Une méthode d'échantillonnage originale
- Un élément clé de la qualité dans l'EDP : l'identification des individus dans les sources
- Intérêt de l'EDP : croiser des sources

  => le cas du panel dans les EAR

Méthode tirage de l'échantillon EDP

- Formellement un sondage en grappe ...

  => un jour de naissance = une grappe

- … avec choix raisonné pour la sélection des grappes (16 jours fixes au début des trimestres)

  => motivations pratiques pour la « collecte » qui jusque aux débuts des années 90 était manuelle

- … en pratique assimil(able/é) à un sondage aléatoire simple

  => taux de sondage 4/365,25 (1,1%) ou 16/365,25 (4,4%)

Les « petits » défauts de l'échantillon

- Âge moyen des EDP différent de leur génération

  => plus jeunes (4J), plus vieux (16J)

- Individus nés à l'étranger un peu sur-échantillonnés

  => 1er jour du mois très particulier

- Répartition uniforme des naissances dans l'année ?

  => planification des naissances (Régnier-Loilier, 2010)

  => samedi/dimanches (Régnier-Loilier, 2010)

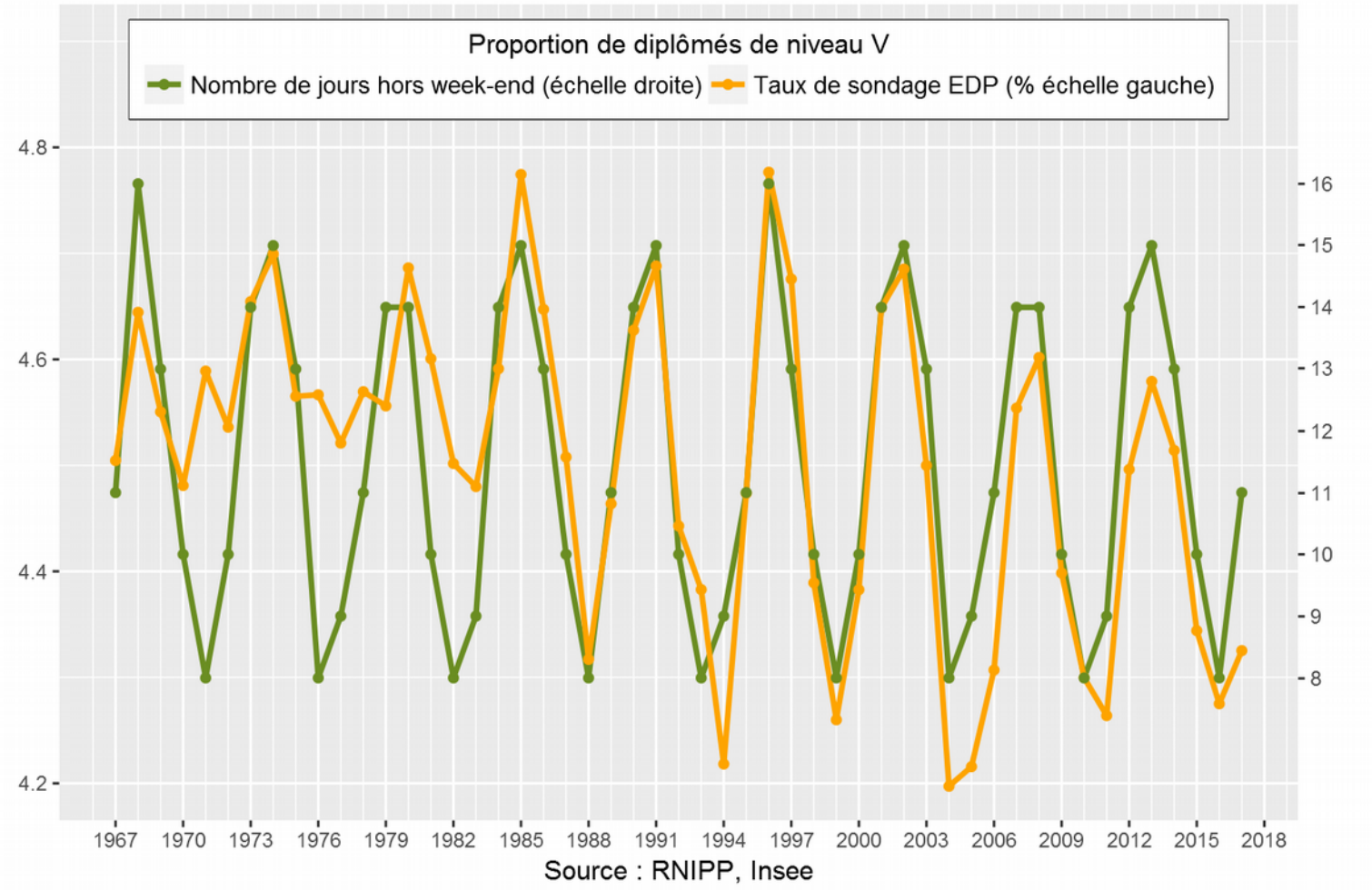
- Influence du jour de naissance sur la vie future ?

  => mortalité (Doblhammer, 2002)

  => niveau de diplôme (Grenet, 2010)

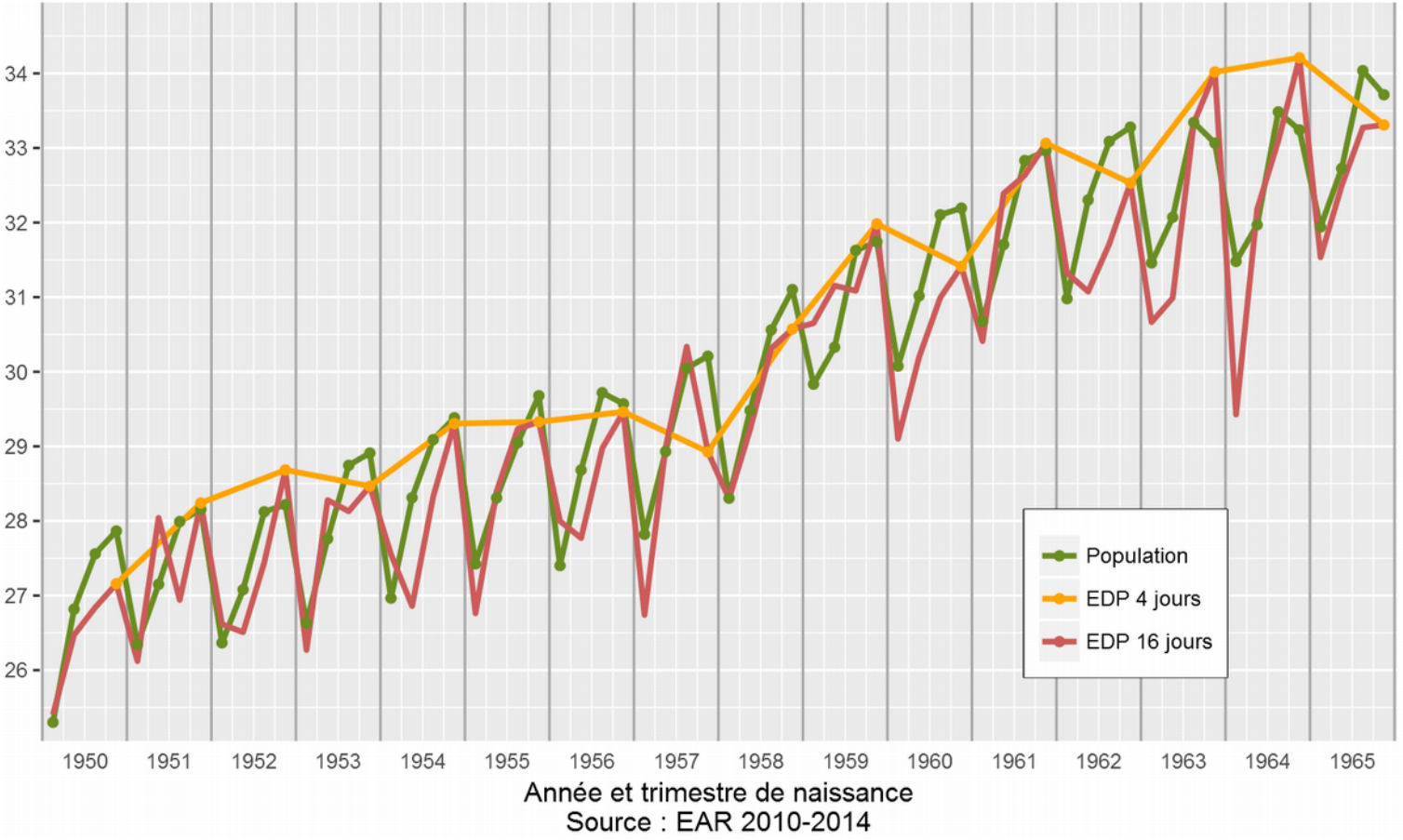
Effet samedi/dimanche

Corrélation entre le taux de sondage EDP et le nombre de jours EDP hors week-end



Effet d'âge relatif : le diplôme

Les individus nés aux 3ème et 4ème trimestres sont plus souvent diplômés de niveau V (CAP/BEP)



Identifier les individus EDP (1)

- Objectif de la « collecte » EDP :

  Retrouver TOUS les individus EDP des sources mobilisées en limitant les erreurs (« fausses trajectoires »)

- Méthode :

  Identifier les individus, qui déclarent être né un jour de naissance EDP, dans le répertoire EDP (≈ RNIPP) au moyen de leurs nom, prénoms et lieu de naissance

- Problèmes :

  Non-réponse totale ou partielle, erreur de déclaration ou de saisie des informations nominatives, répertoire incomplet.

Identifier les individus EDP (2)

- Résultats de l'identification variables selon la source ...

  => données administratives versus données d'enquête

- … et selon les caractéristiques de l'individu (par ordre décroisant de taux d'échec)

  => nés à l'étranger (lieux de naissance non discriminants, répertoire incomplet)

  => les enfants en bas âge et les très âgés

  => les femmes de plus de quarante ans (nom marital)

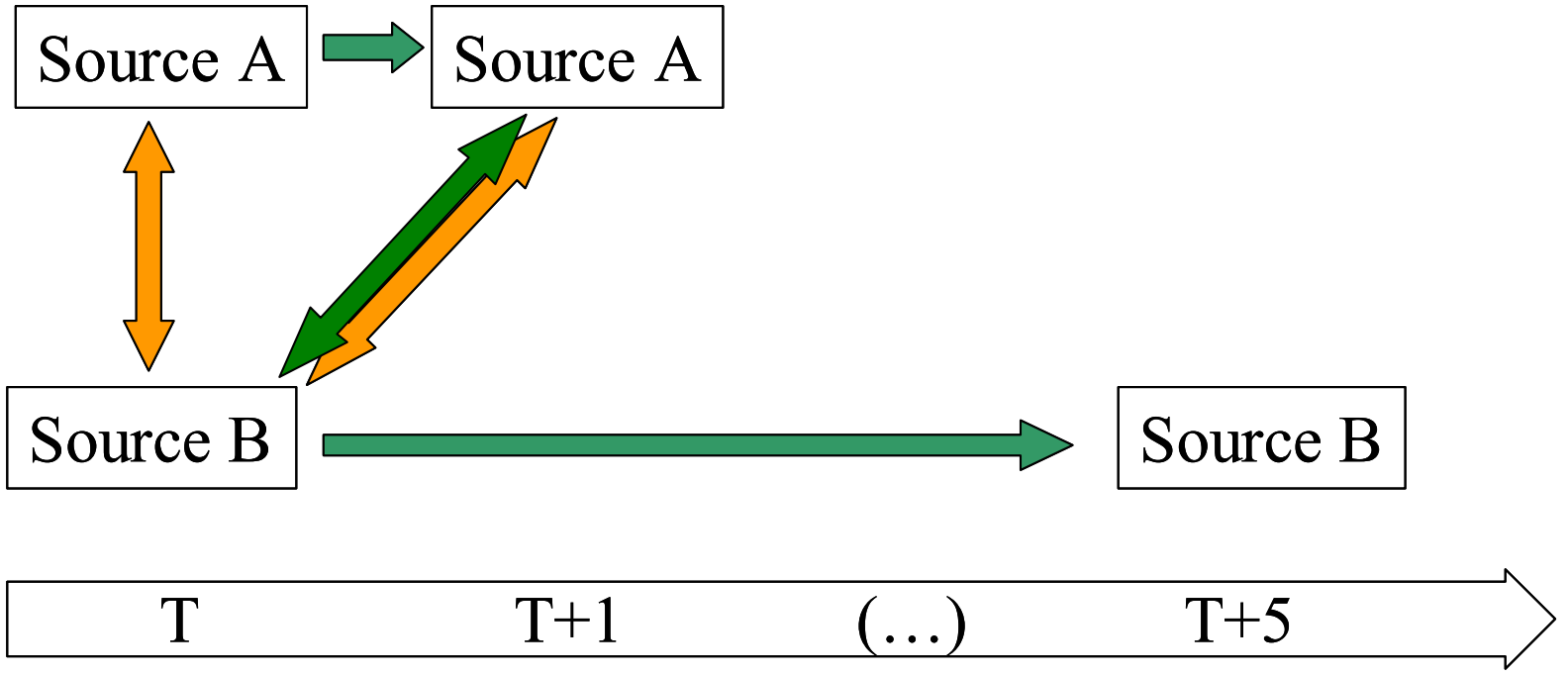
- Solution : considérer l'identification comme une phase supplémentaire de sondage.

  => en tenir compte dans les pondérations (calage)

Intérêt de l'EDP: croiser les sources

=> Enrichir mutuellement les sources utilisées

=> « Panéliser » les sources utilisées



Formulation du croisement de sources

L'échantillon d'individus EDP obtenu par le croisement de deux sources peut être considéré comme le résultat d'un sondage à trois phases, avec comme probabilité d'inclusion pour un individu *i* :

$$\pi_i = \pi_i^{EDP} * \pi_i^{s_1 \cap s_2} * \pi_i^{I_1 \cap I_2}$$

avec : π_i^{EDP} la probabilité d'être un individu EDP

$\pi_i^{s_1 \cap s_2}$ la probabilité d'être dans les deux sources

$\pi_i^{I_1 \cap I_2}$ la probabilité d'être identifié dans chacune des deux sources

Cas général du croisement de sources

- Pour l'ensemble des croisements : $\pi_i^{I_1 \cap I_2} = \pi_i^{I_1} * \pi_i^{I_2}$
  car les identifications sont réalisées indépendamment
- Pour la plupart des croisements : $\pi_i^{S_1 \cap S_2} = \pi_i^{S_1} * \pi_i^{S_2}$
  car les sources sont indépendantes entre elles.
- Pour toutes les sources exhaustives : $\pi_i^{S_{exhaustive}} = 1$
- Dans le cas où une des sources est une EAR, on a π_i^{EAR} fournie par le plan de sondage de l'EAR

  => reste le cas où on croise deux EAR

Le panel EDP dans les EAR

- Avec le passage aux enquêtes annuelles, c'est à dire à un « recensement » rotatif sur cinq ans et par sondage avec des probabilités de tirage inégales, est-il encore possible de réaliser un panel EDP ? Que faire quand on ne retrouve pas un individu EDP ?
- OUI, mais avec les restrictions suivantes :

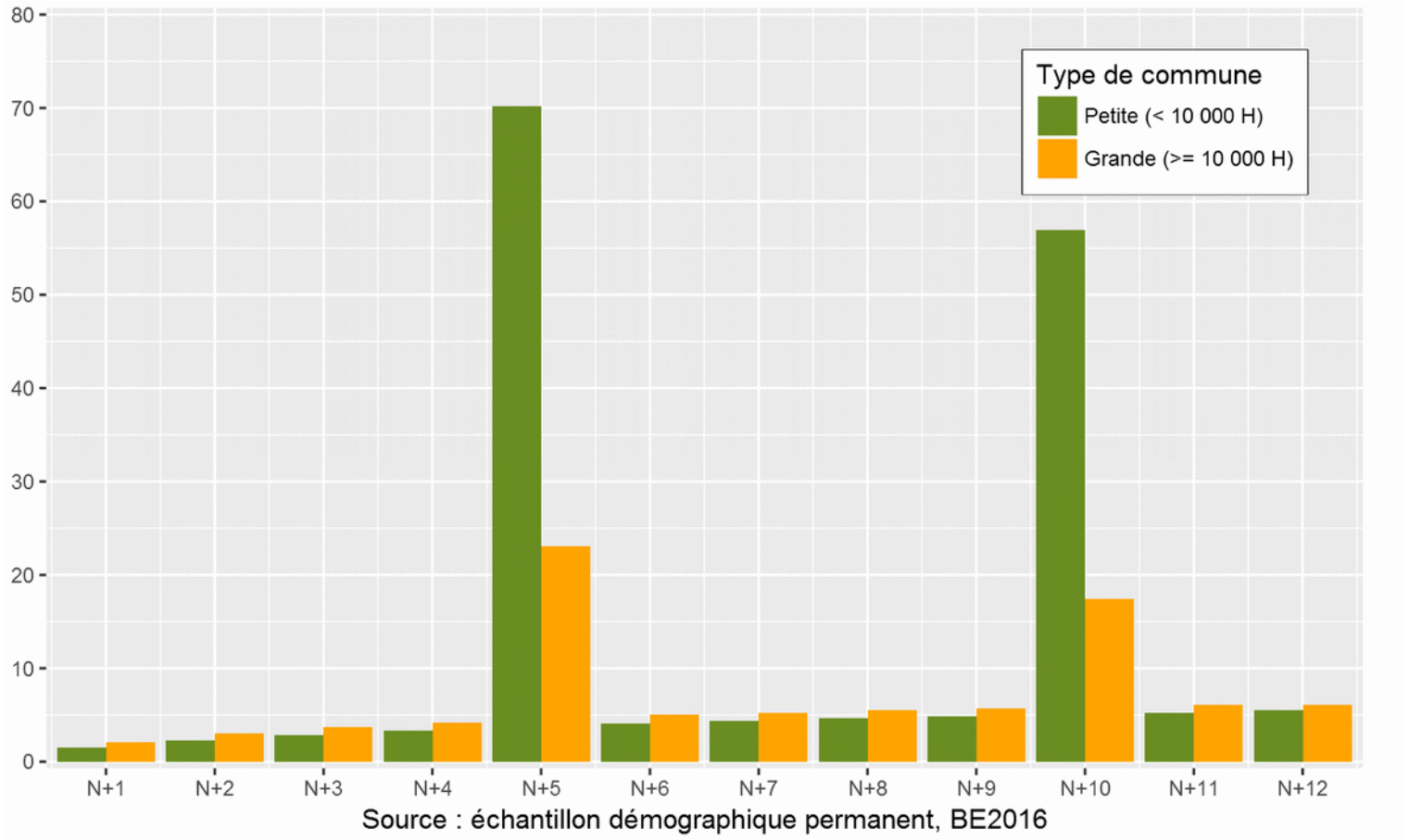
  => le champ est celui des présents aux deux dates

  => se restreindre aux panels N/N+5 (N/N+10, …) où toutes les trajectoires possibles sont représentées

  => utiliser des pondérations adéquates

Sur-représentation des petites communes

Proportion d'individus EDP recensés en N et recensés à nouveau les années d'après selon le type de commune en N



Source : échantillon démographique permanent, BE2016

Pondérations pour le panel EDP-EAR

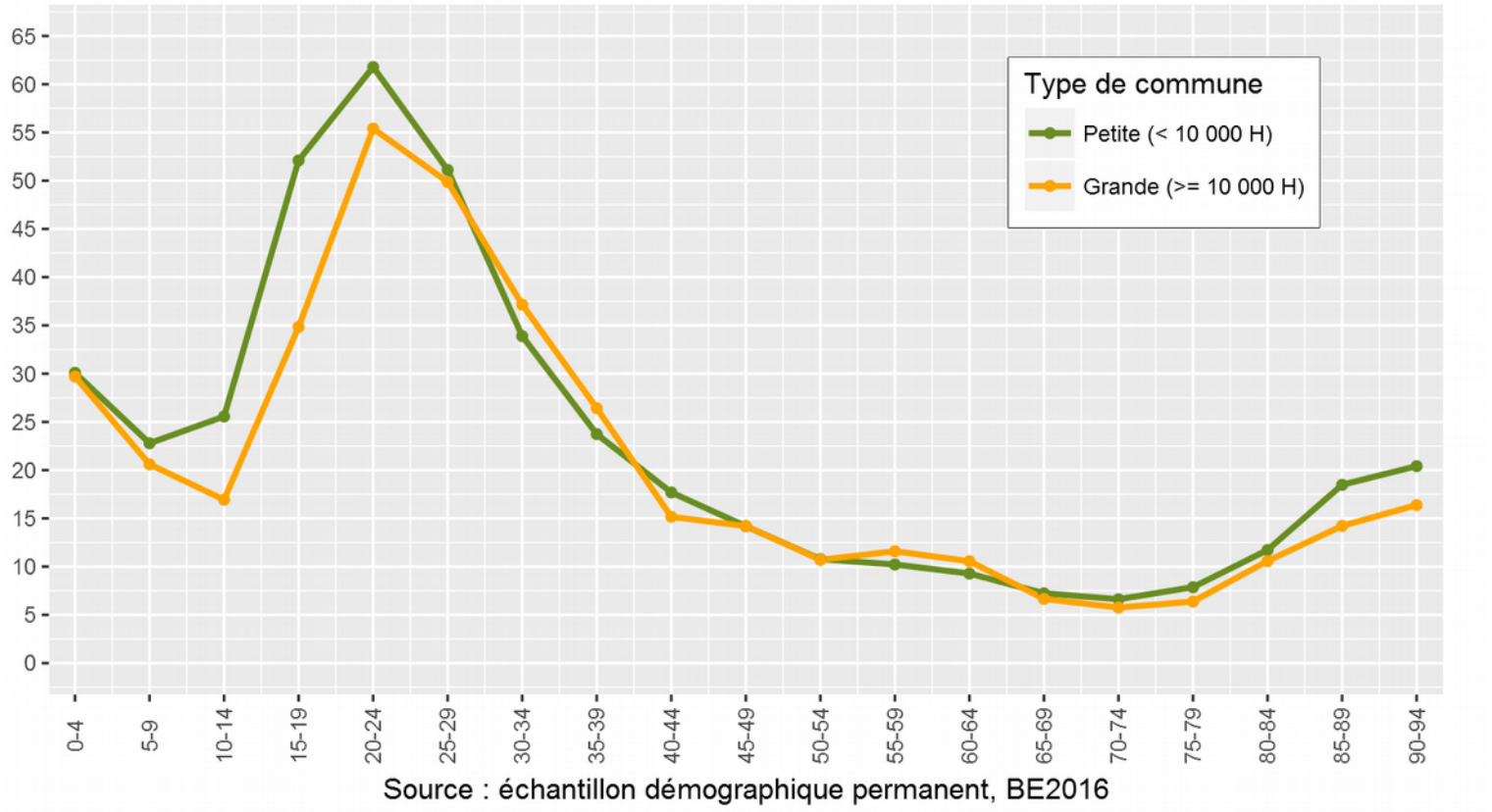
- Pour les EAR, les logements/adresses étant affectés de manière fixe (dans le temps) à un groupe de rotation, l'indépendance des tirages d'un individu à deux dates n'est pas garantie. Il faut alors détailler l'ensemble des cas possibles (Ardilly, 2018), soit en résumé/simplifié/pratique :

  => si les deux logements en N et N+5 sont dans la même PC/même adresse de GC, on prend le produit des poids en N et N+5 divisé 5

  => dans les autres cas, on peut considérer qu'il y a deux tirages indépendants, donc on prend le produit des poids en N et N+5

Étude de la mobilité sur 5 ans

Proportion d'individus changeant de commune entre N et N+5 selon l'âge et le type de commune en N (moyenne des panels 2007-2012, …, 2011-2016)



Source : échantillon démographique permanent, BE2016

Conclusion

- Un sondage original mais qui marche
- Une identification améliorable mais en pratique corrigeable via les pondérations
- Des croisements de source plus compliqués dans le cas des EAR mais avec des pondérations disponibles.