LA VIE EN CONFINEMENT

REDRESSEMENT DE L'ENQUÊTE

Olivier Lê Van Truoc

Grenoble, vendredi 19 juin 2020

REDRESSER L'ÉCHANTILLON

- Fortes sous / sur représentations de certaines catégories de population :
  - Liées à différents facteurs (mode de diffusion de l'enquête, recueil online, longueur du questionnaire, thèmes abordés, période…)
  - L'échantillon brut, en l'état n'est pas représentatif de la population
- Effectuer un redressement (une pondération) par « calage sur marges » :
  - Forme de stratification a posteriori
  - Retrouver une structure d'échantillon conforme à la structure « réelle », sur une sélection de variables clés, choisies car ayant un lien plausible avec les variables d'intérêt de l'étude et permettant de réduire des biais d'échantillonnage.
  - Les données de cadrage sont issues des études de référence de l'INSEE
- Deux objectifs :
  - Corriger certains biais d'échantillonnage pour assurer une meilleure représentativité
  - Améliorer la précision des estimations (pour les variables d'intérêt de l'enquête corrélées aux variables de calage).

LES PRINCIPES

- Le calage sur marges fonctionne par itération. On part d'un poids initial pour chaque individu, ici un poids de 1 (données brutes)
- L'algorithme cherche pour chaque individu, un nouveau poids, qui soit tel que:
  - Le nouveau poids soit le plus proche possible du poids initial
  - Les marges du tableau redressé soient les plus proches possible des objectifs
- Ce poids sera ensuite utilisé dans tous les tris :
  - En schématisant, un individu appartenant à une catégorie de population sous-représentée, aura un poids > 1, et *a contrario* < 1 si la catégorie est sur-représentée.
- Les questions sur lesquelles on cale l'échantillon doivent être renseignées → une non réponse sur l'un des critères conduit à un poids nul.
- Le contrôle de la qualité d'un redressement : convergence vs objectifs, amplitude et distribution des poids, indice d'efficacité du redressement, impact sur des variables non contrôlées par la procédure…

POUR L'ENQUÊTE VICO

- Les atouts :
  - Un échantillon de grande taille et diversifié
  - Des variables bien renseignées
  - Des données de cadrage récentes et fiables
- Les difficultés :
  - De fortes sous ou sur-représentations entraînant *de facto* une amplitude des poids importante
  - Des biais constants : par exemple toutes les tranches d'âges comportent trop de diplômés du supérieur.

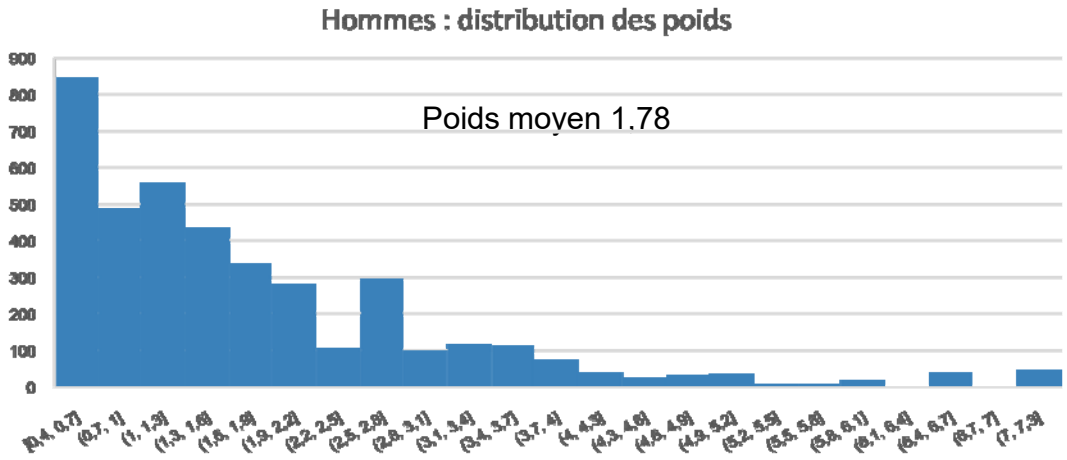
- Une dizaine d'essais effectués :
  - Sélections de variables différentes
  - Regroupements de modalités, …
- Les variables :

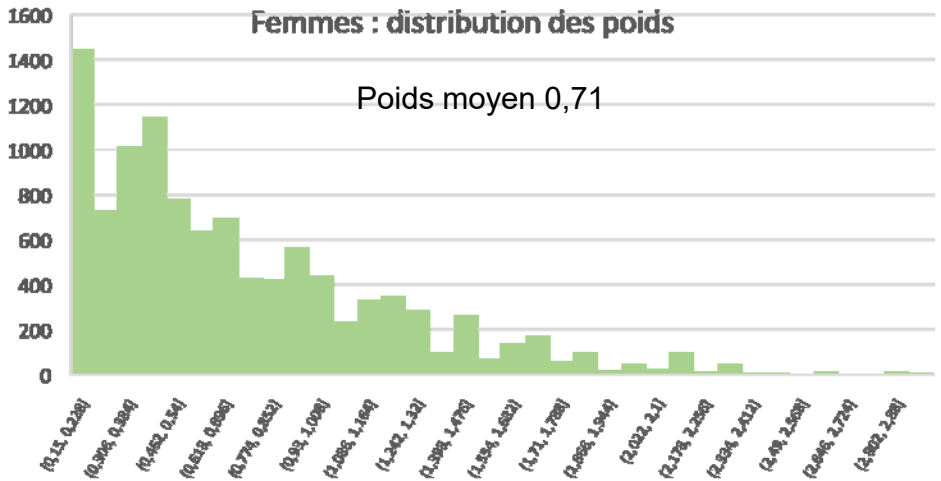
	Systématiquement	Selon les essais
Sexe	X	
Age	X	
PCS		X
Régions	X	
Tranche d'unité urbaine		X
Diplôme		X

- Un jeu de pondération sera proposé pour le Datathon

EXEMPLE : UN DES ESSAIS DE PONDÉRATION

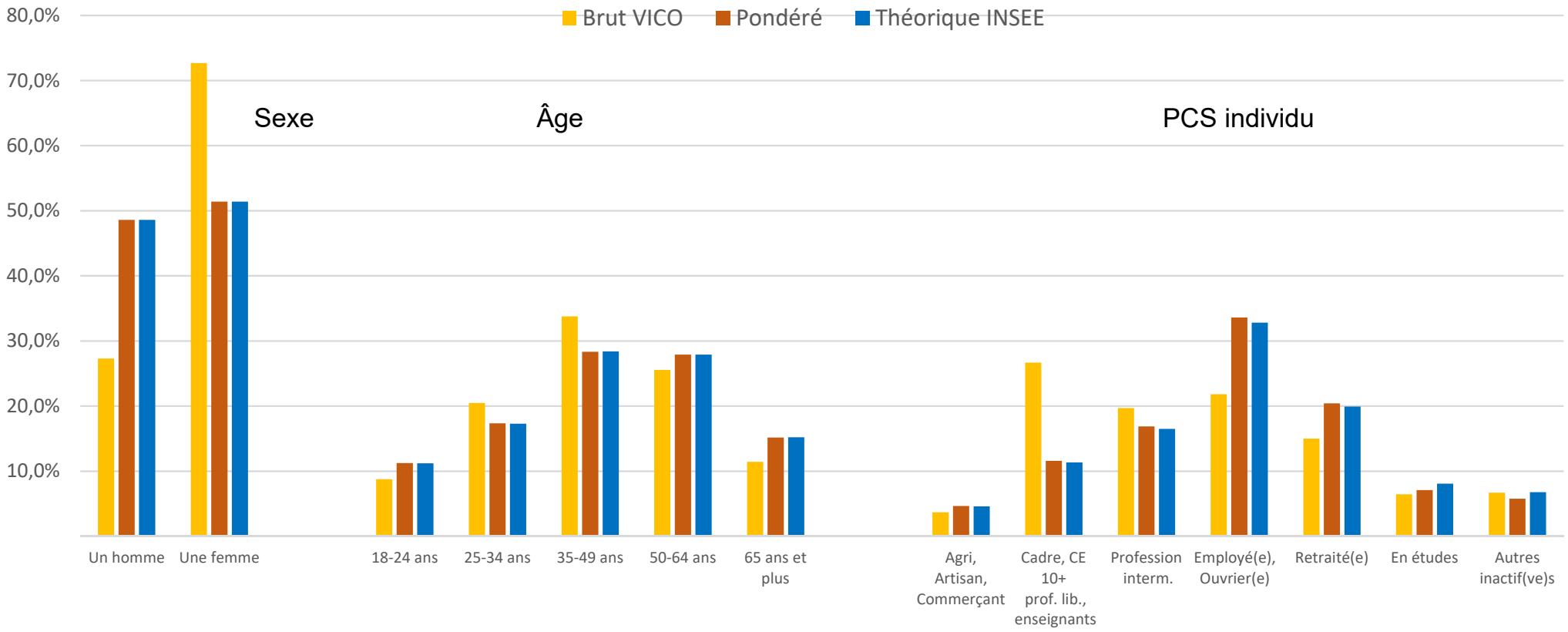
- On a redressé en deux étapes, séparément les hommes et les femmes
  1. Remise à leurs bons poids des hommes et des femmes
  2. Puis pour les hommes d'un côté, et les femmes de l'autre, redressement sur Age (5), PCS(6), Régions (8), TUU (2)
     - Au total 2 x 21 modalités
- En bornant les poids à +/- 4 fois le poids de chaque « strate » (hommes / femmes)
- Efficacité du redressement acceptable (vue l'importance de la correction effectuée) : hommes 63% / femmes 68%





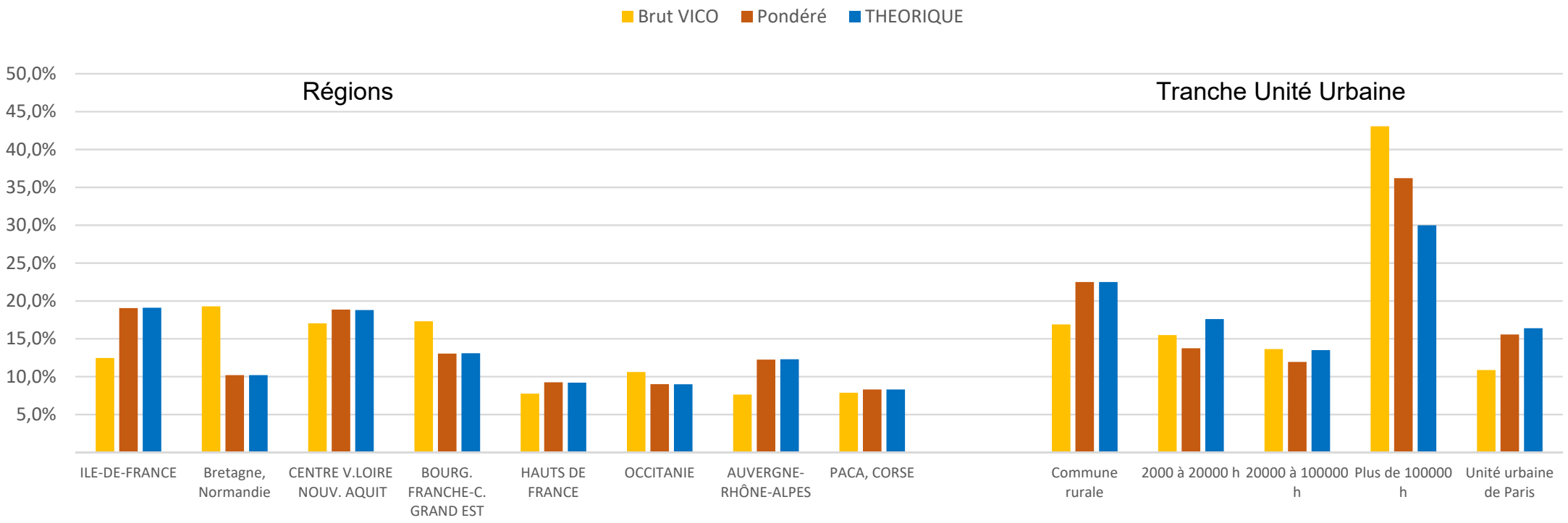
RÉSULTATS SUR LES CRITÈRES PRIS EN COMPTE

- Le fait de borner les poids n'empêche pas le redressement de tendre vers les marges théoriques.



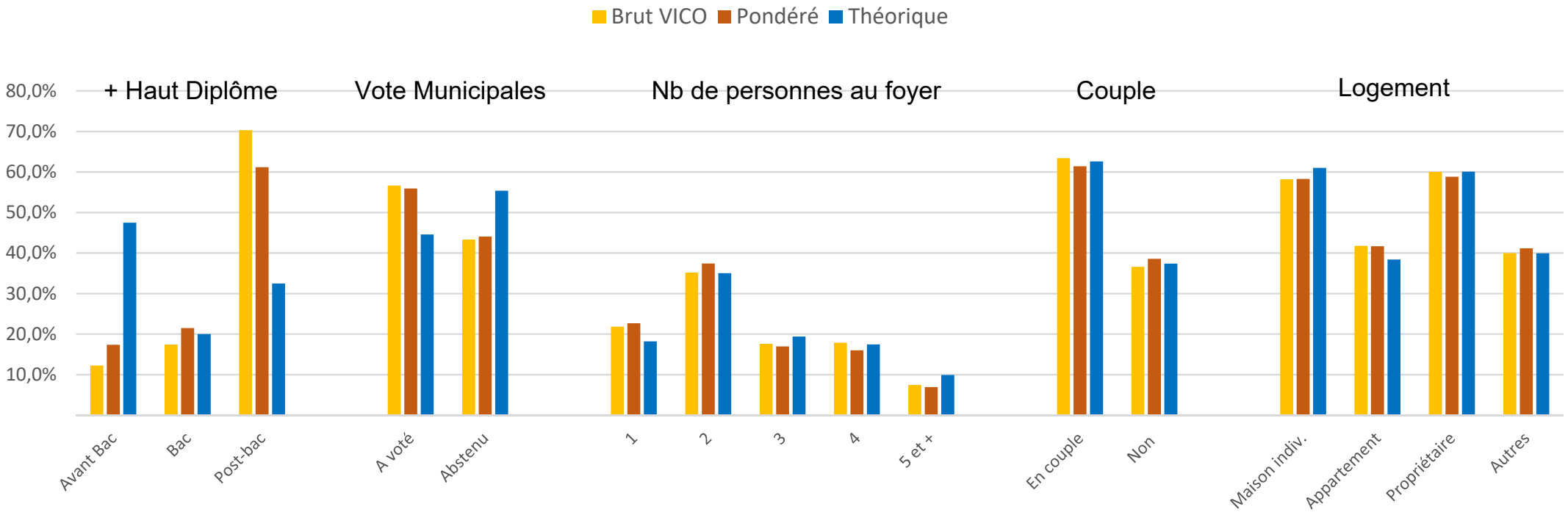
RÉSULTATS SUR LES CRITÈRES PRIS EN COMPTE

- En redressant uniquement sur rural vs urbain (et sur la région), on améliore la Tranche d'Unité Urbaine, mais l'échantillon demeure un peu trop urbain (> 100 000h)



ET SUR QUELQUES CRITÈRES NON MAÎTRISÉS

- Le redressement a un impact positif, mais insuffisant, sur le diplôme. Très léger sur la participation aux municipales. L'échantillon est bien calé (avec ou sans redressement) sur le nombre de personnes au foyer, le statut d'occupation ou le type de logement.



ZOOM SUR LE PLUS HAUT DIPLÔME OBTENU

- Comme on compte trop de diplômés du supérieur quelle que soit la classe d'âge ou le sexe, le redressement effectué corrige insuffisamment les écarts avec la distribution théorique. L'impact de la pondération est plus fort sur les tranches d'âge les plus actives (25-64 ans).