L’utilisation combinée de données d’enquête et de données administratives pour la production des statistiques structurelles d’entreprises

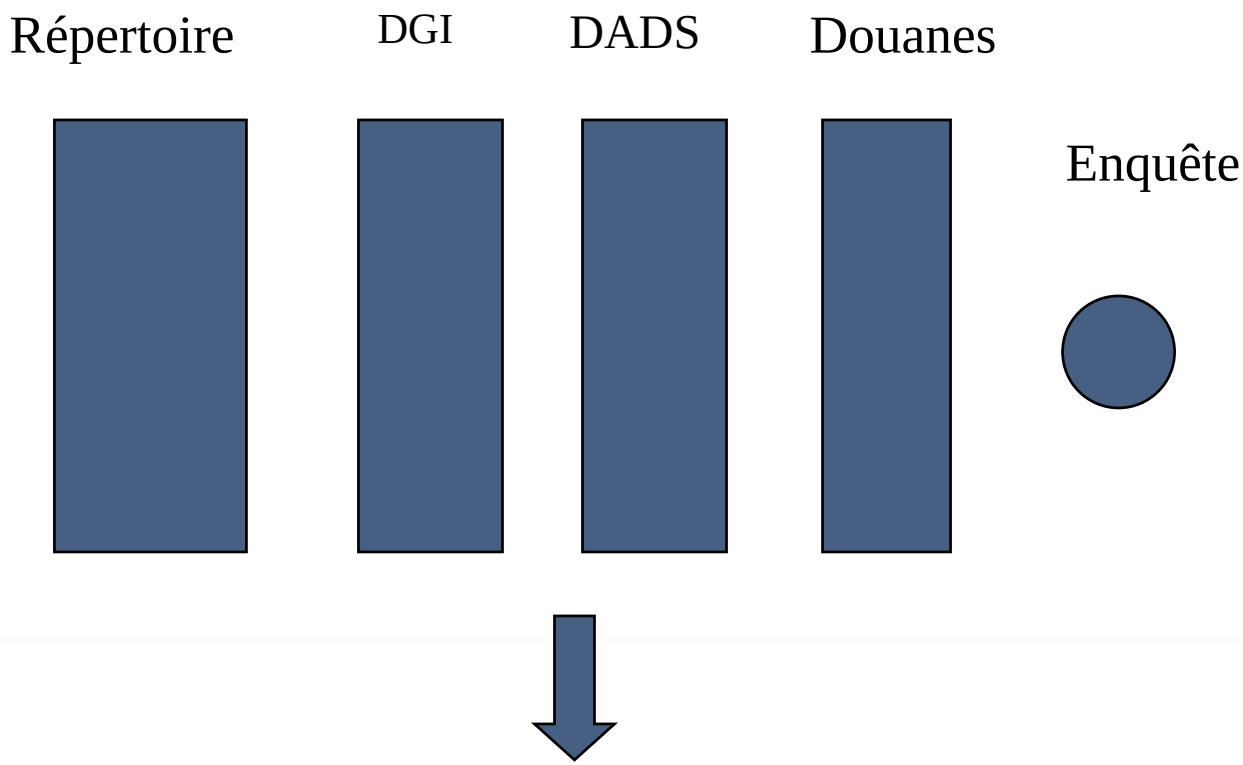
JMS 2009

Ph. Brion
Insee

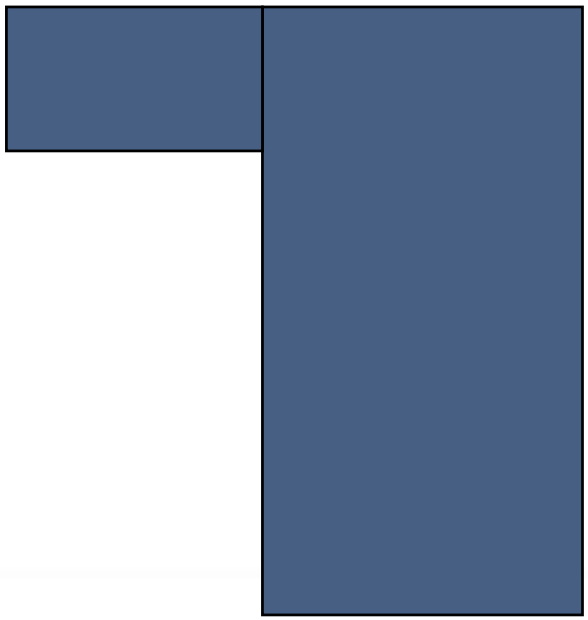
Le précédent dispositif de production des statistiques structurelles d'entreprises

- Deux dispositifs parallèles :
  - des enquêtes (EAE)
  - l'utilisation des données fiscales (SUSE)
- Objectif : produire les données de structure relatives aux entreprises, en particulier sur les variables comptables
- La spécificité des statistiques sectorielles

Les différentes composantes du dispositif ESANE



Deux méthodes possibles pour produire des statistiques



- Partie de droite du dessin : les données administratives
- Partie de gauche : données d'enquête (échantillon)
- Deux familles de méthodes
  - L'imputation de masse
  - Les estimateurs statistiques « combinés »

- *Dans la suite de la présentation, on s'intéresse essentiellement aux questions relatives à la partie échantillonnée*

Les estimateurs statistiques combinés (1)

- Pour les statistiques utilisant des variables obtenues grâce à l'enquête, on utilise l'estimateur standard :

$$\sum_s w_i X_i$$

- Mais on peut utiliser les données administratives, dans un premier temps, pour modifier les poids de façon que (CA étant le chiffre d'affaires) :

$$\sum_s w_i 1_{APErep=X}(i) CA(i) = \sum_U 1_{APErep=X}(i) CA(i)$$

Les estimateurs statistiques combinés (2)

- Pour les statistiques sectorielles (à savoir utilisant le classement sectoriel obtenu à l'enquête), on peut de plus utiliser un estimateur par différence :

$$\sum_U 1_{APErep=X}(i)[CA(i)] + \sum_s w_i (1_{APEenq=X} - 1_{APErep=X})(i)[CA(i)]$$

- Au niveau où on a procédé au calage, les deux estimateurs (c'est-à-dire l'estimateur par différence et l'estimateur « basique ») sont équivalents (pour la variable chiffre d'affaires uniquement, si c'est elle qui est utilisée pour le calage)

Une phase de réconciliation des données

- Cette phase de réconciliation des données « individuelles » peut conduire à modifier la valeur de certaines variables administratives ; au final, un nouvel estimateur sera utilisé, qui permettra donc de faire de l'inférence sur les données administratives :

$$\sum_U 1_{APErep=X}(i)\left[Y_{adm}(i)\right]+\sum_s w_i\left[1_{APEenq=X}(i)Y_{vraie}(i)-1_{APErep=X}(i)Y_{adm}(i)\right]$$

Le contrôle redressement des données

- Il concerne des données arrivant à différentes périodes

- Quelles « normes » pour le contrôle de chacune des sources ?

- Egalement, les poids de sondage seront modifiés pour prendre en compte la non réponse totale

Conclusion

- ESANE : première année d'existence, les premiers résultats sont prévus fin 2009
- D'autres études seront à mener, en particulier sur la production de résultats précoces
- La production de statistiques à partir d'un matériau composite pose des questions plus complexes que celles posées par une « seule » enquête