# Quels usages des données massives pour les statistiques publiques?

*Enjeux, méthodes et perspectives*





Stéphanie Combes et Pauline Givord (DMCSI)

Insee

Mesurer pour comprendre

02/04/2015

# Plan

- Qu'est-ce que le Big Data ?
- Les Big Data pour la statistique publique ?
  - Opportunités et questions
- Quelles compétences pour les statisticiens face à des données massives ?
  - Outils statistiques pour de gros volumes, de grandes dimensions

# Qu'est-ce que le Big Data ?

- Pas de définition fixe mais on peut les caractériser par :
  - Mode de génération : enregistrement automatiques d'activités
  - Volume : élevé qui nécessite le développement d'outils et d'infrastructures adaptés
  - Variété : peu ou pas structurées, différents formats
  - Vélocité : flux continus (en « temps réels »)
- Origine : développement des objets connectés (mobile, capteurs...), acteurs du Web, données de gestion (données de caisse, PMSI) mais aussi Big science (génomique, astronomie)

# Les données massives pour la statistique publique ?

- Réflexions des différents INS sur la possibilité d'utiliser ces données dans leurs productions
- Quelques projets :
  - Données de caisse (Insee) : améliorer la précision, produire des indices de prix localisés, des prix moyens par produits...
  - Données de téléphonie mobile : compléter les statistiques sur le tourisme (Eurostat), la mesure des temps de trajets
  - Recherches internet : améliorer la prévision conjoncturelle (note de conjoncture Insee mars 2015) ; données de réseaux sociaux : indicateur de confiance des ménages (Statistics Netherlands)
  - Imagerie satellite : statistiques agricoles (INS Australie)

# Opportunités

- Disponibilité immédiate de l'information : réduire les temps de publication
- Nombre d'observations très élevé : publication à des échelles territoriales plus fines, des sous-populations, événéments rares
- Réduire la charge d'enquête : compléter les indicateurs existants
  - Temps de transport, dépenses précises
- Réduire les coûts ?
  - Données plutôt complémentaires à la production actuelle (pauvres en caractéristiques sociodémographiques)

# Défis

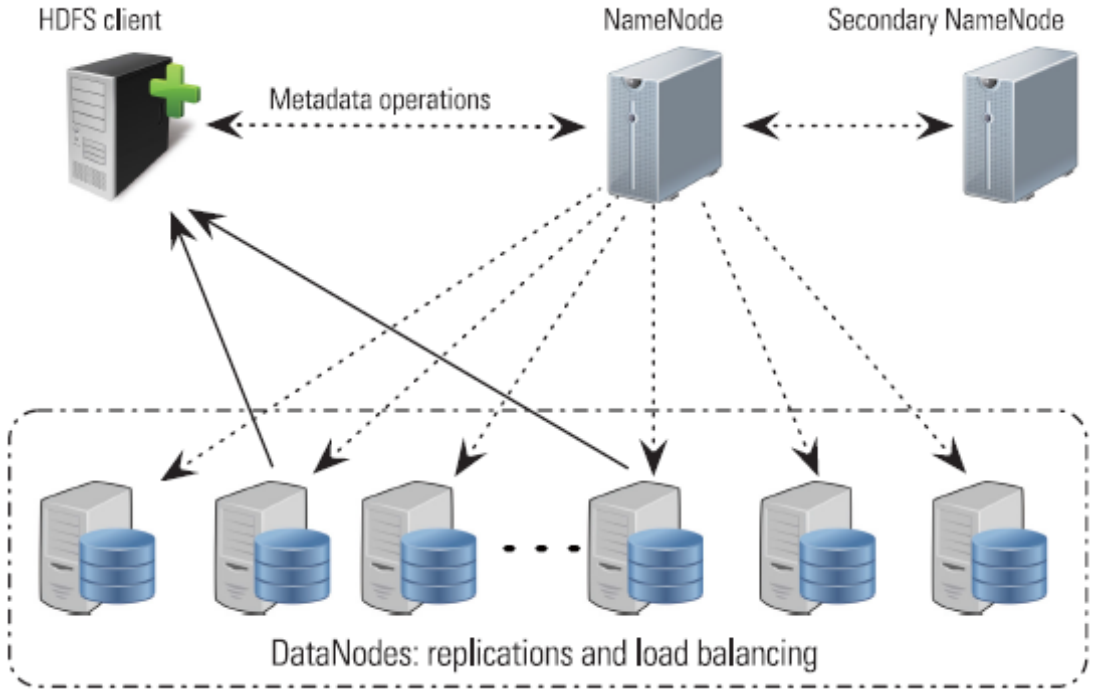
- Protection de la vie privée :
  - Utilisation de données personnelles ; risque de réidentification augmenté par la tentation/nécessité d'appariement des données
- Accès aux données : coût, droit, reproductibilité, garantie d'indépendance
- Concurrence / positionnement de la statistique publique
- Qualité des données
  - Représentativité des données
- Pertinence des outils statistiques utilisés ?

# Quelles compétences /outils pour les statisticiens ?

- Les grands volumes ont conduit au développement d'infrastructures adaptées
  - Données stockées sur des serveurs en parallèle, calculs distribués sur ces serveurs, mais l'interface est transparente pour l'utilisateur
- Utile d'avoir des notions de base pour faire des choix raisonnés
  - objectif de cette (modeste) présentation

# Quels outils pour gérer de gros volumes ?

- Des données stockées sur des fichiers distribués (Hadoop Distributed File System)
- Les calculs peuvent être menés en parallèle sur différents serveurs (Map Reduce)



# Paralléliser des calculs statistiques

- Tout n'est pas échelonnable ! Demande de pouvoir « découper »les opérations sur des parties de la base
- Opération élémentaire comme MCO l'est :
  - Cas où beaucoup d'observations (n grand) réparties sur plusieurs nœuds (indicés par $a$), mais peu de variables ($p << n$)
  - Estimateur MCO : $(X'X)^{-1}X'Y$ : inversion de la matrice X'X qui est « seulement »de taille $(p, p)$
  - Terme générique de la matrice $X'X$ : $X_j'X_k = \sum_{i=1}^{n} X_{i,k}X_{i,j} = \sum_a X_j^{(a)'} X_k^{(a)}$
- Remarque : la plupart des logiciels stats (R, SAS) permettent de travailler sur des bases distribuées
- Ce n'est parce que c'est faisable que c'est optimal !

# Quelles compétences pour le statisticien ?

- Utiliser des technologies de stockage type Hadoop pas toujours optimal selon le volume
- Outils de type Map Reduce adaptés pour des tâches élémentaires, moins pour des tâches plus complexes
- Si les données sont très volumineuses, il peut être utile d'échantillonner pour utiliser des traitements plus complexes (exemple : algorithmes itératifs, validation croisée)
- Logiciels plus adaptés (ex Spark) à ces traitements sont en développement
- Réflexion à avoir sur les méthodes les plus adaptées : dépend de la structure des données, de leur mode de stockage... mais aussi de la nature de la question

# Quelles compétences pour le statisticien ?

- Les données massives correspondent aussi à beaucoup de régresseurs potentiels ($p$ grand)
- Enjeu principal de la statistique : extraire l'information pertinente
- Divers types d'outils qui répondent à des priorités différentes :
  - Datamining, prévision : identifier les covariables essentielles pour minimiser l'erreur de prévision
  - Econométrie : mettre en évidence des relations entre variables

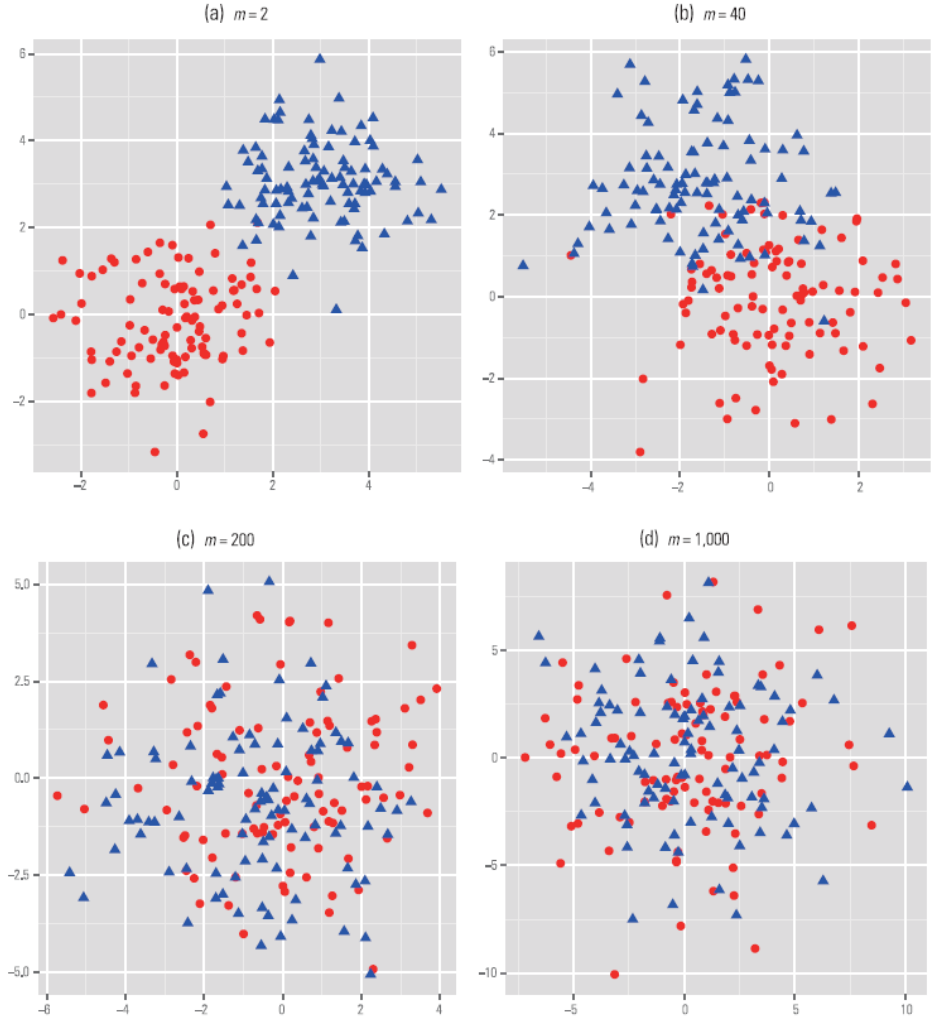
# Quelles compétences pour le statisticien ?

- Nécessaire de dépasser ses réflexes
  - Intérêt des méthodes de réduction de la dimension face à des données massives (sélection des régresseurs pertinents, spécification non linéaire)
  - Meme en prévision / data mining, l'endogénéité peut être un problème
- En grande dimension :
  - la parallélisation peut introduire des problèmes pour la convergence/stabilité/réglage de certains algorithmes $\leftarrow$ importance du choix de la méthode
  - les problèmes classiques (erreurs d'estimations, endogénéité, sur apprentissage) peuvent être amplifiés

# Des problèmes amplifiés par la grande dimension - des estimations bruitées

- Lorsque l'on estime simultanément un grand nombre de paramètres, les erreurs s'accumulent
- Illustration (issue de Fan, Han et Liu, 2014) :
  - On génère deux classes d'observations, seules les dix premières variables sont réellement discriminantes
    $X_1, ..., X_n \sim N_p(\mu_1, I_p)$ et $Y_1, .. Y_n \sim N_p(\mu_2, I_p)$,
    $p = 1000$, $n = 100$
    $\mu_1 = 0$ et seules les dix premières composantes de $\mu_2$ sont non nulles.
  - Résultat d'une ACP selon qu'on utilise m=2, 40, 200, 1000 variables

# Des problèmes amplifiés par la grande dimension - des estimations bruitées



# Des problèmes amplifiés par la grande dimension - des estimations bruitées

- Réduire la dimension peut être une solution
- Hypothèse de sparsité (parcimonie) : parmi les très nombreux régresseurs, seul un nombre limité a un effet non nul
- Différence par rapport à l'analyse économétrique « standard » : on ne sélectionne pas a priori la liste de ces régresseurs
- Estimation par des algorithmes de sélection de variables type LASSO :
  - On introduit un terme de pénalisation, et on minimise : $-QL(\beta) + \lambda \|\beta\|_1$
  - Conduit à forcer l'annulation de nombreux paramètres

# Des problèmes amplifiés par la grande dimension - corrélations et relations fortuites

- Corrélations fortuites :
  - Si on calcule des corrélations sur un nombre important de variables, risque élevé d'obtenir une corrélation statistique par pur hasard
  - Endogénéité incidente : sélection aveugle risque d'introduire un régresseur endogène, ce qui biaise l'inférence
- Sur-apprentissage :
  - Une faible erreur d'estimation peut conduire à une grande erreur de prévision
  - Risque peut être réduit par des méthodes de validation croisée, mais coûteux sur données volumineuses
- Discipline en formation... pas de conclusions définitives

# Conclusion

- L'Insee n'en est qu'au tout début de ses réflexions sur le sujet
- Le projet Données de caisse déjà avancé, d'autres plus expérimentaux devraient se mettre en place
- Sur le plan des méthodes, demande un investissement dans d'autres outils (analyse textuelle, appariements complexes, visualisation...) que ceux évoqués ici
- Les enjeux ne sont évidemment pas que statistiques :
  - nécessité d'une approche/ équipe pluridisciplinaire
  - Data scientist : compétences statistiques, IT, juridique,...
  - Réseaux : nécessité de développer des collaborations

# Quels usages des données massives pour les statistiques publiques?

## Merci de votre attention !

Contact :
M. Stéphanie Combes et Pauline Givord (DMCSI)
Tél. : 01 41 17 66 01
Courriel :pauline.givord@insee.fr

Insee

**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