

# LES DONNÉES NE SONT PAS DONNÉES

Réflexions sur le processus de production  
d'information

*Pascal Rivière, INSEE*

*21 mars 2005*

## Idée générale

- A travers l'exemple de la production de statistiques, on étudiera en quoi consiste le processus de production d'information
- On en tirera des idées générales sur la production de données et la problématique de la qualité des données

## Distinctions importantes

- Information *qualitative* / *quantitative* / *textuelle*
- Information *statistique* (anonyme, agrégée, macro) / information *individuelle* (élémentaire, nommée, micro)
- *Statistique mathématique* (inférentielle, à données préexistantes) / *production statistique* (le problème est de construire les données)
- *Unités* (ou individus, au sens large) / *variables* (qui caractérisent ces mêmes individus)

## Principaux aspects de la production statistique

- Expression du besoin
- Mise au point du support de questionnaire
- Gestion de référentiels
- Mise au point de l'échantillon fondé sur ces référentiels
- Collecte, gestion de la collecte
- Apurement
- Traitements statistiques finaux
- Utilisation des statistiques
- Aspects administratifs et juridiques

## Le(s) besoin(s)

- Pas d'enquête sans objectif : les données collectées dans une enquête n'ont de sens que vis-à-vis d'une demande d'information
- Ce besoin est naturellement exprimé de façon vague. Exemple : « *mieux connaître le comportement des petites entreprises industrielles vis-à-vis de l'innovation* »
- Etroitement lié à un contexte, à des normes, des conventions, des institutions → on ne peut pas demander n'importe quoi, une enquête a un coût

## Discussion sur le besoin

- Evaluation des données existantes, comparaison (éviter de redemander ce qui existe déjà, mais aussi se positionner par rapport à l'existant)
- Opportunité : y a-t-il une vraie demande ?
- Coût de l'opération
- Charge de réponse (= coût pour les répondants)
- Possible remise en question à ce stade : coût élevé, durée de collecte trop longue, confidentialité

## Passage du besoin à la solution

- Construire les données statistiques souhaitées, c'est créer l'ensemble des outils permettant :
  - D'extraire le matériau informationnel (collecte)
  - De le transformer (apurement)
  - De le mettre à disposition sous forme de produits identifiables (documents, CD, pages web, ...)
- Spécificité en statistique : passage individuel – agrégé, et in fine anonymie des individus

## Différence enquête / source administrative

- Pas de maîtrise du processus dans le cas d'une source externe :
  - Toutes les transformations sont exogènes
  - Pas de moyen d'action sur le formulaire d'interrogation
  - Concepts et champ imposés
- Le gain en termes de coût de ne pas effectuer d'enquête n'est pas évident

# Les référentiels : répertoires et nomenclatures

- Exogènes au processus de production d'information ; mais ce sont des moteurs qui le facilitent (en statistique ou ailleurs)
- Fort impact sur la qualité du produit informationnel final
- Nécessité d'une gestion permanente des répertoires et d'une mise en place de nomenclatures (conventions communes, donc données sommables)
- Difficulté de la coordination des identifiants (exemple : Impôts – INS) et des nomenclatures (ex. nomenclatures d'établissements)

## Notion de répertoire

- Ensemble d' « individus », munis d'identifiants, de caractéristiques d'identification (adresse notamment), et d'autres variables d'intérêt, comme des variables de classement
- Le problème des unités de référence (ex : qu'est-ce-qu'une entreprise ?)
- Un répertoire est évolutif : naissances, cessations, changements de caractéristiques
- Gérer un répertoire est très coûteux : en raison de la complexité du circuit d'alimentation de celui-ci et des multiples fonctions attendues
- Exemples : répertoire de clients, d'entreprises, de logements, ...

## Fonctions que doit remplir un répertoire

- Etre aussi complet que possible, cohérent, à jour
- Contenir les unités (statistiques, légales) standard
- Contenir pour chacune : identifiant, données d'identification, données de classement, métadonnées, données de gestion
- Etre doté d'outils facilitant les travaux d'intégration et de fourniture de données
- Avoir un système documentaire complet
- S'appuyer sur un environnement juridique et une organisation solide

## Vie du répertoire

- Circuits de mises à jour : intégré à la gestion, ou bien via des enquêtes ou sources externes
- Qualité : ne pas chercher le 0 défaut. Parties du référentiel moins bonnes que d'autres
- Utilisation : en statistique, à titre de base de sondage, ou du moins de référence pour construire des bases de sondage

# Niveaux d'utilisation effective d'un répertoire

- Distinction entre :
  - Base de gestion du répertoire : vivante, évolue tous les jours, en production courante (souvent, c'est cette base de gestion qui est nommée « répertoire »)
  - Photo du répertoire : tout ce que l'on y trouve à un moment donné
  - Référentiel bidaté : on extrait du répertoire la situation connue à une date  $t'$  relativement à une date  $t$  → outil essentiel pour la diffusion, mais aussi en tant que base de sondage
- Cette distinction est fondamentale sur le plan opérationnel; en pratique, fréquentes confusions

# Nomenclatures

- Nomenclature = emboîtement de partitions. Au contraire du répertoire qui représente un ensemble en extension, la nomenclature décrit en intention
- Exemples de nomenclatures : activités, produits, professions, géographie
- Impossibilité d'une bonne nomenclature générale : elle est toujours liée à un besoin
- Liens entre nomenclatures sur un même sujet : non nécessairement biunivoques
- Surtout limiter les nomenclatures concurrentes

# Le champ

- Il s'agit de délimiter une sous-population d'intérêt (par exemple « les entreprises commerciales de plus de 50 salariés »)
  - Unités statistiques
  - Définition en intention de la sous-population
  - Mise en œuvre opérationnelle du champ
- Statistiques calculées au sein de ce champ
- Délimitation
  - a priori (avant l'enquête, grâce au répertoire),
  - a posteriori (grâce aux données d'enquête)

# Les différentes notions d'« unité »

- Unité statistique : brique de base des statistiques, unité au sujet de laquelle on recueille l'information
- Unité de collecte : unité auprès de laquelle on recueille l'information
- Unité légale : ayant une existence juridique
- Unité économique
- Etc.

## On a donc des « réceptacles à information »

- (unités de collecte, questions) = instrument de collecte
- (unités statistiques, variables) = base de données = ensemble de réceptacles
- Contenu des réceptacles régi par des normes et conventions (identifiants d'unités, nomenclatures, conventions internes)
- Réceptacles non-vides au départ : caractéristiques d'identification, valeurs de l'enquête précédente, sources administratives, ...

## Tirage d'échantillon

- Notion de base de sondage : population de référence, fondée sur le répertoire, structurée grâce aux nomenclatures
- Notion de plan de sondage : procédure définie a priori pour tirer un échantillon ayant des propriétés maîtrisées
- Types de sondage
  - Exhaustif : taux de sondage 100%
  - Exhaustif > seuil
  - Tirage systématique
  - Tirage aléatoire simple
  - Tirage aléatoire simple stratifié
  - Tirages à plusieurs degrés

# Panels

- Plusieurs échantillons successifs dans le temps
- Notion de *mise à jour* du panel
- Evolutions dues à la démographie et à la volonté de « faire tourner » le panel
- Arbitrage sur le taux de rotation :
  - Taux de rotation élevé → évite d'interroger toujours les mêmes, donc lisse la charge de réponse
  - Taux de rotation bas → meilleure qualité des calculs en évolution

# La collecte : mise au point du questionnement

- Collectabilité de l'information / lisibilité des questions
- Coût du travail de réponse
- Possible multiplicité des répondants
- Existence de normes (comptables, par exemple)
- Tests de questionnaires
- Méthodes cognitives
- Ordre des questions
- Réutilisation d'autres sources

## Modes de collecte

- Enquêteur sur le terrain
- Courrier
- Fax
- Téléphone / TDE
- E-mail
- Questionnaire sur le web
- CAPI / CATI / CASI / CAWI / PAPI
- Réutilisation de sources administratives
- Collectes mixtes

## Cibler la collecte

- Trouver le bon interlocuteur, le bon « répondant » potentiel
- Bien choisir la date d'interrogation
- Coût de maintenance (adresses notamment)

## Suivi de la collecte

- Suivi des envois, des retours d'informations
- Echelonnement des interrogations (cf. ciblage)
- Gestion des changements d'adresses (« NPAI »)
- Gestion des unités hors-champ
- Suivi des non-répondants

## Encodage

- Inscrire l'information sur support électronique
- Information brute, information primaire. Respect du matériau initial.
- Scanning : transformation brute, sans structurer l'information → on obtient une image
- Saisie : transformation en tenant compte de la structure → on obtient « des données »
- La saisie se fait parfois pendant la collecte (CAPI, CATI, ...) ; par le répondant (EDI, CAWI, ...)
- Reconnaissance optique de caractères, ou OCR : technique automatisant (partiellement) l'encodage

# Codage

- Information sur support électronique, structurée. Passage de la structure de données de collecte à la structure finale
- Idée sous-jacente : le répondant ne dispose pas de l'information selon les conventions souhaitées
- Recodification de variable qualitative, ou à partir de plusieurs variables (ex. activité principale)
- Codage de variable quantitative (tranche d'âge)
- Codage de variable textuelle ou *libellé* : profession, commune, activité, nationalité, ...

## Codage automatique, codage assisté

- C'est le codage de libellés qui pose problème au niveau de l'automatisation
- Données en entrée : un libellé + éventuellement des variables annexes
- Codage automatique : processus batch, boîte noire. En sortie : soit un code, soit rien. Arbitrage entre efficacité du codage ainsi automatisé et fiabilité.
- Codage assisté : outil interactif facilitant la recherche du « bon » code, en permettant par exemple de naviguer dans une nomenclature

# Data editing

- A structure donnée, on agit cette fois sur le contenu
- La présence d'erreurs dans un questionnaire rempli est non pas l'exception, mais la norme
- Le « data editing » a pour but d'éliminer les erreurs subsistant dans les données afin d'obtenir une base de données « propre »
- C'est en général très coûteux : ne pas éliminer toutes les erreurs

# Principes de fonctionnement

- Distinction entre valeur vraie, correcte, acceptable
- Programme déterminant si un questionnaire (ou une valeur) est acceptable ou non
- Puis vérification manuelle si le questionnaire est jugé « douteux »
- Problème d'arbitrage coût-qualité

## Data editing : la pratique

- Le contrôle intervient à plusieurs stades : au moment même de la collecte, au fur et à mesure des traitements (automatique ou manuel), en fin de processus
- Macroediting, microediting
- Priorisation des contrôles, prise en compte de l'impact sur les statistiques finales
- En statistique, le contrôle final au niveau « macro » joue un rôle essentiel
- Notion de « base de données finale » difficile à cerner

## Data editing : commentaires généraux

- C'est un coût important (40% des coûts d'une enquête), souvent méconnu
- La stratégie de data editing doit être fonction des utilisations qui seront faites des statistiques
- Souvent, en croyant « corriger une erreur », on en ajoute une.
- Ne pas hésiter à laisser des erreurs potentielles
- Importance de la boucle data editing - tabulation

## Traitements statistiques finaux

- Traitement des non-réponses partielles
- Traitement des non-réponses totales (imputation ou repondération)
- Tabulation : prise en compte des poids de sondage, repondération éventuelle des non-réponses
- Calculs en évolution : comparer ce qui est comparable

## Aspects administratifs et juridiques

- La spécificité d'un institut de statistique en tant que collecteur d'informations
- Les enquêtes obligatoires
- Le secret statistique
- L'utilisation de fichiers de données individuelles
- La confidentialité
- Le cadre légal de lancement d'une enquête

# Notion de qualité

- « Ensemble des propriétés et caractéristiques d'un produit ou d'un service qui lui confèrent l'aptitude à satisfaire des besoins exprimés ou implicites » (ISO 8402, 1986)
- Aptitude à l'emploi (« fitness for use ») : Juran
- Définition qui n'a rien de spécifique à un secteur d'activité particulier

# Concept de qualité : idées de base

- La notion de qualité est à associer aux usages, ce n'est pas un concept pur
- Qualité  $\neq$  excellence
- Qualité  $\neq$  conscience professionnelle
- Qualité  $\neq$  conformité
- Maîtrise de la qualité  $\neq$  assurance de la qualité
- La sur-qualité est de la non-qualité
- Triangle coût – qualité - délais

# La qualité en statistique

- Définition de la qualité
- Difficulté à juger ici sur la base du « produit »
- 6 composantes :
  - Pertinence
  - Précision
  - Délais
  - Accessibilité, clarté
  - Comparabilité
  - Cohérence
  - ... et le problème de la charge d'enquête

# La précision en statistique

- Les composantes de l'erreur :
  - Erreur due au sondage
  - Erreur de classification
  - Erreur due à la non-réponse
  - Erreurs dans les données collectées
  - Erreurs dans les traitements élémentaires (saisie, codage)
  - Sur-couverture, sous-couverture

## Qualité des données : critères

- Relecture des critères utilisés en statistique : pertinence, précision, délais, comparabilité, cohérence, lisibilité- accessibilité, complétude
  - Tous peuvent être relus différemment lorsqu'on parle de qualité de données en général, notamment le critère « précision » qui devient « fiabilité »
- On passe inévitablement à la sémantique : les données capturent un sens mais ne sont pas données ; la notion de justesse pose problème
- Autres aspects : conformité, confidentialité, préservation dans le temps

## Qu'est-ce qu'une donnée correcte?

- On peut facilement repérer une donnée non-valide syntaxiquement (non-appartenance au domaine)
- Pas de référentiel absolu pour tester la validité d'une donnée ou d'un ensemble de données
- Possibilité de détecter des incohérences entre données, mais :
  - Si les données A et B sont incohérentes entre elles, rien ne permet de savoir laquelle est bonne
  - Comment assurer que les sémantiques sont les mêmes?
- D'où la question du sens → mode de construction : tout est dans l'élaboration des données

## Construction de l'information : les étapes

- Etapes : saisie, codage, contrôle, calculs, agrégations, reconnaissance de formes, ...
- Au départ, information primaire : telle qu'elle provient de l'extérieur (enregistrement d'un événement, récupération de données externes, ...)
- Niveaux intermédiaires : mises en cohérence de données de différentes sources, transformations purement techniques (fichiers → BD), individuelles ou agrégées, déterministes ou non
- Produits informationnels finaux : utilisables en tant que tels, non pas maillons mais produits

## Construction de l'information : les transformations

- A chaque transformation, perte d'information délibérée pour assurer parallèlement un gain en information (synthèse, meilleure lisibilité)
- Pour chacune d'elles : choix d'une représentation, choix traitement automatique / intervention manuelle, adjonction d'erreurs potentielles
- En permanence, décisions ayant une implication sur les coûts et les aspects de la qualité (fiabilité, lisibilité, comparabilité, cohérence, délais)
- Coûts : de l'erreur, de sa correction, de prévention
- Problème paradoxal de la facilité à fabriquer un pseudo - produit informationnel

# Conclusion

- Les données ne sont pas données : elles résultent d'une élaboration, parfois lente, dont le processus est méconnu
- En cas de production de données de masse, on ne peut industrialiser complètement le processus
- Les erreurs dans les données sont inévitables, elles sont intrinsèques
- La maîtrise de la qualité et du sens des données est étroitement liée à la maîtrise du processus en question, dans lequel les référentiels jouent un rôle central
- Toute étape de la transformation des données est à la fois un gain potentiel et une perte de maîtrise potentielle.
- A chaque étape, on peut avoir à arbitrer (coût-qualité) entre traitements manuel et automatique