RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

Drees
DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES

Évaluation de la qualité de nouvelles données : le Dispositif de Ressources Mensuelles

Séminaire de Méthodologie Statistique & Sciences des données de la DMCSI
Données administratives et statistique : concepts et mises en œuvre

Bureau redistribution et évaluation

Sommaire

1. Introduction au dispositif de ressources mensuelles (DRM)

Le projet ERFS – DRM

- ERFS : individus du 4e trimestre de l'enquête emploi en continu enrichis avec
  - les fichiers fiscaux (déclaration de revenus et du foncier bâti)
  - les prestations sociales versées par les principaux organismes
- Données socio-fiscales annuelles...
- … qui sont ensuite ventilées mensuellement pour simuler l'éligibilité / le recours à des prestations chaque mois dans le modèle Ines
- Appariement de l'ERFS au dispositif de ressources mensuelles (DRM) pour la première fois pour l'ERFS 2020

Présentation du DRM

- Le DRM consiste en l'assemblage de la DSN (déclaration sociale nominative) non redressée et du PASRAU (passage pour les revenus autres)

- La DSN est très connue et utilisée dans la statistique publique (Base Tous Salariés à l'Insee, SISMMO à la Dares) : possible de capitaliser sur l'expérience des utilisateurs du SSP

- PASRAU n'a jamais été utilisée dans le SSP : expertise qualité nécessaire

- Données remontées avec une granularité mensuelle (contre annuelle dans l'ERFS) : amélioration à venir de certains calculs dans le modèle Ines… à condition que les données soient de bonne qualité

Contenu du DRM (1/2)

- Nombreuses origines de ressources disponibles
  - Traitements et salaires
  - ASS - Allocation de solidarité spécifique
  - RSA - Revenu de solidarité active
  - ASPA - Allocation de solidarité aux personnes âgées
  - IJ maternité – Indemnités journalières maternité (paternité et adoption inclus) subrogées
  - …

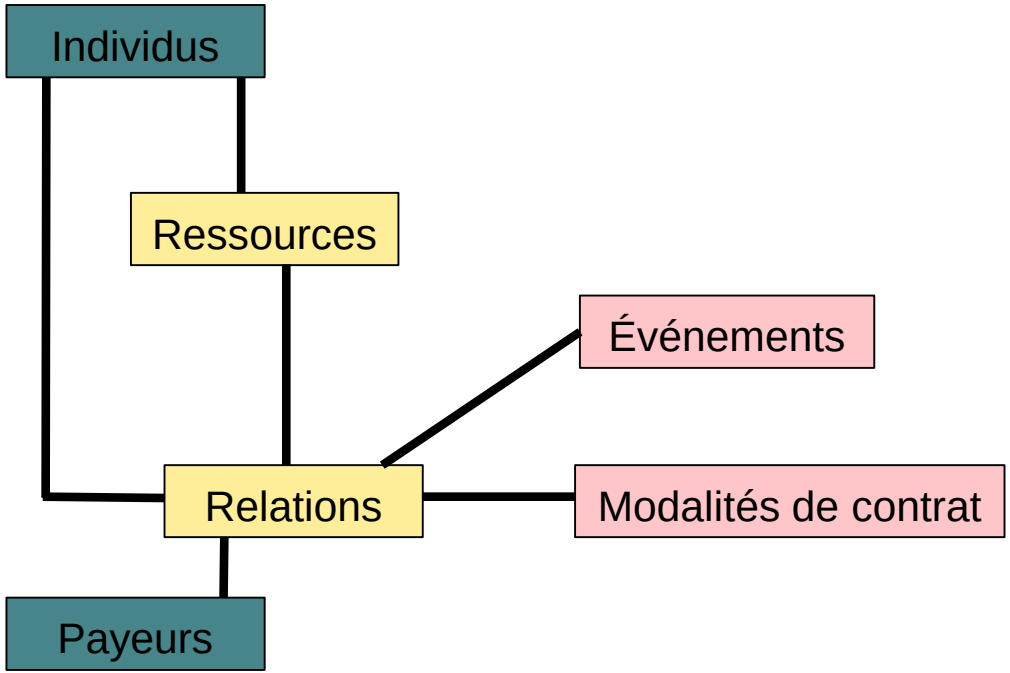
- Détail des ressources perçues par les individus par origine de revenus
  - Revenus nets versés
  - Primes et indemnités brutes par catégories (indemnités de licenciement, indemnité mensuelle de technicité, …)
  - Revenus bruts (salaire de base, heures supplémentaires, jours de RTT monétisés…)
  - Autres éléments de revenus bruts (avantages en nature (repas, logement…), droit d'auteur…)

Contenu du DRM (2/2)

- Événements liés au contrat de travail
  - Arrêt de travail par motif (maladie, maternité, temps partiel thérapeutique…)
  - Fin de contrat par motif (fin de contrat d'apprentissage, licenciement pour faute grave, fin de mission d'interim…)
  - Autre suspension de l'exécution du contrat par motif (chômage intempéries, congé bilan de compétences)

- Modalités du contrat de travail
  - Nature du contrat (CDI, CDD, contrat de mission d'intérim…)
  - Dispositif politique du contrat (contrat d'apprentissage du secteur public, emploi d'avenir…)
  - PCS
  - Quotité de travail
  - …

Dessin de base simplifié



2. L’analyse de la qualité des données

Une source très riche

- Beaucoup de ressources d'origines et de types différents
- Nombreuses lignes
  - 33 millions de lignes dans la table de ressources
  - 2,8 millions de lignes dans la table de relations
  - alors qu'il n'y a que 75 000 individus
- Comment quantifier les erreurs et identifier celles à traiter en priorité ?
- Comment évaluer la qualité de la base ?

Un triple niveau de vérifications

- Cohérence macro : comparaison des estimations de masses de prestation versées avec des cibles externes

- Cohérence micro : comparaison des données annuelles de l'ERFS avec les données mensuelles agrégées par année du DRM au niveau individuel

- Cohérence interne : les versements entre un payeur et un individu suivent souvent un cycle

Identification des erreurs au niveau micro

- Sélection de 100 individus par sondage aléatoire simple
- Étude détaillée de ces 100 individus pour identifier toutes les anomalies repérables au niveau micro pour ces individus
- Possibilité d'estimer la fréquence de chaque type d'erreur, nécessaire pour la priorisation
- Construction d'une base redressée à la main pour les 100 individus, qui a été comparée sur ce champ à la base complète redressée par des fonctions
  - Identification du taux de données non redressées (faux négatifs)
  - Identification du taux de données redressées à tort (faux positifs)
- Construction d'une table annexe avec le motif de chaque redressement manuel
- Calcul d'indicateurs sur la qualité des redressements à chaque étape des redressements

Exemple d'évaluation de la qualité des redressements

- Les relations issues de Pasrau génèrent de trop nombreuses lignes pour un même droit
  - Sur 5 019 relations dans la table relations issue de 100 individus, il n'y a en réalité que 584 droits ou contrats
  - Après redressement
    - 150 droits ou contrats sont correctement reconstitués sur une cible de 189 droits ou contrats : 21 % de faux négatifs
    - 1 droit ou contrat est redressé à tort : 0,2 % de faux positifs
  - Algorithme de redressement conservatif

- Les relations issues de Pasrau ont régulièrement des dates de début erronées
  - Sur 584 droits ou contrats, 156 dates de début de droit ou de contrat ne correspondent pas à la date réelle de début
  - Après redressement
    - 129 dates de début sont correctement redressées : 17 % de faux négatifs
    - 6 dates de début ont été imputées à tort : 1 % de faux positifs

3. Le choix des objets statistiques

Le choix de l'architecture de la base

- Choix du DRM d'assembler toutes les ressources dans la même table : empilement de ressources brutes, nettes, en nature, etc.

- Que doit faire le statisticien ?
  - Conserver une table qui assemble des lignes qui ne sont pas de même nature ?
  - Constituer une table par type de revenus ?
  - Où arrêter la séparation des ressources en différentes tables ?
    - Si on constitue une table de revenus bruts, doit-on séparer salaire de base et heures supplémentaires ?

- Choix de conserver toutes les ressources dans la même table pour la phase de redressements
  - Pratique pour l'application de correctifs et la détection d'erreurs

Un risque d'erreur accru et une nécessité de maîtriser le contenu de la base

Exemple simplifié de la table de ressources

Identifiant de ressource	Date de début de la ressource	Date de fin de la ressource	Catégorie	Type de ressources	Origine de la ressouce	Montant	Identifiant de relation
1	2020-09-01		Net	Net fiscal	Retraite	392	rel_001
2	2020-09-01		Net	Net versé	Retraite	379	rel_001
3	2020-09-01	2020-09-30	Brut	Non catégorisé	Retraite	424	rel_001
4	2020-10-01		Net	Net fiscal	Retraite	392	rel_001
5	2020-10-01		Net	Net versé	Retraite	379	rel_001
6	2020-10-01	2020-10-31	Brut	Non catégorisé	Retraite	424	rel_001
…							

La somme des montants nets n'a pas de sens économique

La colonne date de début a une signification différente selon la catégorie de montants

RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

Drees
DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES

Merci pour votre attention !

Bureau redistribution et évaluation