

Mise au point de critères d'arrêt pour la chaîne de traitement EPURE

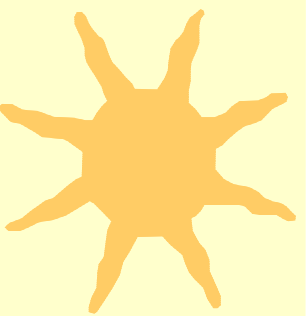
N. Caron et S. Grcic, Insee DSDS

Présentation aux Journées de Méthodologie Statistique

Mercredi 16 mars 2005

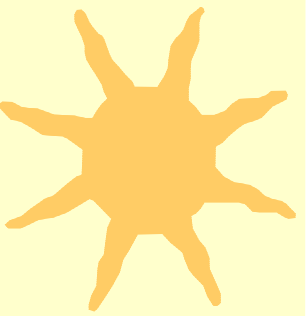


I. La problématique



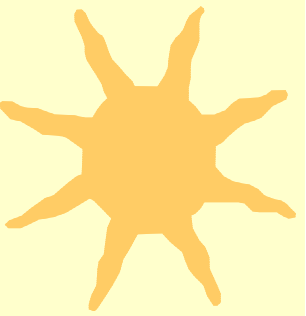
★ Traitement d'une source administrative

- La source : bordereaux récapitulatifs de cotisations recueillies par URSSAF
- Réception trimestrielle des BRC par l'Insee
 - objectif : suivi conjoncturel d'évolution de l'emploi (régional et national)



★ Les contrôles :

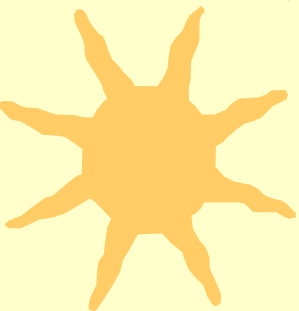
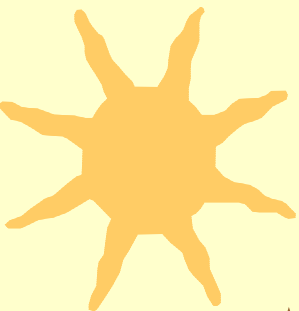
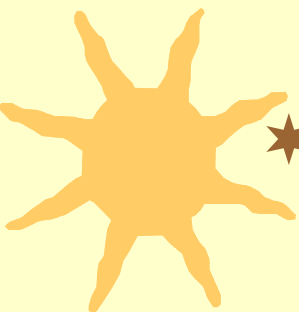
- Chaîne de traitement « automatisée » EPURE
- Plus précisément, « le redresseur » :
 - détecte des anomalies
 - et propose des données corrigées
- Les macro-contrôles





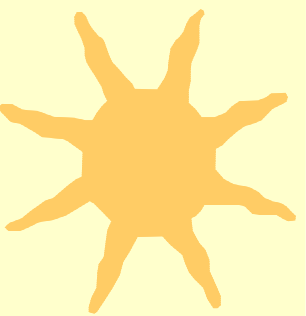
I. La problématique (suite)

- ★ Les variables modifiées par le redresseur sont vérifiées par des gestionnaires
 - 75 gestionnaires EPURE répartis dans toutes les DR
 - 10 semaines
 - **OR, impossible de traiter toutes les anomalies :**
 - 1 900 000 établissements et 350 000 anomalies (18%)
- ★ Organisation actuelle du travail :
 - Objectif national : les établissements > 50 salariés prioritaires
 - Objectif régional propre à chaque DR
 - 30% des anomalies traitées

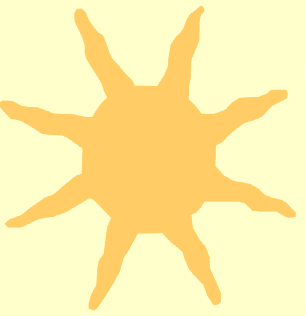




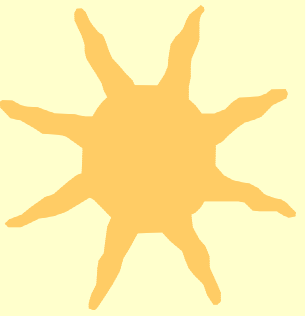
I. La problématique (suite)



⇒ lancement d'une opération « de qualité »
sur la chaîne de production EPURE.



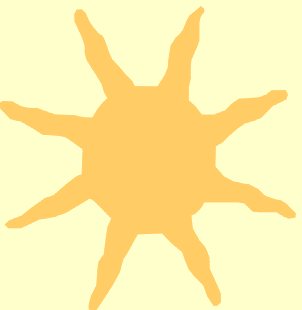
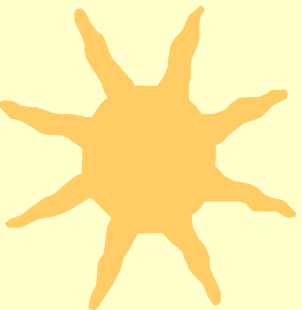
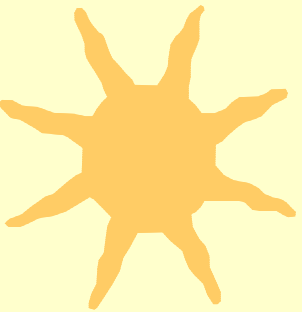
objectif : « rationaliser » l'expertise en
direction régionale



- rendre plus efficaces les traitements des gestionnaires
- et déterminer un seuil (critère d'arrêt) en delà duquel on peut considérer que l'on fait de la « sur-qualité »



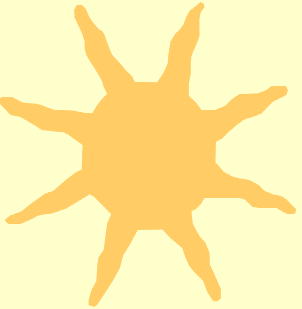
II. Une démarche en 2 étapes



- déterminer un ordre de classement des établissements par anomalie
- définir un seuil d'arrêt des traitements et déterminer la « qualité » obtenue



II.1. Classement des établissements



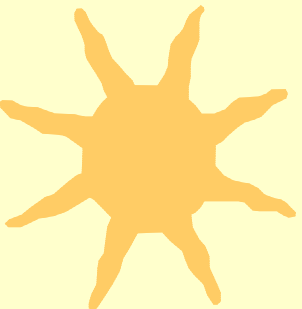
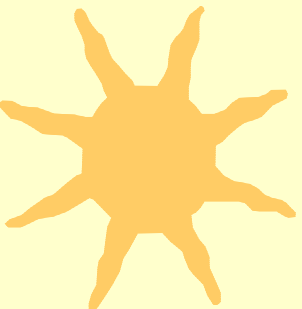
★ **Idée** : Modéliser l'écart entre l'effectif « redresseur » et l'effectif « vrai »

– Régression linéaire sur $(\text{Eff}_{\text{redresseur}} - \text{Eff}^*)$

avec $\text{Eff}^* = \text{valeur « vraie »}$

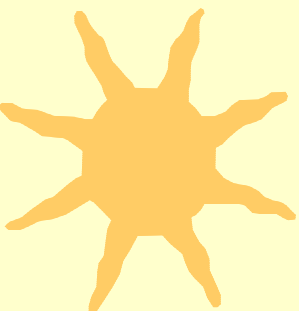
– Variables :

- note du redresseur, taille, taux couverture
- contribution de l'établissement à l'évolution





II.1. Classement des établissements



★ Les problèmes

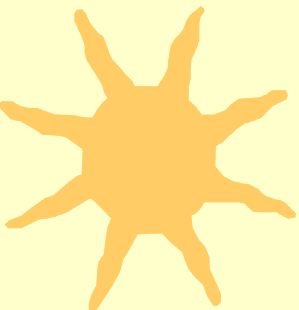
– notion d'effectif « vrai » ?

→ le test « qualité »

- T3 2003 : Bretagne, Haute-Normandie, Picardie, Midi-Pyrénées (1 dépt)
- T4 2003 : les mêmes... + Limousin, Pays de Loire (2 dépts), Poitou-Charentes et Centre (2 secteurs)

– Quel niveau?

- NES16 * dépt (1368 « cases » France entière)





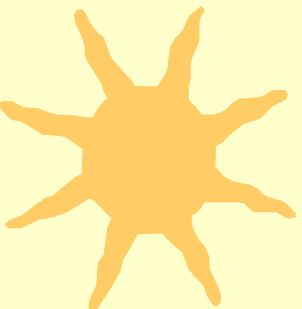
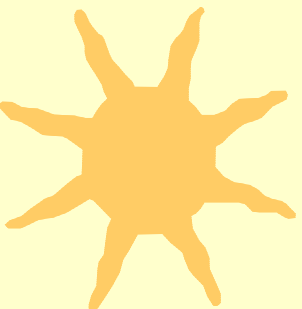
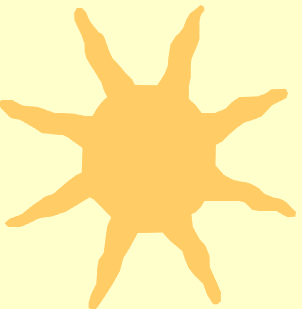
II.2. Les critères d'arrêt

★ Plusieurs approches possibles

- fixer un seuil S^* pour quantifier l'erreur admissible, selon :

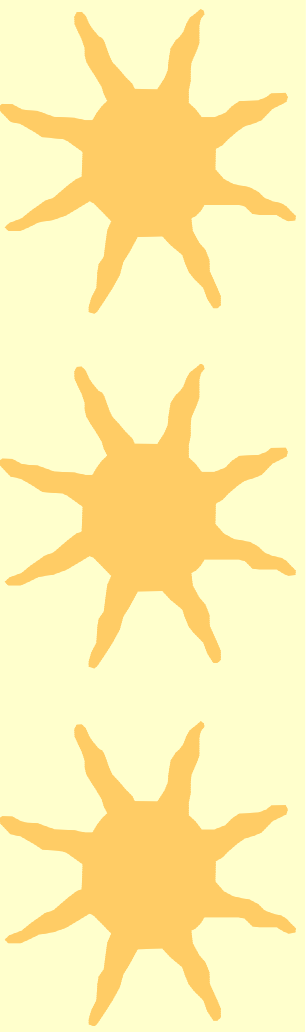
$$\frac{\left| \sum_{NV} (Y_i - Y_i^*) \right|}{\sum_{i=1}^N Y_i^*} \leq S^*$$

- où N = taille totale de la population.
- NV = ensemble des établissements non vérifiés
- Arrêt des contrôles dès que inégalité satisfaite (pas d'impact a priori des établissements non corrigés)





III. *Une mise en pratique pas si simple!!*



- ★ Impact du travail des gestionnaires sur l'indice d'évolution de l'emploi ...

Évolution T/T-1 brute	Évolution T/T-1 après redressement automatique	Évolution T/T-1 après reprise manuelle
-1,9	-0,1	-0,3

- ★ ... pour des effectifs modifiés pour seulement 12% des établissements en anomalie



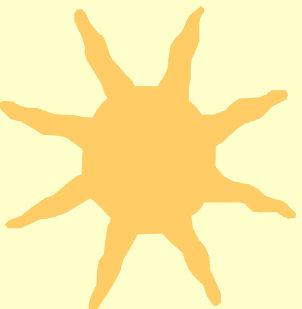
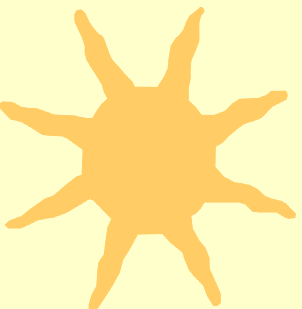
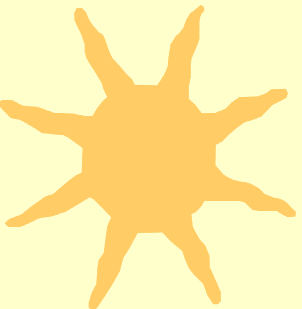
Classement issu de la régression

- ★ Un ordre de classement insuffisamment efficace
 - Dans 88% des cas le gestionnaire confirme la proposition du redresseur => 88% de valeurs nulles

T3-2003					
	Indice initial avant travail manuel	Indice final avec 100 % des anomalies traitées	Indice avec 25 % des anomalies traitées	Écart	Écart en effectif
Ensemble des régions du test	-0,09	-0,31	-0,21	0,10	2 131



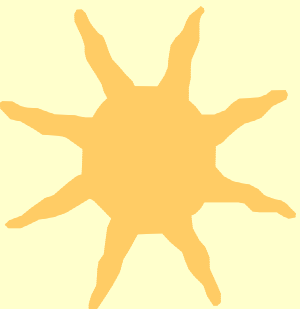
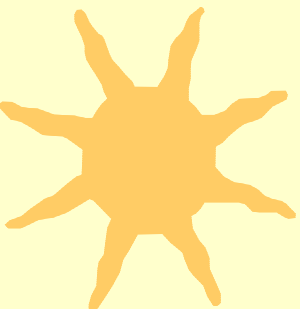
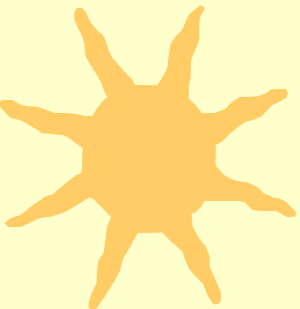
Poursuite des travaux ...



- ★ En classant les établissements d'une case Département * Nes16 :
 - principalement selon la contribution à l'indice ;
 - l'évolution T/T-1 ;
 - la note (en tranche) ;
 - la taille (en tranche) ;
 - l'indicateur d'anomalie sur les effectifs du 3ème mois.



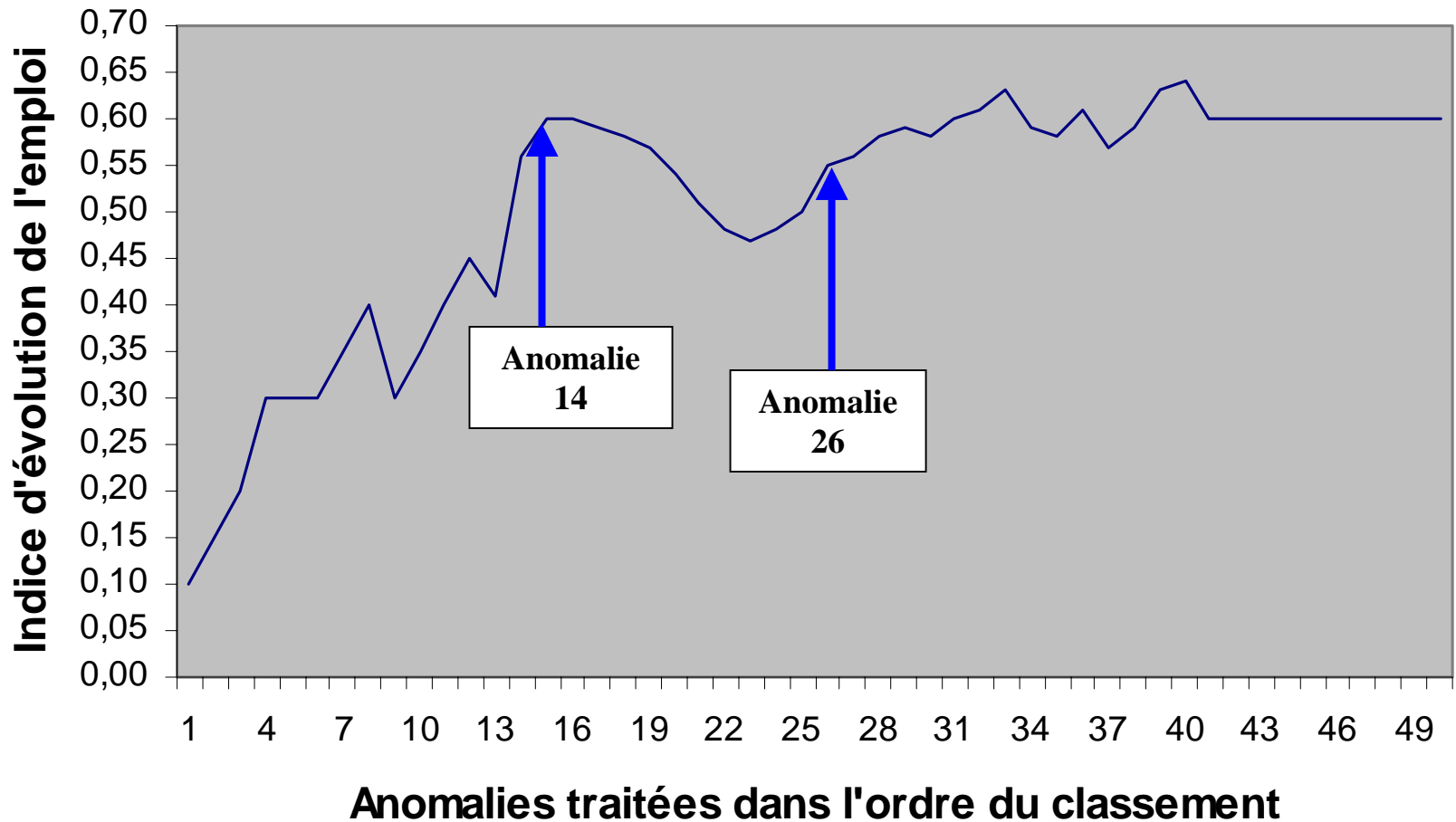
Construction des critères d'arrêt



- ★ Définir l'écart maximal d'indice acceptable et mesurer où cela conduit à arrêter le traitement
- ★ Cela conduit à mettre en pratique le critère d'arrêt exposé précédemment avec $S^* = 0,05$



Construction des critères d'arrêt



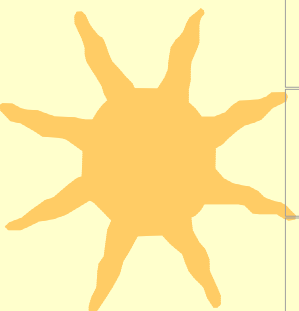
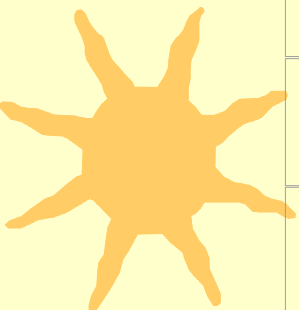
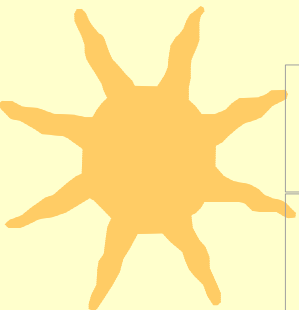


Construction des critères d'arrêt

Classement issu :	de la régression		du classement direct	
	% d'anomalies à traiter pour un indice identique	% d'anomalies à traiter pour un indice proche à moins de 0,05 point	% d'anomalies à traiter pour un indice identique	% d'anomalies à traiter pour un indice proche à moins de 0,05 point
T3 2003				
Total régions du test	90,4	37,5	80,8	27,4



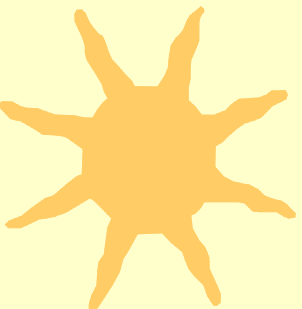
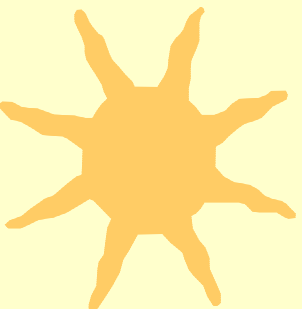
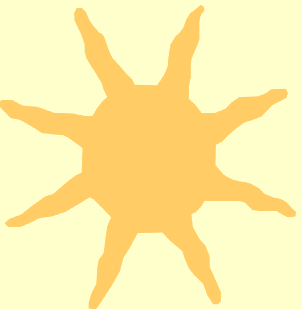
Construction des critères d'arrêt



	T3-2003
Classement direct	Dispersion du % d'anomalies à traiter par « case »
Max	100,0
75%	51,1
50%	10,6
25%	4,1
Min	0,0



En conclusion



- ★ Sur plusieurs trimestres :
 - Le « meilleur » classement se vérifie ;
 - Les critères d'arrêt par case « ne sont pas stables ».
- ★ Encore toute une réflexion à mener pour la mise en production