

JMS 2012

Qualité du codage de l'activité et de la profession dans le recensement de la population

Elodie MARTAL – Yves JACQUIN
Insee-DSDS



PLAN DE LA PRESENTATION

1. LE CODAGE DE L'ACTIVITE ET DE LA PROFESSION DANS LE RP
2. LA MESURE DE LA QUALITE
3. LES PRINCIPAUX RESULTATS DES OPERATIONS QUALITE de 2006 à 2009
4. QUELS ENSEIGNEMENTS EN TIRER?

1. LE CODAGE DE L'ACTIVITE ET DE LA PROFESSION DANS LE RP (1)

Le processus de codification :

Traitement automatique : de mai à début septembre en fonction de la saisie des questionnaires

Traitement centralisé au Centre national informatique d'Orléans

Traitement de reprise manuelle des échecs de codage pour trois variables (activité/établissement employeur, profession actuelle, profession antérieure)

120 gestionnaires environ dans toutes les Directions régionales de l'Insee, sauf en DOM

1 800 000 libellés à coder

Application RECAP : REcherche et Codage de l'Activité et de la Profession

1. LE CODAGE DE L'ACTIVITE ET DE LA PROFESSION DANS LE RP (2)

La reprise manuelle :

Activité

→ Codage à plusieurs niveaux possibles, a minima au niveau division (2 positions) de la nomenclature d'activité NAF (88 postes)
Sinon : codage à blanc → redressement

Profession

→ Profession actuelle : niveau détaillé uniquement de la nomenclature des professions et catégories socioprofessionnelles PCS (486 postes d'actifs)
sinon codage à blanc → redressement

2. LA MESURE DE LA QUALITE (1)

- **Opération spécifique RECAP QUALITE de mesure de la qualité du codage de l'activité et de la profession,
→ à la fois pour le processus automatique et le processus de reprise en DR avec RECAP**
- **Mise en place en 2006
et renouvelée chaque année depuis (sauf en 2010 pour raisons budgétaires)**
- **S'appuie sur une version spécifique de l'application RECAP,
dénommée RECAP Qualité**

2. LA MESURE DE LA QUALITE (2)

1^{ère} phase : Le 2^{ème} codage

sur un échantillon d'environ 100 000 individus

8 à 9 Directions régionales de l'Insee (DR) « volontaires » entre novembre et décembre

Environ 30 gestionnaires pendant 1 mois et demi, sur une période creuse en DR pour le RP

Conditions identiques à celles de la campagne RECAP

Les gestionnaires de RECAP Qualité codent un échantillon de bulletins individuels sans connaître le premier codage

A l'issue de cette phase : si le premier et le deuxième codages sont identiques **au niveau le plus fin**, on parle de convergence, dans le cas contraire de divergence.

2. LA MESURE DE LA QUALITE (3)

2ème phase : l'arbitrage

Réalisé en janvier et février de l'année suivante dans 3 DR (environ 15 gestionnaires experts)

Expertise de tous les codages divergents et de 10% des codages convergents (sauf en 2006 où les codages convergents n'avaient pas été arbitrés)

L'arbitre connaît les deux premiers codages attribués et peut choisir entre ces deux derniers, ou proposer un troisième codage différent.

Il a plus de temps pour coder que les 2 premiers codeurs

Postulat : on considère que le code arbitré est correct

2. LA MESURE DE LA QUALITE (4)

La structure de l'échantillon qualité

Population totale					
	Codage automatique		Codage RECAP		TOTAL
Activité	774 000	43%	1 026 000	57%	1 800 000
Profession	1 368 000	76%	432 000	24%	1 800 000

Echantillon					
	Codage automatique		Codage RECAP		TOTAL
Activité	12 000	24%	39 000	76%	51 000
Profession	6 000	15%	34 000	85%	40 000

- Population totale : les 1 800 000 actifs occupés de l'exploitation complémentaire (1/4 des individus des petites communes, tous les individus des grandes communes d'une enquête annuelle de recensement)
- Echantillon : sur-représentation du codage manuel

2. LA MESURE DE LA QUALITE (6)

Méthode de calcul du taux d'erreur de codage :

1. Stratification de l'échantillon par processus de codage
2. Calcul des taux d'erreur de codage selon le processus et selon le caractère convergents/divergents
→ erreur si code arbitré \neq code initial, selon le niveau de codage considéré
3. Extrapolation à la totalité de l'échantillon (car convergents non tous arbitrés), puis du champ de la population codée selon les clés de répartition par processus (codage automatique/codage manuel)

LES PRINCIPAUX RESULTATS sur la qualité du codage de l'activité et de la profession

3. LES PRINCIPAUX RESULTATS (1)

Taux d'erreur de codage : l'activité en 2009 (niveau division)

Echantillon qualité activité : 51 079 individus

Processus de codage	Echantillon		Echantillon	Taux d'erreur de codage de l'activité %	
Codage automatique	12 000	Convergenents	10 041	0,4	7,8
		Divergenents	1 959	31,6	
Codage manuel	39 079	Convergenents	19 027	3,3	9,5
		Divergenents	20 052	27,9	
Total		des convergenents	29 068	2,3	
		des divergenents	22 011	28,2	
TOTAL ECHANTILLON			51 079		

3. LES PRINCIPAUX RESULTATS (2)

Taux d'erreur de codage : l'activité en 2009 (niveau division)

Population totale à coder : 1 811 000 actifs occupés –
Exploitation complémentaire de l'enquête annuelle de recensement 2009

Processus de codage	Population à coder		Taux d'erreur de codage de l'activité %	Taux d'erreur global %
Codage automatique	798 000	44%	7,8	8,7
Codage manuel	1 013 000	56%	9,5	

3. LES PRINCIPAUX RESULTATS (3)

Taux d'erreur de codage : la profession actuelle en 2009 (niveau CS 42)

Echantillon profession : 39 593 individus

Processus de codage	Echantillon		Echantillon	Taux d'erreur de codage de la profession actuelle %	
Codage automatique	6 000	Convergents	4 842	1,7	7,2
		Divergents	1 158	30,6	
Codage manuel	33 593	Convergents	17 791	3,3	19,1
		Divergents	15 802	36,9	
TOTAL		des convergents	22 633	3,0	
		des divergents	16 960	36,5	
TOTAL ECHANTILLON			39 593		

3. LES PRINCIPAUX RESULTATS (4)

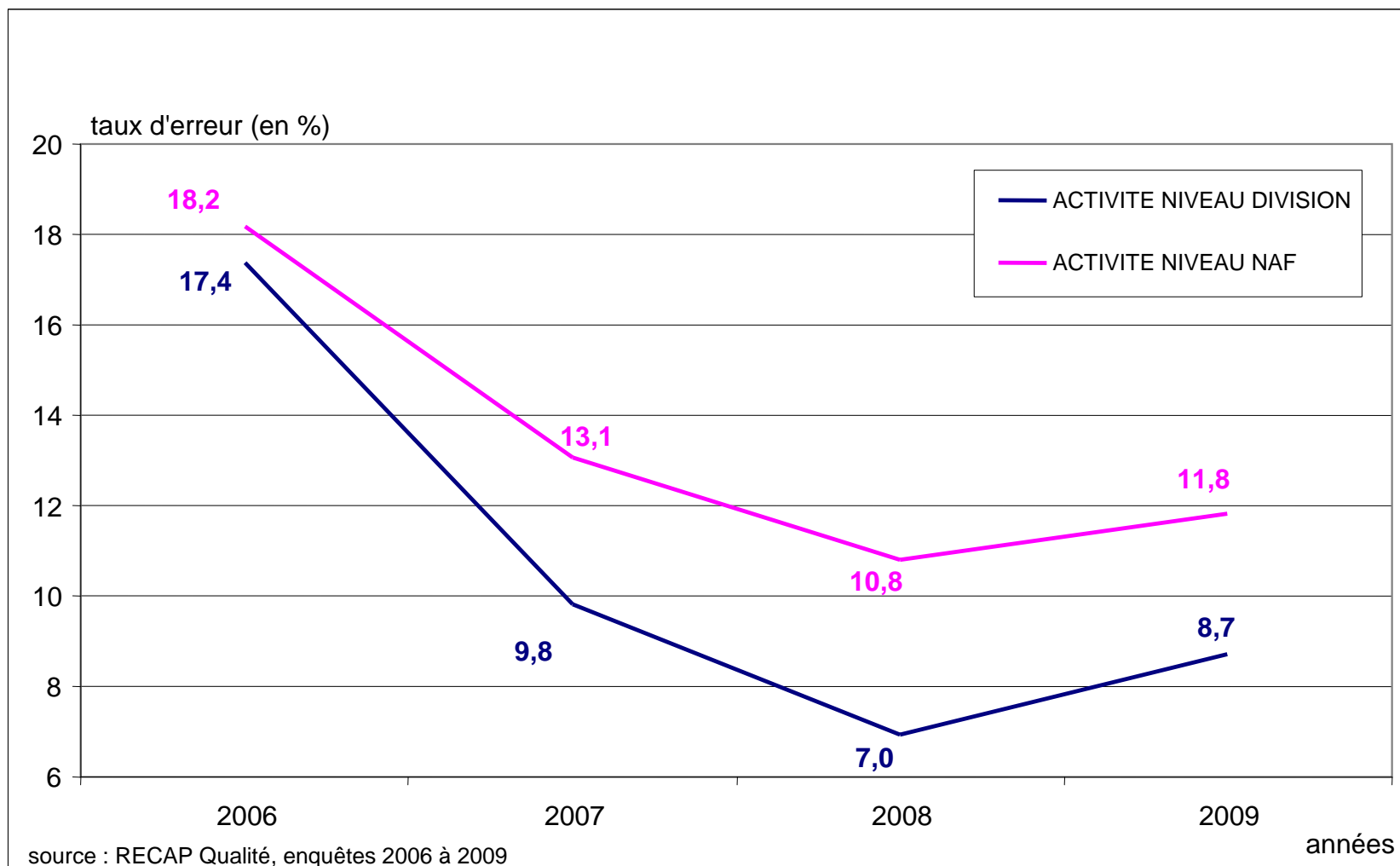
Taux d'erreur de codage : la profession actuelle en 2009 (niveau CS 42)

Population totale à coder : 1 760 000 actifs occupés –
Exploitation complémentaire de l'enquête annuelle de recensement 2009

Processus de codage	Population à coder		Taux d'erreur de codage de la profession actuelle %	Taux d'erreur global %
Codage automatique	1 330 000	76%	7,2	10,1
Codage manuel	430 000	24%	19,1	

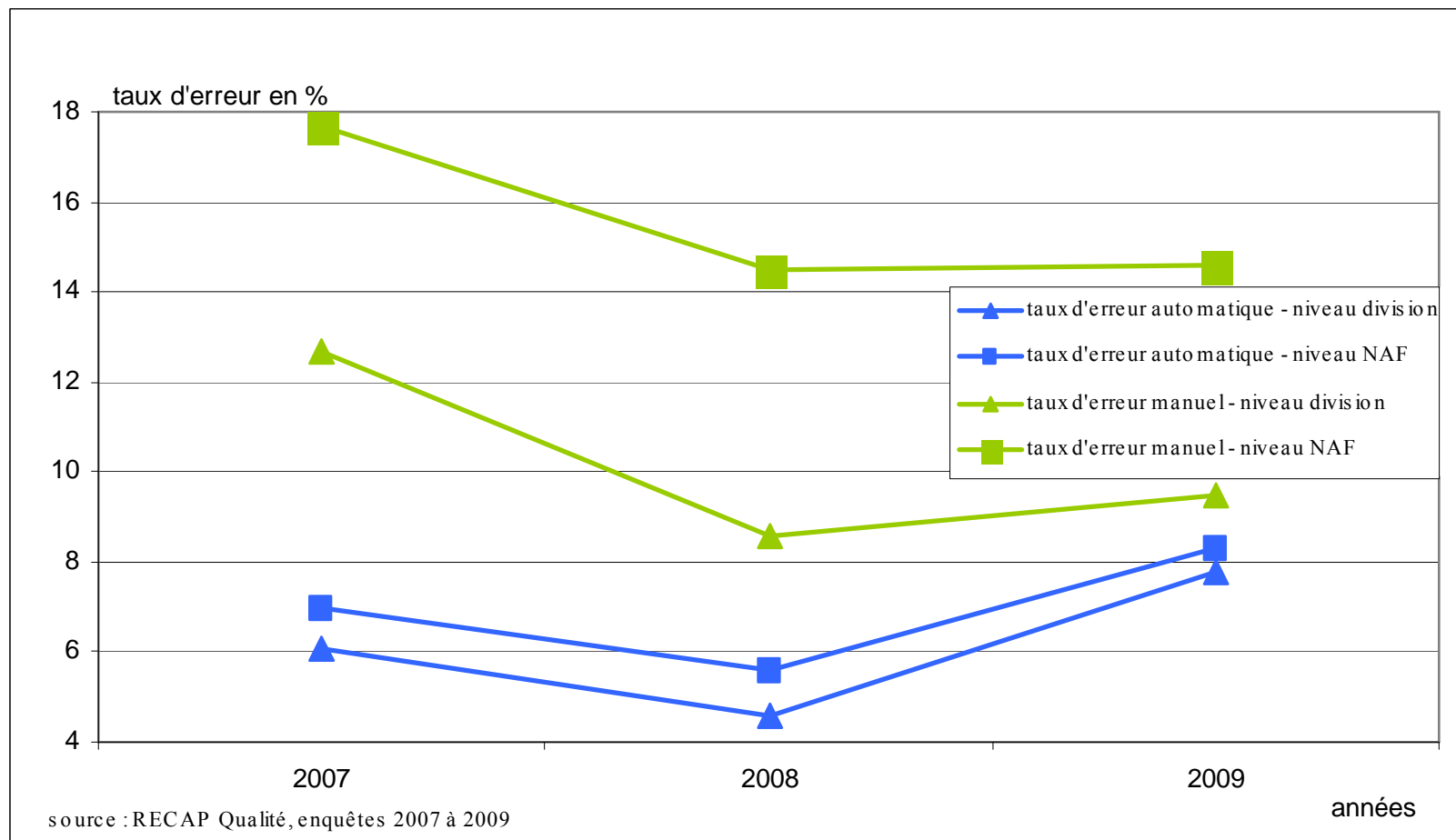
3. LES PRINCIPAUX RESULTATS (5)

Evolution du taux d'erreur de codage de l'activité selon le niveau de nomenclature



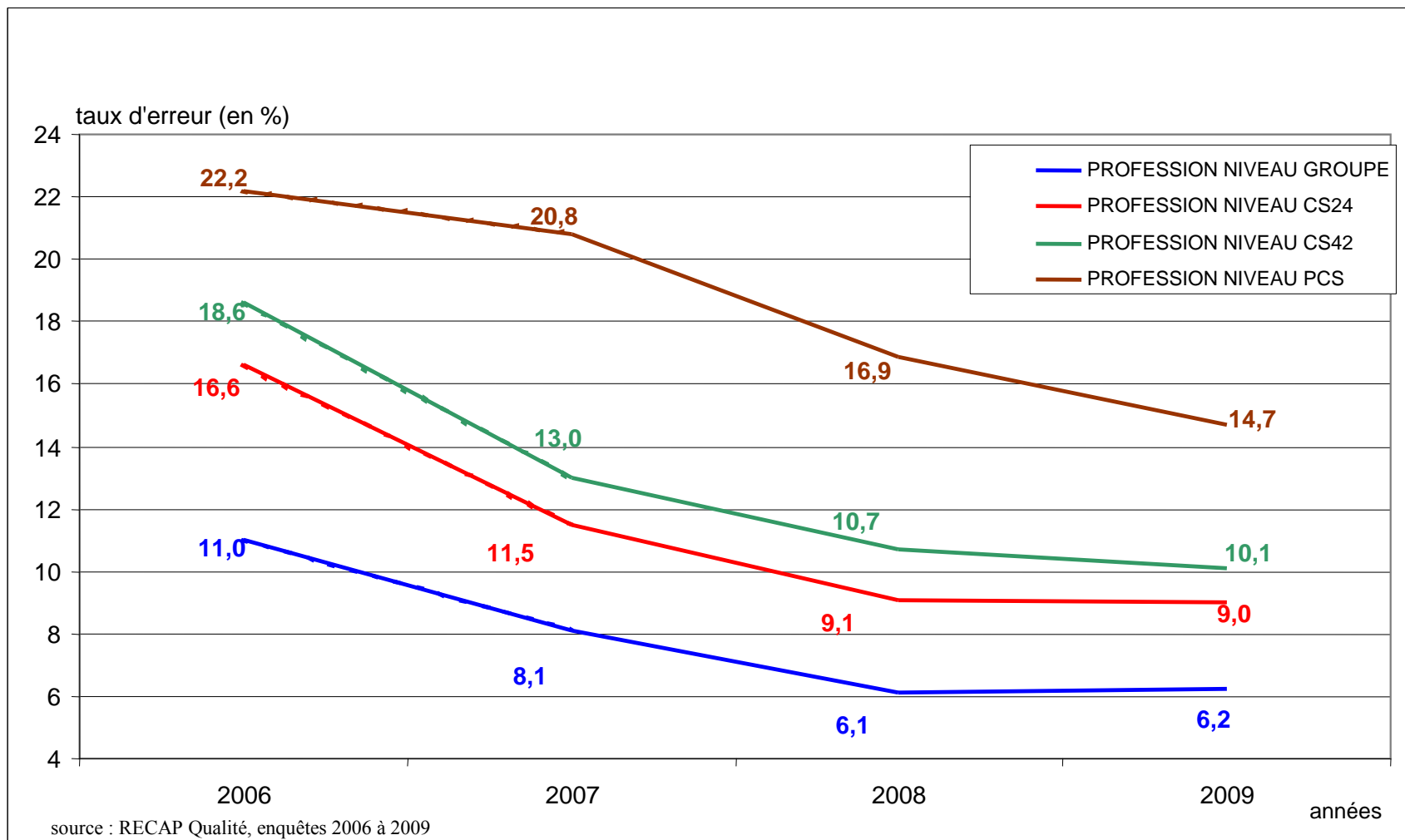
3. LES PRINCIPAUX RESULTATS (6)

Evolution du taux d'erreur de codage de l'activité selon le processus de codage et le niveau de nomenclature



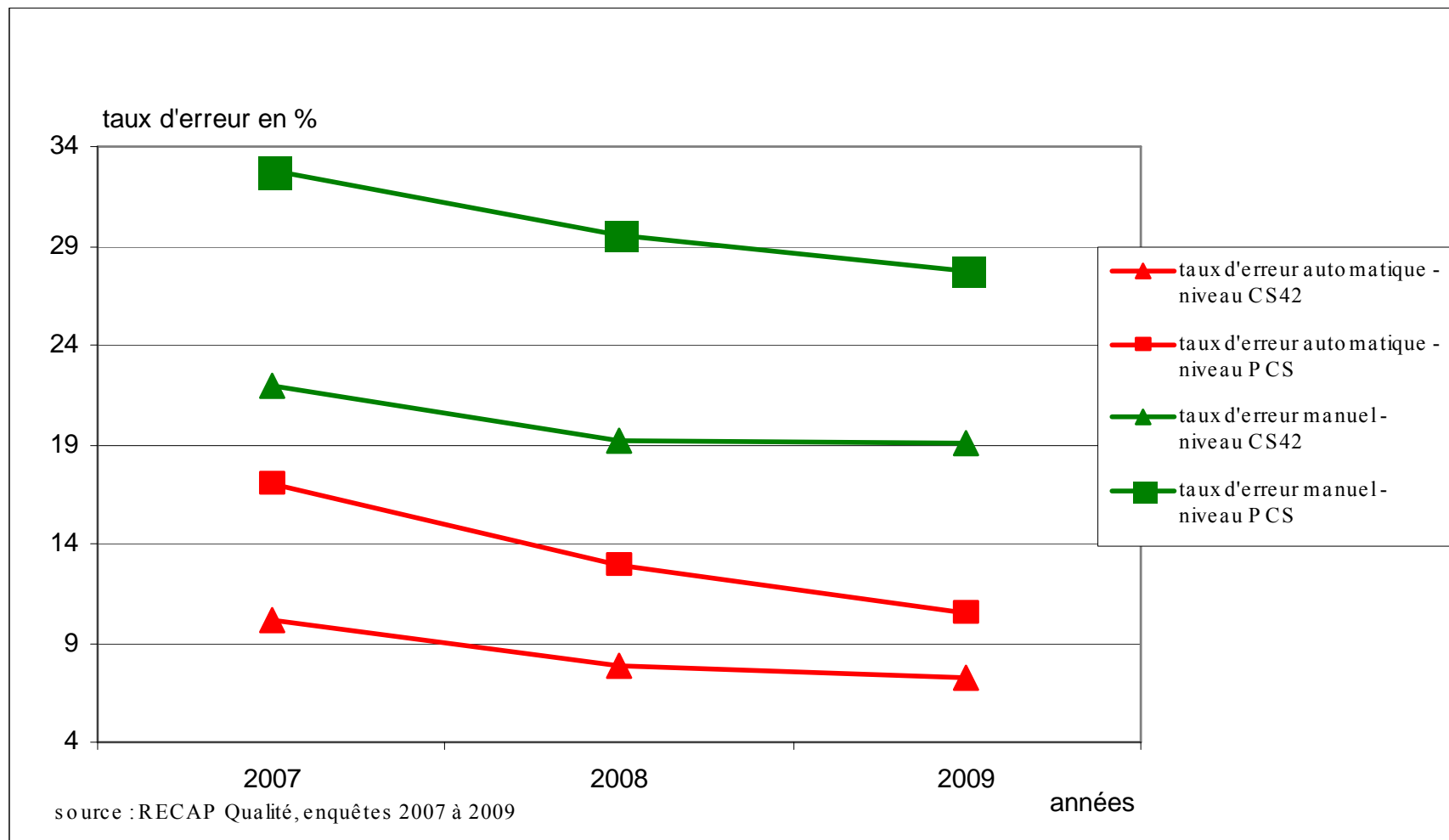
3. LES PRINCIPAUX RESULTATS (7)

Evolution du taux d'erreur de codage de la profession actuelle selon le niveau de nomenclature



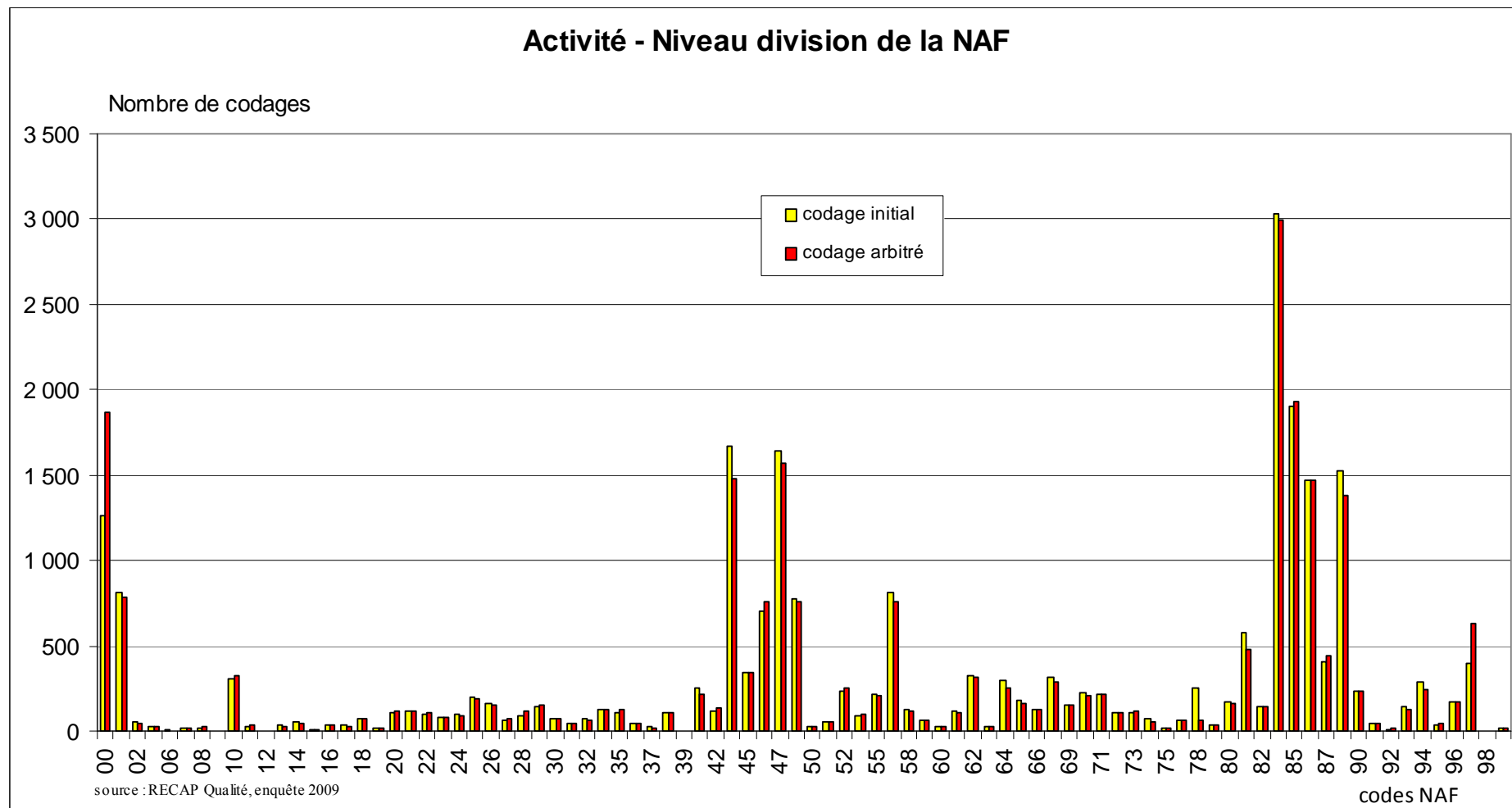
3. LES PRINCIPAUX RESULTATS (8)

Evolution du taux d'erreur de codage de la profession actuelle selon le processus de codage et le niveau de nomenclature



3. LES PRINCIPAUX RESULTATS (9)

Distribution statistique de l'activité en 2009 : codage initial et codage arbitré



3. LES PRINCIPAUX RESULTATS (10)

Distribution statistique de l'activité

Quelques postes de divisions NAF plus fragiles :

Division 43 : travaux de construction spécialisés

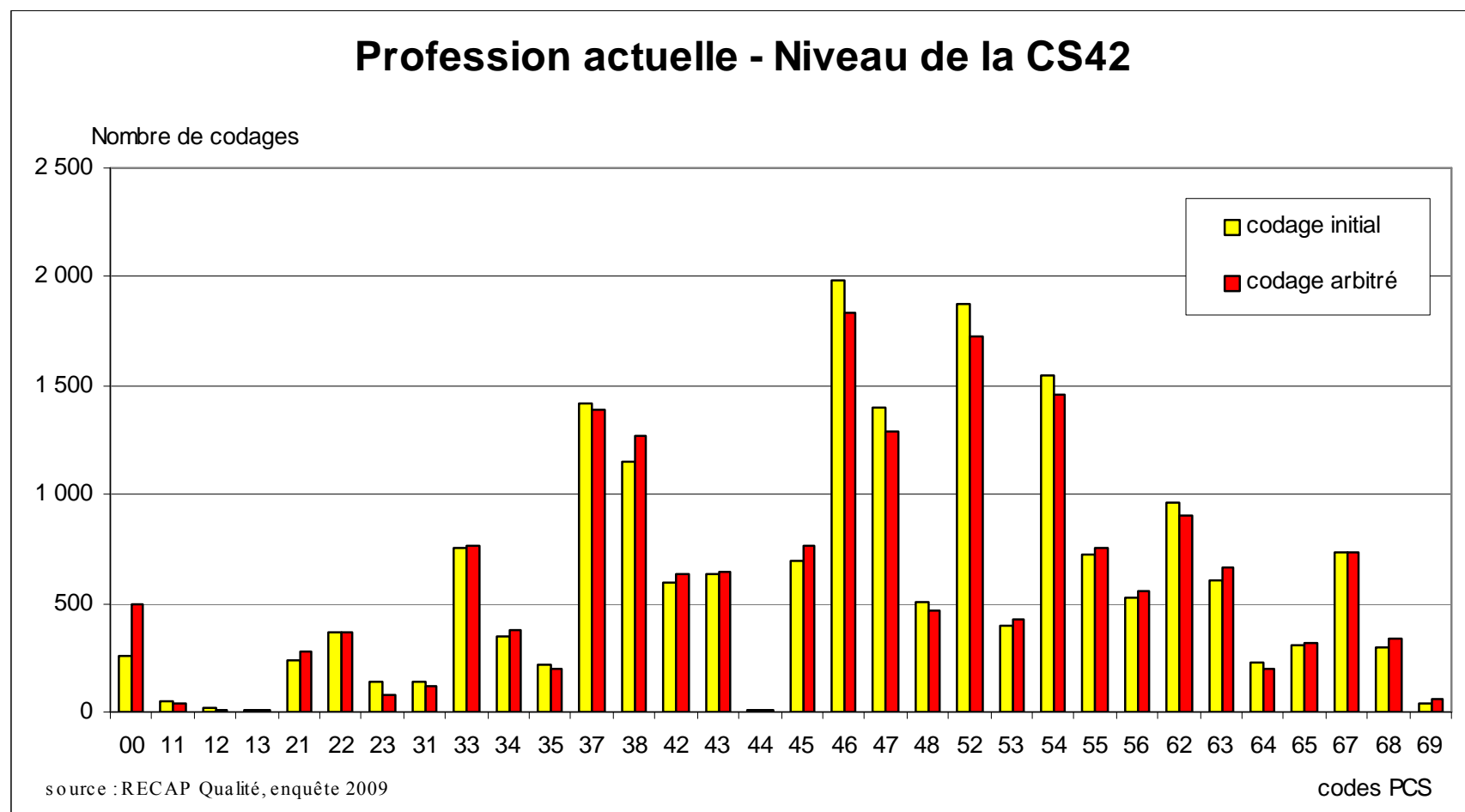
Divisions 46-47 : commerce de gros/de détail, à l'exception des automobiles et des motocycles

Divisions 88-97 : action sociale sans hébergement/activité des ménages en tant qu'employeurs de personnel domestique (dont activités d'aide à domicile)

Effets de transferts croisés entre postes qui atténuent l'impact au niveau macro

3. LES PRINCIPAUX RESULTATS (11)

Distribution statistique de la profession actuelle en 2009 :
codage initial et codage arbitré



3. LES PRINCIPAUX RESULTATS (12)

Distribution statistique de la profession

Quelques postes de CS plus fragiles :

37 (cadres administratifs et commerciaux d'entreprise)
et 38 (ingénieurs et cadres techniques d'entreprise)

46 (professions intermédiaires administratives et
commerciales des entreprises) et 47 (techniciens)

52 (employés civils et agents de services de la fonction
publique)

54 (employés administratifs d'entreprises)

Effets de transferts croisés entre postes qui atténuent l'impact au niveau macro

4. QUELS ENSEIGNEMENTS EN TIRER?

- Des taux d'erreurs de codage un peu élevés, mais de niveaux acceptables et en baisse régulière
→ confirmation attendue de RECAP qualité 2011 après l'interruption de 2010.
- Une distribution statistique corrigée proche de la distribution initiale pour la profession au niveau CS (31 postes d'actifs) et pour l'activité au niveau division (88 postes), mais faisant apparaître quelques postes plus fragiles.
- Une amélioration de la qualité du codage manuel de l'activité et de la profession au fil des campagnes de recensement avec un niveau d'expertise croissant des gestionnaires – mais aussi des arbitres - lié à un affinement des consignes de codage.
- RECAP qualité : un outil pour contrôler et faire évoluer :
 - * les outils de codage automatique (outil de recherche dans SIRENE pour l'activité, SICORE-PCS pour la profession)
 - * Les consignes de codage aux DR dans RECAP

JMS 2012

Qualité du codage de l'activité et de la profession dans le RP

Merci de votre attention !