Pondération de l'enquête sur les revenus et la consommation (ERC): Exemple d'application d'un calage

- Introduction
- Pondération sans calage
- Pondération avec calage
- Sources externes
- Variables de calage
- Méthode de calage utilisée
- Résultats avec calage
- Conclusions

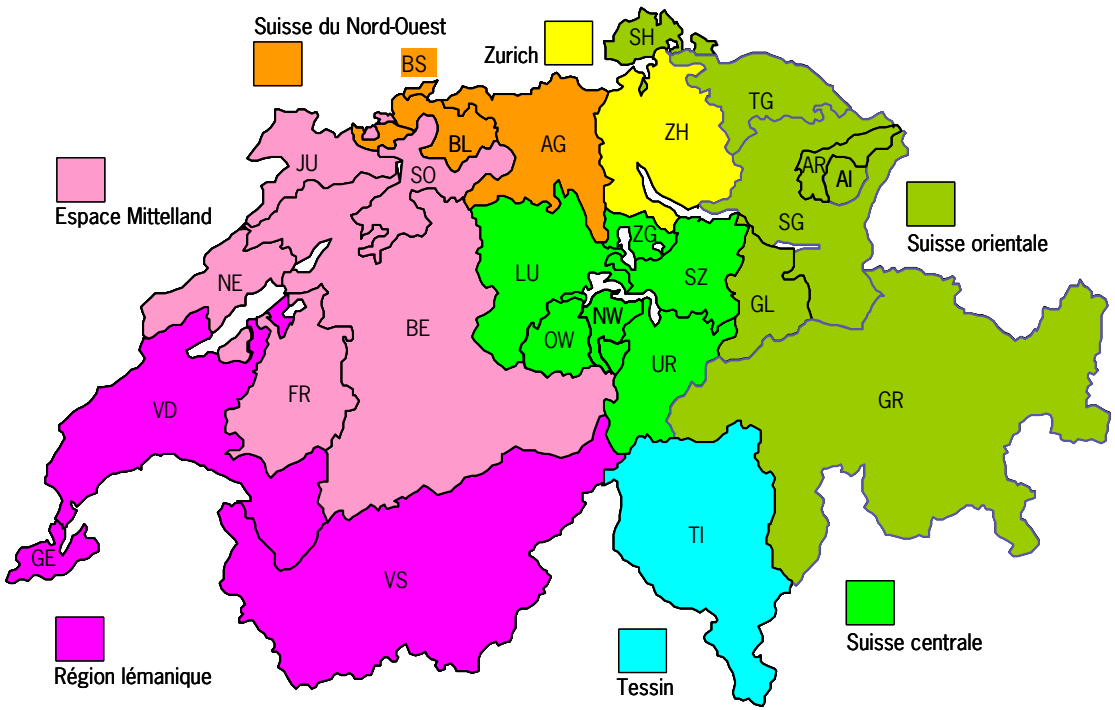
L'enquête sur les revenus et la consommation

- But: estimation des revenus et de la consommation des ménages
- Enquête auprès des ménages privés depuis 1998
- Taille de l'échantillon (à partir de 2005): environ 3000
- Déroulement en 12 vagues mensuelles durant l'année
- Effectuée sous mandat de l'OFS par un institut privé
- Saisie des données dans des carnets, suivi téléphonique (CATI)
- Charge des ménages assez lourde (1 mois)
- Taux de réponse: environ un tiers

Plan d'échantillonnage

- Échantillon stratifié simple: 7 strates
- Tirage des échantillons dans le SRH (annuaire Swisscom)
- Stratification selon les grandes régions:

Région lémanique:
(18.5 %): 17.1 %
Espace Mittelland:
(23.5 %): 21.8 %
Suisse du Nord-Ouest:
(13.8 %): 12.9 %
Zurich:
(18.4 %): 17.1 %
Suisse orientale:
(13.6 %): 12.6 %
Suisse centrale:
(8.1 %): 8.1 %
Tessin:
(4.1 %): 10.4 %



Déroulement de l'enquête (pertes)

1) Lettre envoyée par l'OFS (7 semaines avant le mois)

2) Interview de recrutement (6 à 3 semaines avant le mois):
   - pertes: non atteints, refus, incapacité –> niveau 0
   - pertes: interviews non-réponse –> niveau 1

3) Mois sous revue (période d'inscriptions):
   - pertes: abandons, refus –> niveau 1

4) Saisie, suivi téléphonique:
   - pertes: abandons, refus –> niveau 1

5) Interview finale (2 à 12 semaines après le mois):
   - pertes: abandons, refus, non atteints –> niveau 1
   - ménages répondants –> niveau 2

Taux de réponse de l'ERC

- Taux de non réponse plutôt élevé: nécessité de corriger le biais introduit par la non réponse à l'aide d'un modèle de pondération
- Utilisation de l'information des ménages qui n'ont pas participé à l'enquête mais qui ont donné leurs caractéristiques de base

	ERC 1998*		ERC 2000*		ERC 2001		ERC 2002		ERC 2003	
Adresses activées	30920		12930		12130		12580		11020	
Adresses sans information	15486	50.1%	6868	53.1%	5225	43.1%	5175	41.1%	4339	39.4%
Ménages avec informations	**15434**	**49.9%**	**6062**	**46.9%**	**6905**	**56.9%**	**7405**	**58.9%**	**6681**	**60.6%**
Non participants	*6139*	*39.8%*	*2420*	*39.9%*	*3165*	*45.8%*	*3679*	*49.7%*	*3206*	*48.0%*
Participants	***9295***	***60.2%***	***3642***	***60.1%***	***3740***	***54.2%***	***3726***	***50.3%***	***3475***	***52.0%***

* Pas d'interview non réponse

Le modèle de pondération sans calage

Deux étapes:

1) Probabilité d'inclusion = probabilité pour un ménage d'être tiré
   –> définie par le plan d'échantillonnage: π_k

2) Modèle de non réponse = probabilité pour un ménage de répondre
   –> au niveau 1 (sans modèle): τ_k
   –> au niveau 2 (modèle logistique de non réponse): ρ_k

- Poids:

$$w_k = \frac{1}{\pi_k \cdot \tau_k \cdot \rho_k}$$

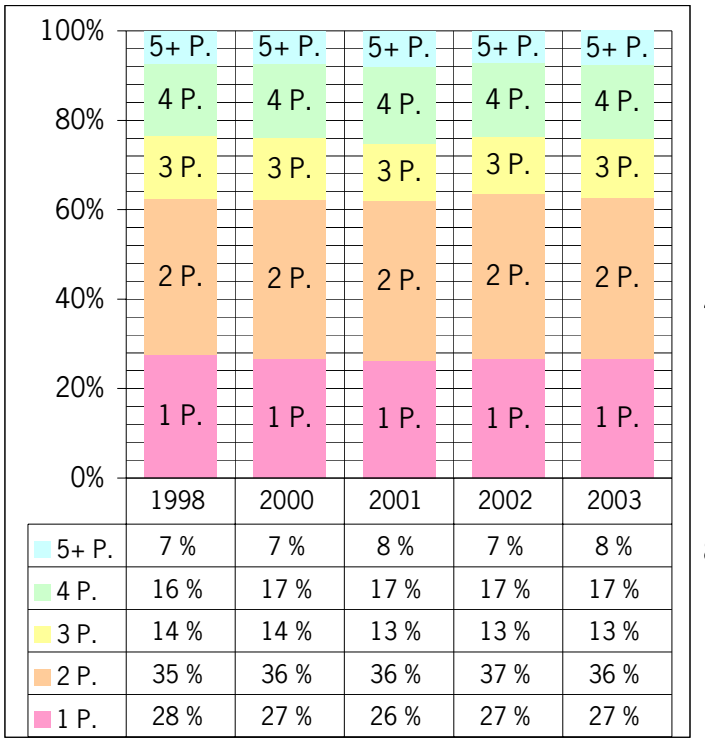
$$\rho_k = \frac{\exp\left(\sum_j x_{jk}\beta_j\right)}{1 + \exp\left(\sum_j x_{jk}\beta_j\right)}$$

- Prise en compte du cadre de sondage et des résultats de l'enquête
- Aucune autre source d'informations utilisée

Sans calage: taille des ménages

- Taille moyenne des ménages de l'ERC pondérée sans calage supérieure à celle observée par le recensement

Taille des ménages



	1998	2000	2001	2002	2003
5+ P.	7 %	7 %	8 %	7 %	8 %
4 P.	16 %	17 %	17 %	17 %	17 %
3 P.	14 %	14 %	13 %	13 %	13 %
2 P.	35 %	36 %	36 %	37 %	36 %
1 P.	28 %	27 %	26 %	27 %	27 %

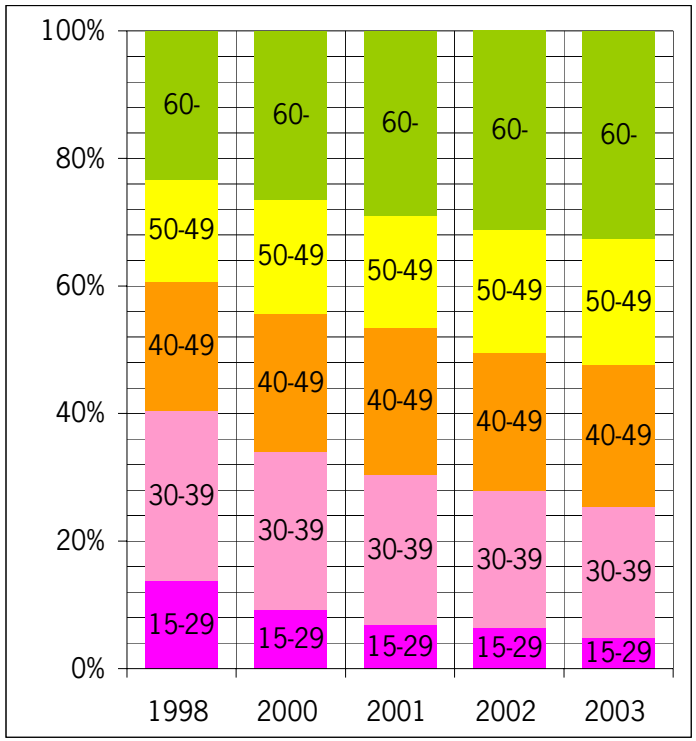
ERC 2003:
taille moyenne 2.45

Recensement fédéral (RFP) 2000:
taille moyenne 2.24

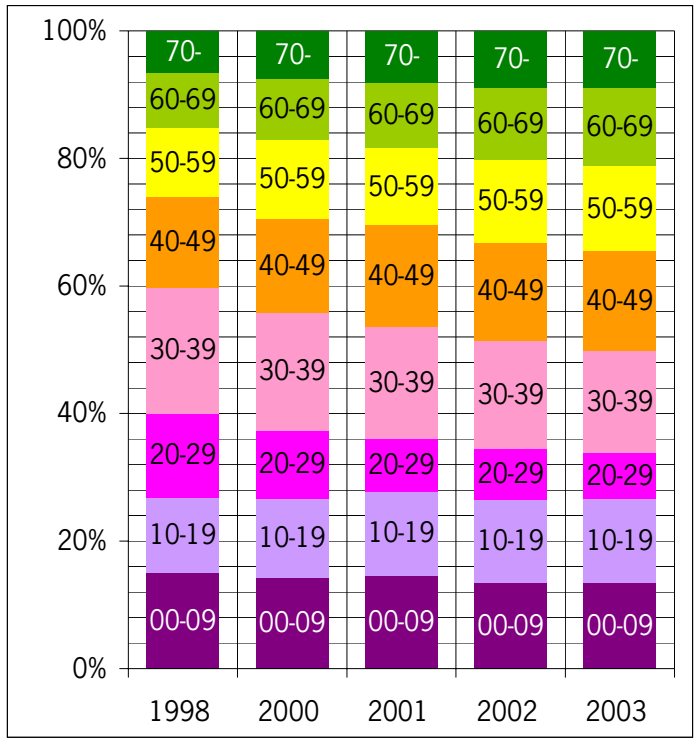
Sans calage: tranches d'âge

- Évolution accentuée dans les tranches d'âge: diminution de la part des jeunes (15-29)

Âge de la personne de référence



Âge des personnes



Pondération: trois premières étapes

Les trois premières étapes:

1) Probabilité d'**inclusion**:
   –> comme le modèle sans calage

2) Correction sur les strates et les mois:
   - constat: différences entre le cadre de sondage (SRH) et le RFP,
     différences entre les vagues du SRH

   - but: réduire ces différences en utilisant les sources externes

   - effet sur les distributions: faible

3) Modèle de **non réponse**:
   –> comme le modèle sans calage

Pondération: calage

Les deux étapes supplémentaires:

4) **Calage** sur les caractéristiques des personnes:
   - constat: différences avec les sources externes
   - but: réduire ces différences en utilisant les sources externes
   - méthode: calage sur des variables (disponibles) au vu du taux de réponse et des estimateurs importants, –> calage sur les personnes, avec des poids ménage
   - effet sur les distributions et les estimateurs: important

5) **Winsorisation** des poids:
   - constat: plus grande étendue de la distribution des poids après calage –> peut poser des problèmes pour l'estimation
   - but: limiter cette étendue par une winsorisation
   - effet sur les distributions et les estimateurs: très faible

Le modèle de pondération avec calage

Résumé des cinq étapes:

1) Probabilité d'inclusion
   définie par le plan d'échantillonnage

2) Correction sur les strates et les mois
   avec les données ESPOP adaptées avec le recensement

3) Modèle de non réponse
   modèle logistique de non réponse

4) Calage sur les caractéristiques des personnes
   sur les données ESPOP adaptées avec le recensement

5) Winsorisation des poids
   en coupant les extrêmes à l'aide de la médiane

Recensement fédéral de la population (RFP)

Avantages:

- mêmes définitions du ménage privé et des types de ménages dans l'ERC et le RFP
- beaucoup de variables comparables:

  Ménages privés: taille, type, variables géographiques

  Personnes vivant dans les ménages privés: âge, sexe, nationalité

Inconvénients:

- effectué tous les 10 ans
- structure fixe pendant 10 ans

ESPOP

ESPOP: enquête de structure de la population

Avantage:

- source de données actuelle sur l'évolution de la population

Inconvénients:

- statistique sur les personnes, sans information sur les ménages
- statistique contenant toutes les personnes, y compris celles vivant dans les ménages collectifs

Combinaison recensement et ESPOP

–> corriger l'ESPOP avec les données du recensement pour obtenir uniquement les personnes vivant dans les ménages privés

Avantages:

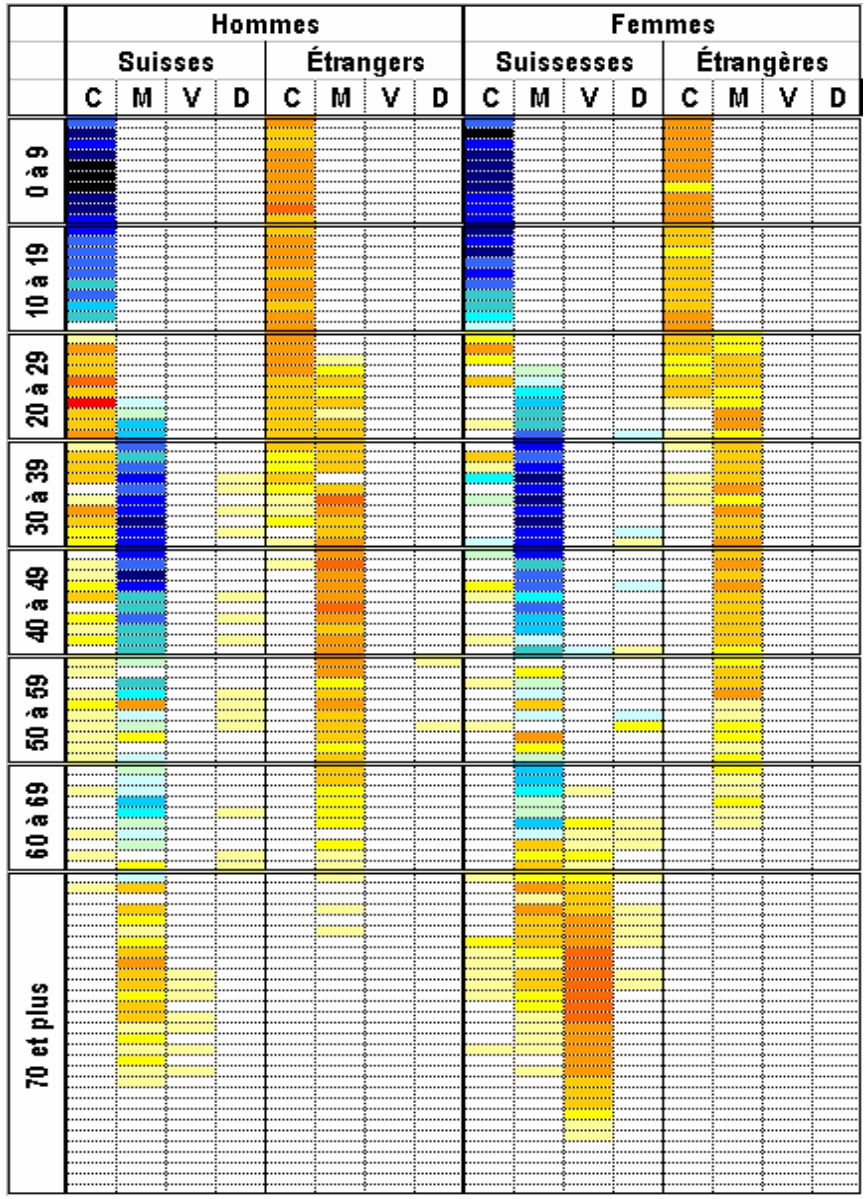
- calage uniquement sur les personnes des ménages privés
- pas de redressement trop important des personnes très âgées (> 80 ans: environ un tiers dans les ménages collectifs)

Inconvénients:

- rapport collectif / privé constant par tranche d'âge jusqu'au prochain recensement
- informations du RFP sur les ménages pas prises en compte

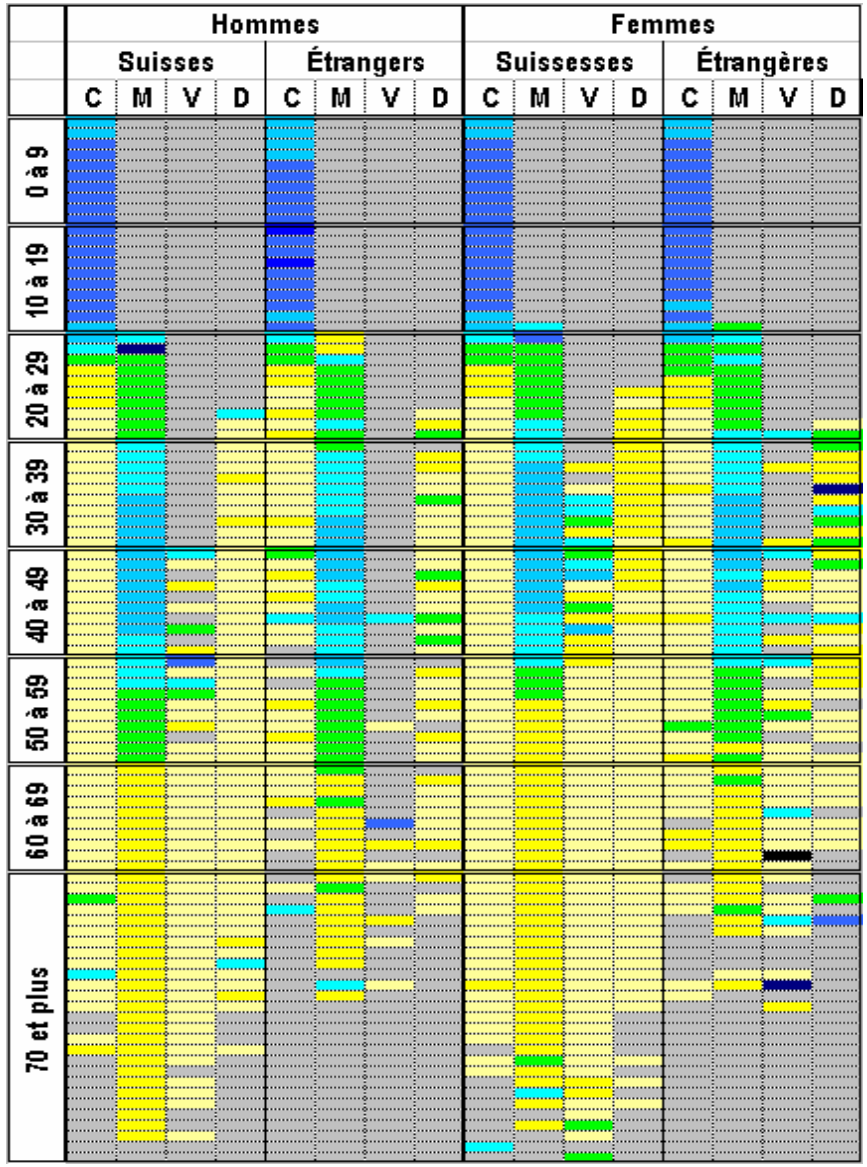
Choix des variables: taux de réponse

- But: regrouper au mieux les personnes avec un taux de réponse similaire
- Analyse de la **différence des répartitions** entre le RFP et les échantillons des ERC 2001 à 2003
- Couleurs:
  - bleu: surreprésentation
  - blanc: pas significatif
  - rouge: sous représentation



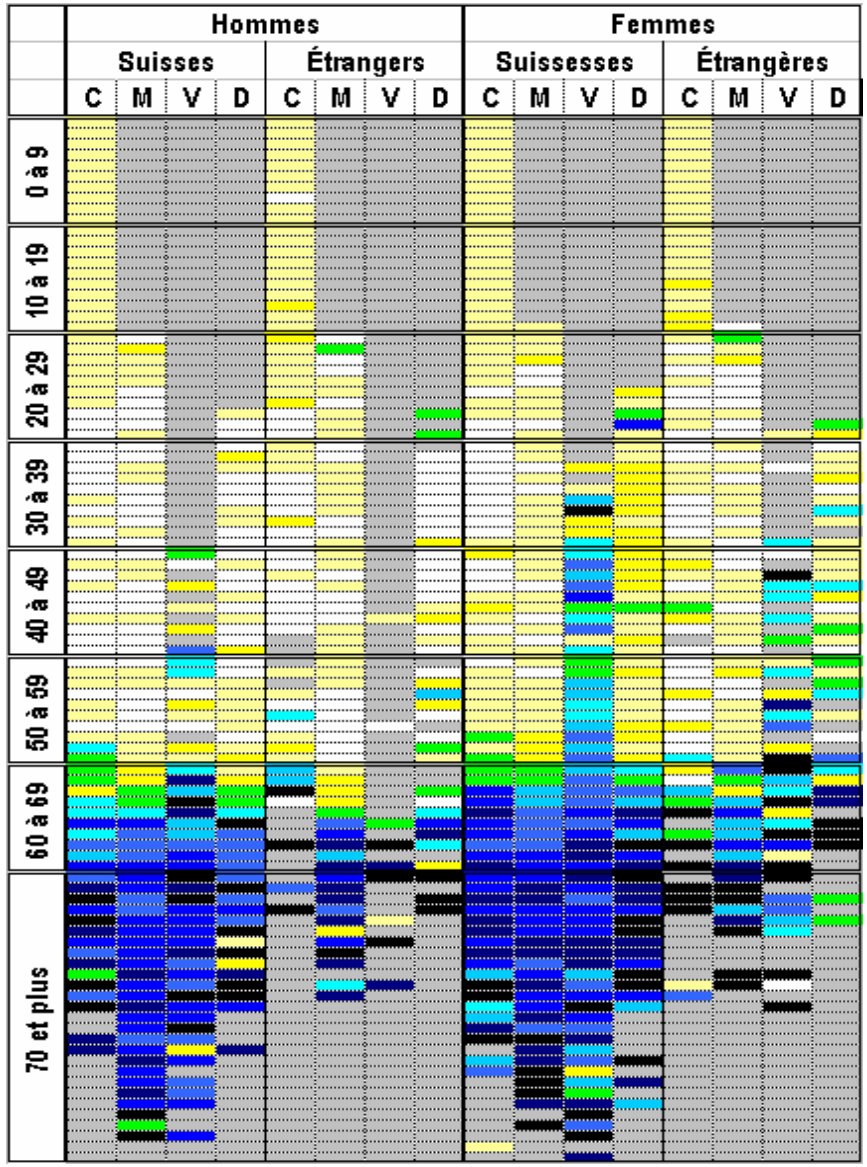
Choix des variables: corrélation avec la taille

- But: regrouper au mieux les personnes avec des niveaux d'estimateurs proches
- **Taille moyenne** par groupe de personnes des ERC 1998 à 2002
- Couleurs:
  - jaune clair: 1 à 1.75
  - jaune: 1.75 à 2.25
  - vert: 2.25 à 2.75
  - bleu clair: 2.75 à 3.75
  - bleu foncé: 3.75 et plus



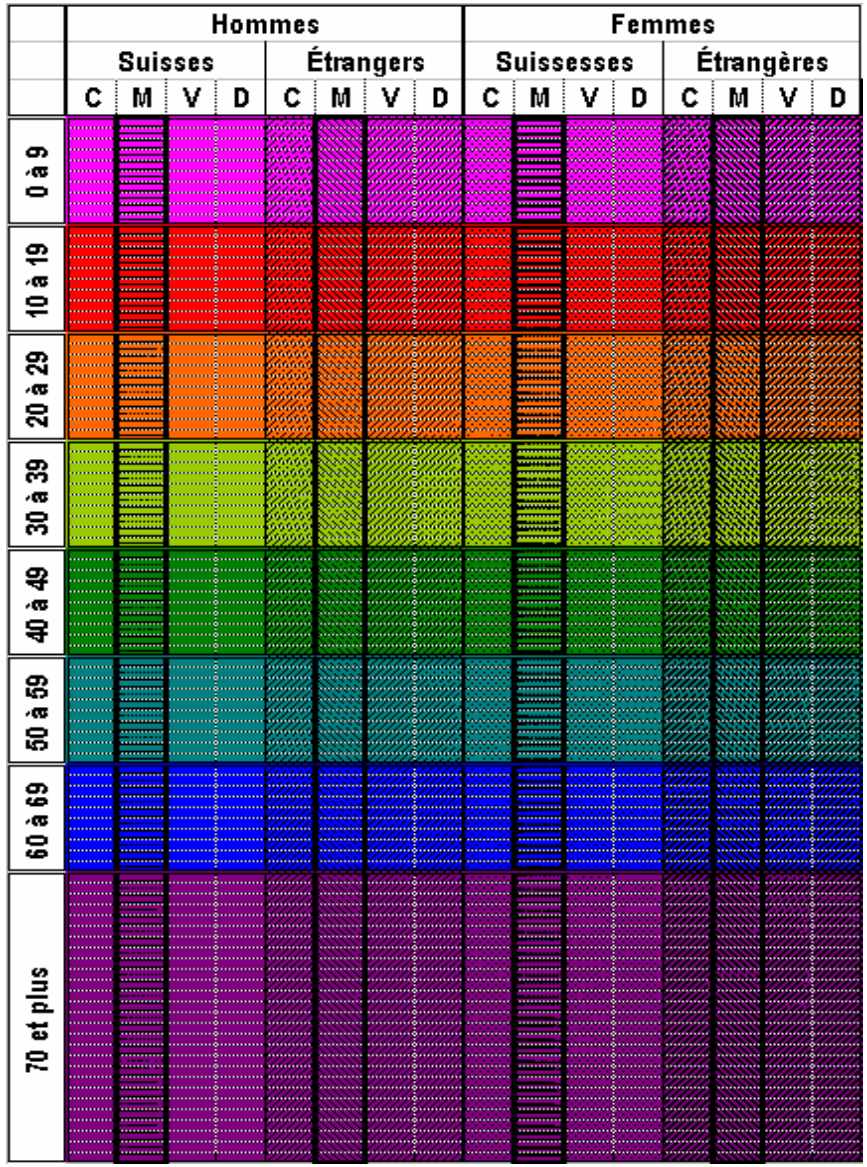
Choix des variables: corrélation avec les revenus de transfert

- But: regrouper au mieux les personnes avec des niveaux d'estimateurs proches
- **Part des revenus de transfert** par groupe de personnes des ERC 1998 à 2002
- Couleurs:
  - blanc: < 10%
  - jaune: 10% à 30%
  - vert - bleu clair: 30% à 50%
  - bleu foncé: 50% à 80%
  - noir: >80%



Choix des variables: regroupement

- Regroupement retenu avec 64 classes
- Groupes de personnes similaires:
  - taux de réponse
  - niveau des estimateurs
- Découpage rectangulaire
- Tranches d'âge par 10 ans
- Célibataires (C), veufs (V) et divorcés (D) regroupés; mariés (M) séparés



Choix du regroupement: modalités

- Plusieurs regroupements testés:
  - 5 classes (plus petite classe avec 348 / 9292 personnes)
  - ...
  - 17 classes (plus petite classe avec 123 / 9292 personnes)
  - 64 classes (plus petite classe avec 3 / 9292 personnes)
- Distributions des poids et nombre d'itérations peu influencés par le choix du regroupement
- Regroupement avec «64 classes» retenu pour obtenir des distributions de personnes plus précises
- Les modalités sans observations dans l'échantillon (p.ex. mariés < 20 ans) sont supprimés aussi dans les marges

Calage sur les personnes <--> ménages

- Unité d'enquête: **ménage**
- Marges de calage: fréquences des **personnes**

–> Contrainte: même poids pour les personnes du même ménage

- Approche: comptage des personnes par modalité et par ménage

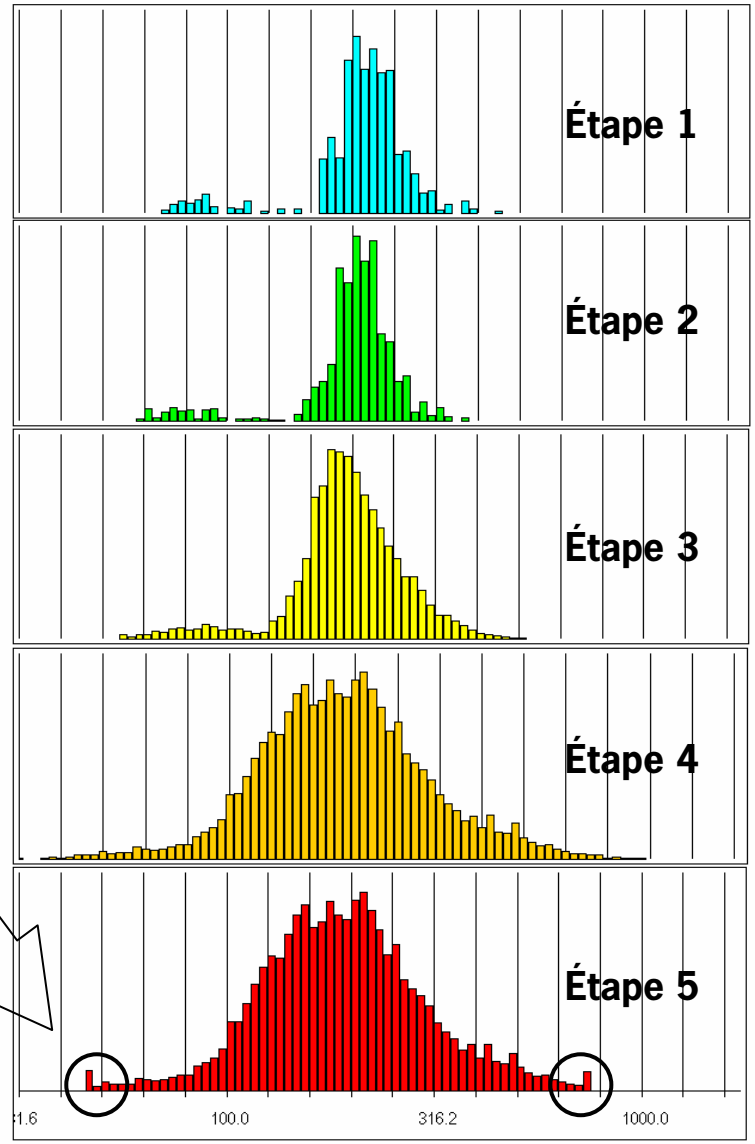
		Modalité 1	Modalité 2	Modalité 3	Modalité 4	...	Modalité m
Ménage 1	*Personne 1*	*1*				...	
Ménage 1	*Personne 2*				*1*	...	
Ménage 1	**- Total -**	**1**	**0**	**0**	**1**	**...**	**0**
Ménage 2	*Personne 1*		*1*			...	
Ménage 2	**- Total -**	**0**	**1**	**0**	**0**	**...**	**0**
Ménage 3	*Personne 1*	*1*				...	
Ménage 3	*Personne 2*	*1*					
Ménage 3	*Personne 3*		*1*				
Ménage 3	*Personne 4*						*1*
Ménage 3	**- Total -**	**2**	**1**	**0**	**0**	**...**	**1**
...	...	...	...	...	...	...	...
Ménage n	*Personne 1*		*1*			...	
Ménage n	*Personne 2*			*1*			
Ménage n	**- Total -**	**0**	**1**	**1**	**0**	**...**	**0**

Programme de calage utilisé: CALMAR

- Programme **SAS** développé à l'INSEE par M. Sautory:
  –> voir présentation vendredi
- Redresse un échantillon par calage sur les **marges**
  en minimisant la distance entre les poids initiaux
  et les poids finaux par itération
- Méthode utilisée dans cette application:
  –> "2: raking ratio"
- Normalement 6 à 9 itérations

Winsorisation des poids:

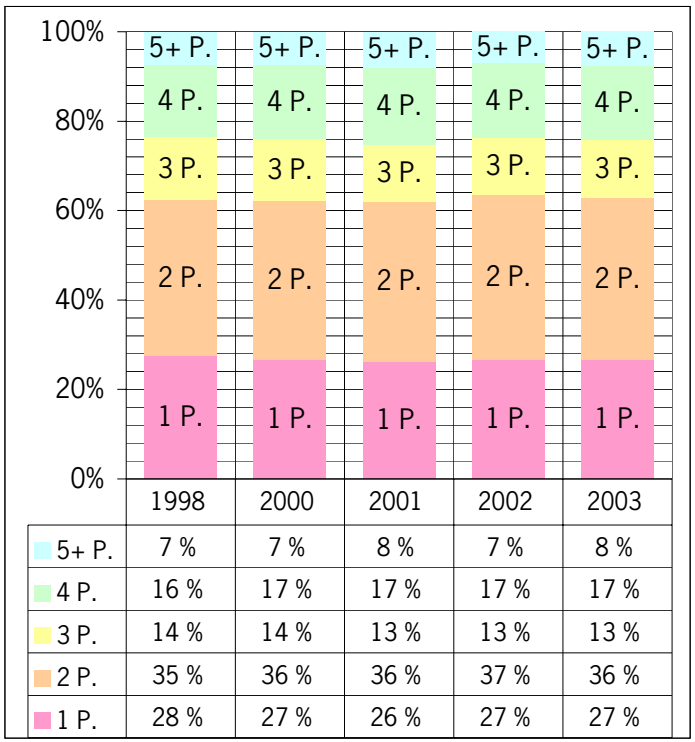
- Distribution des poids après calage: plus large (rapport max / min > 60)
- Un tel rapport est problématique pour les estimateurs:
  –> winsorisation des extrêmes
- Les extrêmes coupées:
  - minimum = médiane / 4
  - maximum = médiane * 4
- Distributions des ménages et des personnes peu touchées par cette winsorisation



Effets du calage: taille des ménages

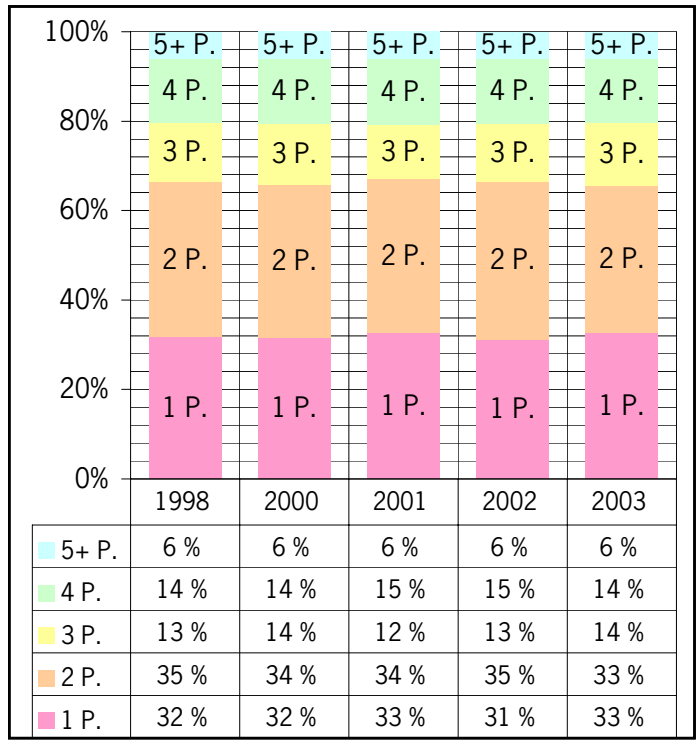
- Taille moyenne des ménages de l'ERC réduite: de 2.45 à 2.30
- Plus proche de celle observée par le recensement: 2.24

Modèle traditionnel



	1998	2000	2001	2002	2003
5+ P.	7 %	7 %	8 %	7 %	8 %
4 P.	16 %	17 %	17 %	17 %	17 %
3 P.	14 %	14 %	13 %	13 %	13 %
2 P.	35 %	36 %	36 %	37 %	36 %
1 P.	28 %	27 %	26 %	27 %	27 %

Modèle avec calage

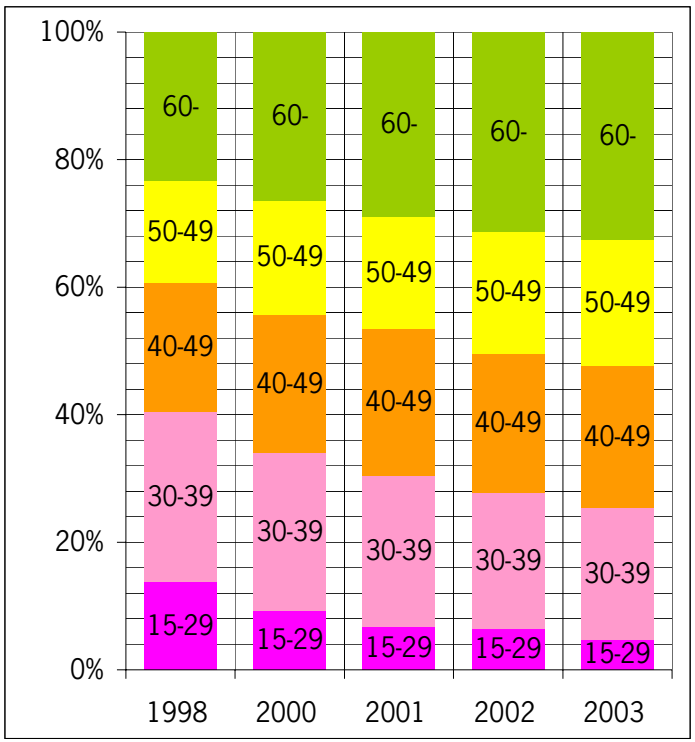


	1998	2000	2001	2002	2003
5+ P.	6 %	6 %	6 %	6 %	6 %
4 P.	14 %	14 %	15 %	15 %	14 %
3 P.	13 %	14 %	12 %	13 %	14 %
2 P.	35 %	34 %	34 %	35 %	33 %
1 P.	32 %	32 %	33 %	31 %	33 %

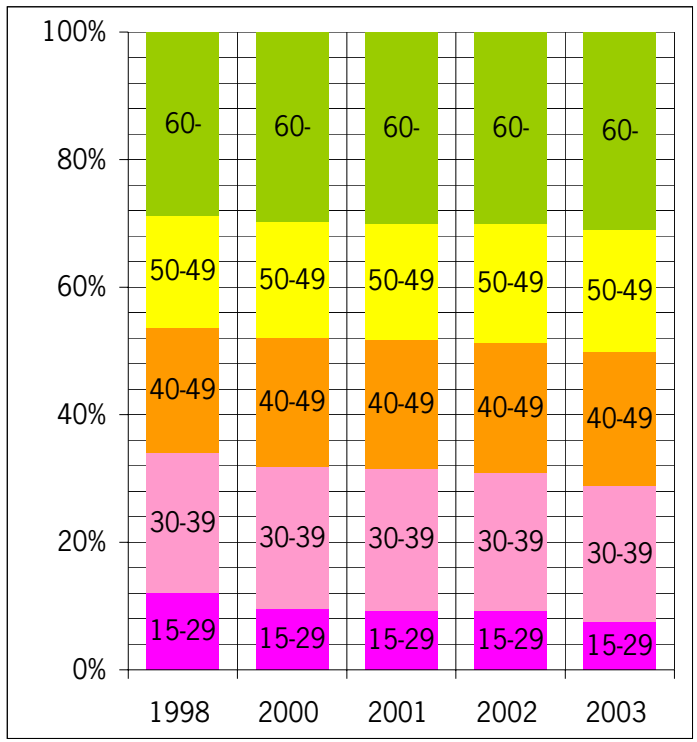
Effets du calage: évolution des tranches d'âge

- Tranches d'âge des personnes de référence: évolution atténuée
- Part des jeunes réduite en 1998 et augmentée en 2003

Modèle traditionnel

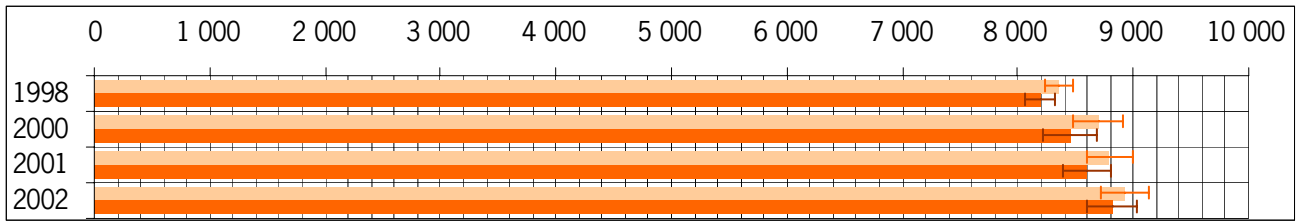


Modèle avec calage

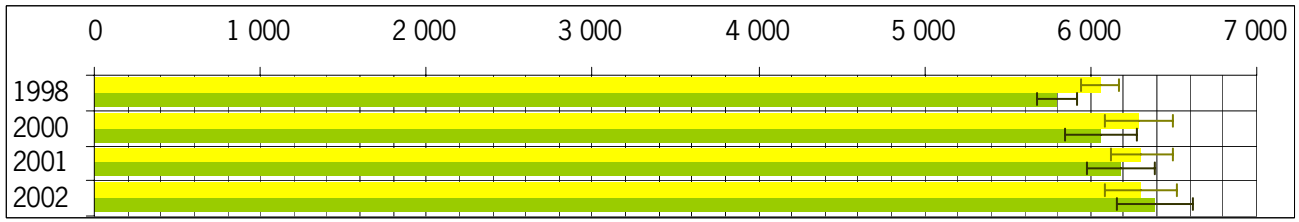


Effets du calage: estimateurs revenus

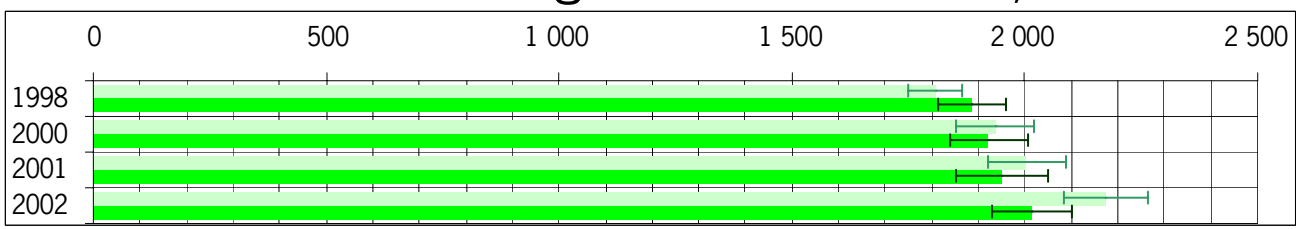
- Revenu total: réduction



- Revenus du travail: réduction en 1998, augmentation en 2002



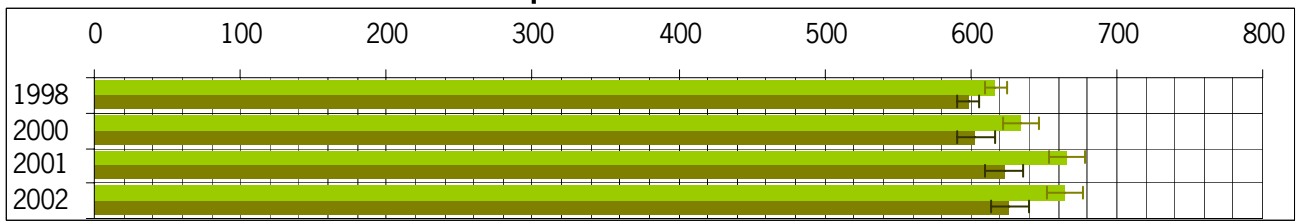
- Revenus de transfert: augmentation en 1998, réduction en 2002



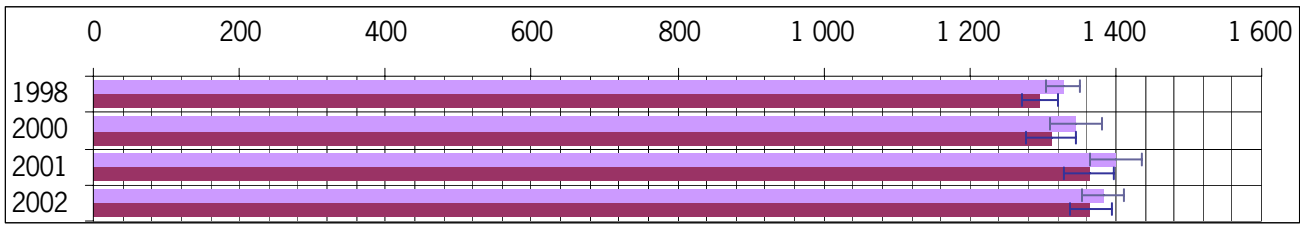
Couleurs: claires = modèle traditionnel / foncées = modèle avec calage

Effets du calage: estimateurs dépenses

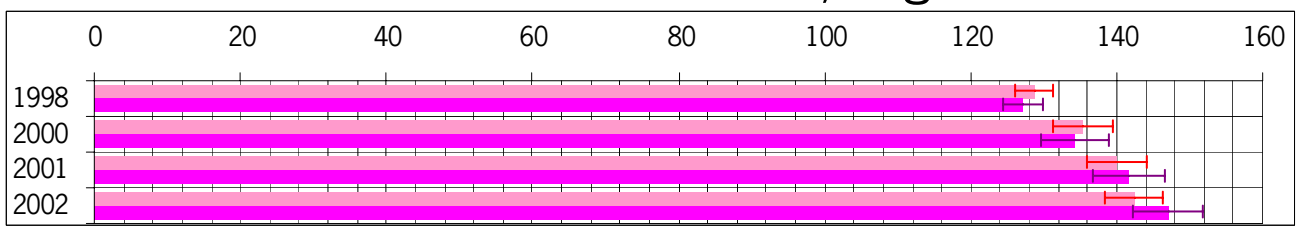
- Alimentation: réduction plus forte en 2001 et 2002



- Logement: réduction



- Communications: réduction en 1998, augmentation en 2002



Couleurs: claires = modèle traditionnel / foncées = modèle avec calage

Avantages du calage choisi

- Distributions des ménages plus proches des sources externes:
  –> taille des ménages plus petite (RFP 2000)
  –> augmentation du nombre de ménages locataires (RFP 2000)
  –> évolution selon la tranche d'âge plus plausible (ESPOP)
- Stabilisation (et redressement) des distributions au fil des années
- Prise en compte de la non-réponse des ménages sans information:
  –> ceux qui ont complètement refusé de participer
  –> ceux qui n'ont pas été atteints
  –> ceux qui ne sont pas dans l'annuaire
- Ménages collectifs exclus de la source externe
- Amélioration de la qualité des estimateurs

Questions, points ouverts, discussion

- Effets du calage sur l'estimation de la variance ?
- Limites, critères, «garde-fous» pour le nombre de modalités ?
- Winsorisation: est-ce qu'il existe une contrainte dans CALMAR afin de limiter le rapport maximum / minimum ?
- Avantages des autres méthodes par rapport au «raking ratio» ?