**INSTITUT NATIONAL DE LA STATISTIQUE**
**ET DES ETUDES ECONOMIQUES**
Série des Documents de Travail
de la
**DIRECTION DES STATISTIQUES DEMOGRAPHIQUES ET SOCIALES**
Département de la Démographie

N° F0207

# ESTIMATIONS LOCALES DANS LE CADRE DE L'ENQUÊTE HID

Christine COUET
Division « Enquêtes et Etudes Démographiques »
22 Novembre 2002

# Estimations locales dans le cadre de l'enquête HID

RESUME

Ce document retrace la démarche suivie par le groupe de travail chargé de la réalisation d'estimations locales dans le cadre de l'enquête HID (Handicaps-Incapacités-Dépendance).

L'enquête HID a été réalisée fin 1999 auprès de 16 945 individus vivant en domicile ordinaire en France métropolitaine. Ceux-ci ont été tirés parmi les 360 000 personnes ayant répondu à la préenquête VQS (Vie Quotidienne et Santé) associée au recensement de mars 1999. Cette enquête de filtrage comportait une extension de son échantillon dans huit zones géographiques (les départements des Bouches du Rhône, de l'Hérault, de l'Ille et Vilaine, de la Loire, du Pas de Calais, de la Seine et Marne et du Val d'Oise, et la région Haute-Normandie). Ces extensions, d'un coût très inférieur à celui de l'enquête complète, ont servi de base aux estimations présentées ici.

Le groupe de travail spécialisé qui a mis au point la méthode a utilisé la technique dite « d'estimations sur petits domaines ». Son avantage est de fournir des estimations proches de celles qu'aurait fournies une enquête auprès de 16 900 personnes dans chaque zone locale, sans en supporter les contraintes matérielles et financières.

La validité du modèle a pu être testée sur le département de l'Hérault, où le Conseil général a souhaité financer une extension départementale de l'enquête HID, permettant une exploitation directe de l'enquête à ce niveau géographique. La confrontation des résultats entre estimations directe et indirecte a guidé la recherche d'améliorations des performances du modèle.

Ce travail a permis d'estimer la prévalence des handicaps des populations dans les huit collectivités locales concernées. Une estimation de la précision des résultats est également proposée.

MOTS CLES

Méthode « d'estimations sur petits domaines », Estimation locale, Handicap, Incapacité, Dépendance.

# La réalisation d'estimations locales dans le cadre de l'enquête "Handicaps-Incapacités-Dépendance"

## 1 . Les estimations locales, un sous-produit de l'enquête nationale

L'objectif central de l'enquête HID est de donner une vue d'ensemble sur la réalité du handicap : dénombrer les personnes handicapées ou dépendantes à l'échelon national et évaluer les flux d'entrée et de sortie en incapacité. Le but de cette enquête est de chiffrer l'importance du phénomène mais aussi de fournir une description détaillée des situations afin de connaître l'ampleur des aides existantes et des besoins non satisfaits.

Parallèlement, l'enquête doit aussi répondre aux multiples interrogations qui se posent à l'échelon local, au niveau du département ou de la région, là où se prennent le plus souvent les décisions. C'est en effet aux gestionnaires locaux qu'il incombe de connaître les besoins actuels et d'en prévoir les évolutions. Ces exigences se font d'autant plus sentir qu'elles s'inscrivent dans un contexte de décentralisation des décisions et de développement des pouvoirs locaux, à l'heure où les besoins sont en fort développement du fait du vieillissement de la population. C'est pourquoi l'enquête HID comporte dans son cahier des charges la fourniture d'estimations de résultats concernant la population des ménages pour certains départements ou régions. Pour répondre à cette demande il fallait disposer d'un outil capable de produire des résultats localement, tout en garantissant une cohérence d'ensemble.

L'enquête HID ne peut donner de réponses directes à ces interrogations locales pour des raisons évidentes de coûts. De fait, les échantillons, au niveau du département par exemple, sont trop restreints pour assurer aux résultats une fiabilité suffisante. Par ailleurs, si on peut envisager d'appliquer systématiquement à l'échelon local des prévalences d'incapacité relevées au niveau national, cette solution risque de donner une vision simplifiée et de passer à côté de la diversité des situations locales.

L'insuffisance de l'échantillon local au moins dans le cas des départements conduit à recourir à des techniques d'estimations sur petits domaines. L'idée sous-jacente à cette méthode est de se servir de l'ensemble de l'échantillon de l'enquête HID nationale pour garantir une bonne précision aux estimations tout en adaptant ces données à la diversité des situations locales. Cette adaptation repose sur des hypothèses et donc sur la construction d'un modèle appelé modèle d'estimation sur « petits domaines ».

Par des conventions avec plusieurs collectivités locales L'INSEE s'est engagé à fournir des fichiers HID comportant une variable de pondération adaptée à certaines collectivités locales. C'est autour de cet objectif prioritaire que c'est constitué un groupe de travail, dont la mission était de définir une démarche conduisant à la production de statistiques locales et à la réalisation de plusieurs publications. Ce groupe comprenait, en plus de quelques membres de l'équipe HID, des « méthodologues » et plusieurs responsables régionaux concernés (cf. composition détaillée du groupe, en note 2, page 8).

Toutefois ce travail ne concerne qu'un des volets de l'enquête qui au total en compte quatre (**cf. annexe I**). En ne se rapportant qu'au premier passage de l'enquête réalisée auprès des ménages en 1999, les estimations présentées ici laissent de côté la population handicapée

hébergée en institutions ; très minoritaire au total (1,12 % de l'ensemble de la population métropolitaine), celle-ci représente cependant une part importante des handicaps les plus lourds.

On peut envisager de compléter ces évaluations en y ajoutant les populations handicapées des institutions des départements ou régions, par une adaptation des résultats du premier passage de l'enquête en institutions (enquête 1998). La méthode d'estimation peut être inspirée de celle utilisée ici. Elle repose à l'évidence sur une bonne connaissance de la nature des établissements et et du nombre des résidents à l'échelon local.

## 2 . Les éléments disponibles pour produire des estimations locales

Décliner des résultats nationaux au niveau d'un département ou d'une région suppose de pouvoir adapter des données d'enquête nationale – donc recueillies sur un domaine plus large que la zone considérée – à une situation locale particulière qu'il convient de caractériser. Il faut donc disposer à la fois du fichier national de l'enquête HID et d'informations auxiliaires caractérisant au mieux la zone d'étude.

Le premier élément, le fichier national, est le produit de l'enquête consacrée aux ménages vivant en France métropolitaine. Les principales étapes de sa réalisation - tirage et redressement de l'échantillon – sont décrites en **annexe II**.

Son complément, les particularités locales, est constitué de statistiques dont la production avait été prévue au moment de la conception de l'enquête, dans le souci de mieux cerner les structures locales. La réalisation de ces statistiques a nécessité parfois certains aménagements du dispositif de l'enquête.

Tout d'abord, grâce à la proximité du recensement de population de mars 1999, on a bénéficié d'informations de bonne qualité - parce que récente - sur la composition des populations locales, sur lesquelles on a pu ensuite s'appuyer pour construire les estimations locales des populations handicapées.

Ensuite, il est apparu souhaitable de tirer parti au mieux de la première phase de l'enquête auprès des ménages : l'enquête de filtrage, dite « Vie Quotidienne et Santé » (VQS). Les Conseils généraux et régionaux désireux d'informations chiffrées ont été sollicités pour financer des extensions de l'enquête VQS sur leur zone de compétence. Huit d'entre eux – sept départements et une région - ont accepté de signer des conventions dans ce sens. Ces extensions devaient permettre d'obtenir de bonnes estimations des variables issues de VQS, parce que l'échantillon était de taille suffisante. L'échantillon national initialement prévu pour une taille de près de 300 000 personnes a ainsi légèrement dépassé les 400 000 au total. Ce premier éclairage sur la situation locale du handicap devait permettre d'améliorer, d'une façon encore mieux ciblée qu'à travers le RP99, l'estimation des variables de l'enquête HID à un niveau infra-national. C'est sur ces huit zones – voir tableau ci-dessous - qu'a été expérimentée la méthode des « petits domaines », notamment grâce aux améliorations apportées par l'extension des échantillons VQS.

**Taille des échantillons VQS et HID dans les zones avec et sans extension**

| | répondants VQS | répondants HID |
|---|---|---|
| **- Zones avec extension VQS** | | |
| *7 départements* | | |
| Bouches-du-Rhône (13) | 20 490 | 682 |
| Hérault (34) | 16 172 | 1 479 |
| Ille-et-Vilaine (35) | 20 196 | 400 |
| Loire (42) | 17 856 | 207 |
| Pas-de-Calais (62) | 33 481 | 397 |
| Seine-et-Marne (77) | 20 413 | 613 |
| Val-d'Oise (95) | 16 074 | 270 |
| *1 région* | | |
| Haute-Normandie (27 et 76) | 18 299 | 468 |
| **- Zones sans extension** | | |
| *87 départements* | 196 029 | 12 429 |
| **Total** | **359 010** | **16 945** |

Enfin, cette méthode d'estimation indirecte nécessite une très grande prudence dans l'élaboration des résultats. La validité du modèle sous-jacent a pu être testée sur le département de l'Hérault, où le Conseil général a souhaité financer une extension départementale de l'enquête HID proprement dite. Alors qu'à un département correspond, en moyenne, moins de 200 répondants HID, l'échantillon de l'Hérault atteint 1 479 individus sur les 16 945 répondants que compte l'enquête.

## 3 . La méthodologie des « petits domaines »

Le développement de méthodes visant à exploiter les enquêtes nationales pour produire des estimations à des niveaux géographiques régionaux - ou infra-régionaux - est assez récent à l'INSEE. Ce type de démarche a principalement été expérimenté sur les données de l'enquête annuelle sur l'emploi de 1996[1].

Au-delà des estimateurs locaux directs utilisant les données de l'enquête en provenance exclusivement de la zone étudiée, qui manquent souvent de précision parce que la taille de l'échantillon est trop restreinte, il existe toute une gamme d'estimateurs indirects reposant sur

[1] K. Attal-Toubert et O. Sautory, « Estimation de données régionales à l'aide de techniques d'analyse multidimensionnelle », Méthodologie Statistique, Document de travail n°9807.

la construction de modèles (voir la liste d'articles traitant de méthodes d'estimation sur petits domaines dans l'**annexe** bibliographique **III)**.

C'est de cette deuxième catégorie d'estimateurs que se sont inspirés les travaux du groupe[2] chargé de définir une méthode d'estimation locale. Celle-ci devait permettre de produire des évaluations sur les thèmes de l'enquête HID dans les huit zones géographiques (sept départements et une région) pour lesquelles on disposait d'extension de l'enquête VQS : les Bouches du Rhône, l'Hérault, l'Ille et Vilaine, la Loire, le Pas de Calais, la Seine et Marne, le Val d'Oise et la région Haute-Normandie.

### 3.1 Le modèle d'estimation sur « petits domaines »

A la base, il y a la reconnaissance de l'influence de certains facteurs sur la fréquence et la sévérité du handicap. L'influence de l'âge est certainement l'exemple le plus intuitif. On part d'un constat simple, par exemple que la prévalence[3] d'une incapacité est moins élevée chez les sujets jeunes que chez les sujets âgés.

D'une façon plus générale, la population peut se répartir en sous-groupes ou « post-strates »[4] - définis par le croisement de critères socio-démographiques ou d'autres indicateurs liés au phénomène étudié - ayant chacune des prévalences bien spécifiques, homogènes dans la sous-population concernée. Or c'est sur l'échantillon le plus vaste, l'échantillon national, que la mesure de ces prévalences est la plus fiable. En résumé, l'hypothèse de base du modèle d'estimation sur «petits domaines » est d'admettre que le «comportement moyen » dans une zone – département ou région - à l'intérieur d'une post-strate est identique au comportement moyen national dans cette post-strate.

Autrement dit, la proportion de mal-entendants parmi les femmes de telle tranche d'âge, habitant une commune de tel type d'unité urbaine et ayant été classées selon leurs réponses à VQS dans tel "groupe de handicap VQS"… est supposée ne pas dépendre de la zone géographique.

En conséquence, le choix des critères les plus discriminants à l'égard du handicap et la recherche de la partition de la population la meilleure, définissant un nombre de post-strates optimum - ni trop ni trop peu –, doivent faire l'objet du plus grand soin.

On notera que parmi les critères retenus, outre des facteurs socio-démographiques «ayant une influence sur les prévalences de handicap », en premier rang desquels l'âge joue un rôle essentiel, figure également un indicateur direct des prévalences de handicap, obtenu grâce à la réalisation d'une préenquête légère et résumant à l'échelle locale les réponses de la poulation :

[2] Composition du groupe de travail : Pascal Ardilly (DR de Lyon), François Clanché (DEE), Christine Couet (DEE), Jean-Claude Deville (DSDS), Claude Gissot (DREES), Jean-Luc Le Toqueux (DAR), David Levy (DR de Lyon), Claude Michel (DR de Montpellier), Pierre Mormiche (DEE), Lionel Qualité (DR de Saint-Quentin), Christian Robert (GENES), Frédéric Tardieu (DR de Rennes), Laurent Wilms (UMS).

[3] En épidémiologie, la prévalence désigne simplement la proportion de sujets qui dans une population donnée souffre de telle pathologie ou de tel handicap.

[4] Appelées ainsi car elles sont définies après la réalisation du plan de sondage.

le « groupe VQS ». On mesure l'importance qu'a revêtu pour le travail d'estimation sur petits domaines la réalisation des extensions locales de l'enquête VQS, avec notamment une meilleure définition de la structure des populations locales par strates qui en découle.

### 3.2 La formalisation du modèle

Laurent Wilms a proposé une définition de l'estimation sur petits domaines basée sur un modèle de comportement et un rappel de ses propriétés statistiques (**cf. annexe IV**). Le groupe de travail en a adopté le principe que l'on peut résumer de la façon suivante.

Appelons Y la variable d'intérêt HID dont on veut estimer la moyenne $\overline{Y}_R$ à un niveau régional (ou départemental). On se propose par exemple d'estimer la proportion de mal-entendants dans la région de Haute-Normandie.

Sachant que la population se répartie en H post-strates qui correspondent à autant de variétés dans la prévalence des incapacités, l'hypothèse de comportement consiste ici à postuler que la proportion de personnes mal-entendantes est constante au sein de la post-strate *h*, quelle que soit la zone géographique considérée. Alors, l'estimateur post-stratifié de cette proportion s'écrit :

$$\hat{\overline{Y}}_R = \sum_{h=1..H} \frac{\hat{N}_{Rh}}{\hat{N}_R} \hat{\overline{Y}}_h$$

où $\hat{\overline{Y}}_h$ représente l'estimateur de la moyenne de la variable *Y* dans la post-strate *h* calculée sur l'échantillon HID national, soit :

$$\hat{\overline{Y}}_h = \frac{\sum_{k \in s_{HID} \cap postrate h} \frac{y_k}{p_k}}{\sum_{k \in s_{HID} \cap postrate h} \frac{1}{p_k}}$$

où $p_k$ est la probabilité d'inclusion de l'individu k dans l'échantillon HID

et $\hat{N}_{Rh}$ et $\hat{N}_R$ sont les estimateurs respectifs de l'effectif régional global et de la post-strate h. Ils sont calculés à partir de l'échantillon VQS de la région *R*, échantillon noté VQSR, à partir des formules :

$$\hat{N}_{Rh} = \sum_{k \in s_{VQSR} \cap postrate h} \frac{1}{p'_k} \qquad \text{et} \qquad \hat{N}_R = \sum_{k \in s_{VQSR}} \frac{1}{p'_k}$$

avec $p'_k$ la probabilité d'inclusion de l'individu *k* dans l'échantillon VQS.

### 3.3 Choix des critères de post-stratification

L'hypothèse dite «de comportement homogène » sera d'autant mieux vérifiée que les post-strates auront été convenablement définies.

Un des premiers travaux a donc consisté à étudier quels critères avaient une influence sur l'état du handicap au niveau national, à sélectionner les facteurs les plus influents et à éprouver leur pertinence quelle que soit la nature du handicap, enfin à vérifier si leur effet était sensiblement analogue sur les différentes parties du territoire. On a largement utilisé les procédures logistiques à cet effet (on trouvera de nombreuses données sur ce point dans le rapport de stage de Valérie Albouy – ENSAE, été 2000). Un résumé de ces travaux est présenté en **annexe V**.

Sur quel support géographique doit-on réaliser l'étude ? On a choisi de partager la France en huit zones, chacune constituée de plusieurs départements regroupant un effectif de répondants HID suffisant, de l'ordre de 1 500 à 2 000 personnes.

La démarche consiste à trouver l'ensemble de variables qui expliquent au mieux l'état du handicap sur chacune des zones géographiques, au point de rendre négligeable l'effet d'appartenance à telle ou telle zone. Elle suppose qu'il existe des critères qui partagent la population en sous-groupes présentant des comportements comparables en matière de risque de handicap, quelle que soit la zone d'étude. Le modèle de comportement sélectionné devra rendre compte au mieux des disparités locales, à travers l'inégale représentation des sous-populations dans les différentes zones.

Cette approche reconnaît implicitement que :

- si le modèle de comportement est vérifié sur des zones élargies, on suppose qu'il garde toute sa pertinence à des niveaux géographiques plus fins (département ou région). Toutefois la confrontation des résultats ainsi obtenus avec ceux provenant d'une estimation directe dans le département de l'Hérault (dont la fiabilité est admise grâce à l'extension HID) devrait conforter cette hypothèse ;

- bien que le modèle soit établi uniquement à partir de quelques variables judicieusement choisies de l'enquête HID, on admet qu'il est également vrai sur l'ensemble des thèmes abordés par l'enquête.

Ces travaux montrent la difficulté à expliquer la totalité des disparités régionales à travers les variables socio-démographiques « classiques » (sexe, âge, CS,…). D'autres facteurs, pour lesquels on ne dispose pas toujours d'informations statistiques, peuvent exercer une influence.

- Par exemple, l'inégale répartition géographique des places offertes en institutions a probablement des effets, du fait de son caractère complémentaire, sur la prévalence des incapacités des populations vivant en ménage.

- En outre, divers facteurs tenant à l'environnement et au mode de vie - tels que les habitudes alimentaires ou les loisirs - peuvent avoir une influence à l'échelon local.

- Enfin, la perception même du handicap, dans la mesure où elle relève de comportements culturels, peut se traduire aussi par des disparités régionales dans les réponses saisies par l'enquête.

Toutes ces données sont difficiles à intégrer dans le modèle. En définitive, il se dégage de cette étude que les deux critères les plus influents sur le handicap sont l'âge et le groupe VQS. Deux critères supplémentaires ont été retenus : le sexe et la tranche d'unité urbaine. Le milieu social joue également un rôle important. Toutefois, pour des raisons de disponibilité d'information, il n'a pu intervenir au moment de la post-stratification.

En effet, il est indispensable que les critères retenus soient disponibles :

- d'une part dans les données de l'enquête (pour pouvoir calculer des comportements moyens nationaux dans les strates qu'ils définissent) ;

- d'autre part dans les données de structure (pour disposer des effectifs de populations - nationale et locales - pour ces mêmes strates).

La réalisation de la seconde condition tient à l'utilisation simultanée des résultats du recensement de mars 1999 et des extensions locales de l'enquête VQS. Ce rapprochement a permis d'établir les effectifs des six groupes VQS dans chacune des zones disposant d'une extension de l'enquête et de chiffrer les populations aux croisements de ce critère avec les autres.

Ces deux conditions étaient réunies pour quatre des cinq critères influents : le sexe, l'âge, la tranche d'unité urbaine et le groupe VQS. Comme aucune question relative au milieu social n'est posée dans l'enquête VQS, ce facteur n'a pas été retenu pour la post-stratification mais ses effets seront pris en compte ultérieurement dans une dernière phase de calage.

### 3.4 Réduction du nombre de post-strates

Le croisement des quatre critères de « post-stratification » retenus (sexe, classe d'âge, tranche d'unité urbaine et groupe VQS) conduisait à la définition de 180 strates. L'inconvénient de ce nombre relativement élevé est de risquer d'obtenir des effectifs parfois fort réduits de l'échantillon HID national dans certaines strates et en conséquence des estimations de comportements entachées d'une assez forte incertitude.

On a donc considéré qu'il était nécessaire - et le groupe de travail a jugé légitime - de regrouper un certain nombre de strates selon deux critères :

- d'une part déclarer le regroupement d'une strate souhaitable d'après son effectif dans l'échantillon ;

- d'autre part déclarer le regroupement de deux strates acceptable en raison de la proximité du comportement de leurs populations.

Cette proximité a été testée par des modèles de type LOGIT reliant quelques variables d'intérêt de HID aux caractéristiques socio-démographiques intervenant dans la post-stratification (**cf. annexe VI**).

Les regroupements sont surtout réalisés chez les individus à faible risque : essentiellement les populations jeunes n'appartenant pas aux groupes VQS n°6. En définitive, le nombre des post-strates a été réduit à 52. Il correspond aux croisements des modalités - plus au moins agrégées - des quatre critères. Le traitement du département des Bouches du Rhône constitue un cas particulier puisque le nombre de post-strates y est abaissé à 46, du fait de l'absence d'individus VQS vivant en commune rurale et appartenant au groupe VQS n°6.

### 3.5 Modification des poids individuels du fichier national

Une contrainte pesait sur le choix de l'estimateur. Etant donnée l'importance du nombre de variables dans l'enquête HID, il importait que les pondérations des fichiers locaux soient indépendantes de la variable d'intérêt à estimer. Or l'avantage d'une approche de type « petits domaines » est de satisfaire cette condition. On montre qu'une adaptation des poids individuels de l'échantillon national à chacune des situations locales permet de traiter localement l'ensemble des variables de l'enquête, tout en réalisant des estimations conformes au choix méthodologique du groupe de travail. Cette opération n'a de sens que si la post-stratification retenue garde toute sa pertinence quelle que soit la nature du handicap étudié.

Ainsi, de la définition d'un estimateur post-stratifié de type « petit domaine »,

$$\hat{\bar{Y}}_R = \sum_{h=1..H} \frac{\hat{N}_{Rh}}{\hat{N}_R} \hat{\bar{Y}}_h = \sum_{h=1..H} \frac{\sum_{k \in VQSR \cap postrateh} \frac{1}{p'_k}}{\sum_{k \in VQSR} \frac{1}{p'_k}} \frac{\sum_{k \in HID \cap postrateh} \frac{y_k}{p_k}}{\sum_{k \in HID \cap postrateh} \frac{1}{p_k}}$$

on aboutit à une redéfinition des poids des individus dans l'échantillon national, puisque :

$$\hat{\bar{Y}}_R = \frac{1}{\sum_{k \in VQSR} \frac{1}{p'_k}} \sum_{h \in 1..H} \left[ \sum_{k \in HID \cap postrateh} \left( \frac{1}{p_k} \cdot \frac{\sum_{k \in VQSR \cap postrateh} \frac{1}{p'_k}}{\sum_{k \in HID \cap postrateh} \frac{1}{p_k}} \right) y_k \right] = \frac{\sum_{k \in HID} \left( \frac{1}{p_k} \cdot \frac{\sum_{k \in VQSR \cap postrateh} \frac{1}{p'_k}}{\sum_{k \in HID \cap postrateh} \frac{1}{p_k}} \right) y_k}{\sum_{k \in VQSR} \frac{1}{p'_k}}$$

où le rapport $\left( \frac{1}{p_k} \cdot \frac{\sum_{k \in VQSR \cap postrateh} \frac{1}{p'_k}}{\sum_{k \in HID \cap postrateh} \frac{1}{p_k}} \right)$ représente les nouveaux poids individuels du fichier national, adaptés aux situations locales.

En d'autres termes, le calcul d'une estimation de type « petits domaines » passe en pratique par la modification des pondérations du fichier national.

La suite de la démarche a consisté à déterminer la précision de cet estimateur post-stratifié puis à le tester en comparant ses résultats à ceux d'une estimation directe, soit sur des sous-échantillons régionaux de l'enquête HID - de taille suffisamment importante pour que le

l'estimation directe soit fiable - soit sur le département de l'Hérault qui comprend une extension de son échantillon HID.

### 3.6 Précision des résultats

L'estimation « indirecte » ainsi définie gagne en précision par rapport à une estimation « directe » basée uniquement sur le petit nombre d'observations appartenant à la zone d'étude. Cette conviction de bon sens ne doit pas faire oublier notre ignorance de la mesure exacte de la variance d'un estimateur post-stratifié indirect.

La complexité du plan de sondage de l'enquête HID rend déjà difficile la connaissance de l'expression de la variance de l'estimateur à l'échelon national. Cette expression est à fortiori méconnue dans le cas d'un estimateur local de type «petit domaine », pour lequel ont été introduits une post-stratification et un modèle de comportement.

Toutefois, le groupe de travail a jugé acceptable l'approximation proposée par Laurent Wilms et Valérie Albouy[5] (cf **annexe VII**). En résumé, ils se sont appuyés sur la formule d'un estimateur de Horvitz-Thompson, qu'ils ont adapté au cas d'un estimateur post-stratifié – en remplaçant dans son expression la variable d'intérêt par sa moyenne sur la strate – et dans laquelle ils ont introduit un modèle de comportement.

L'expression ainsi obtenue ne pouvant être calculée, ils se sont servis pour l'approcher d'une formule d'approximation proposée par Jean-Claude Deville pour le cas de sondage à probabilités inégales. Ces diverses adaptations ont conduit à l'expression suivante :

$$\hat{V}_2(\hat{\bar{Z}}_{HT}) = \frac{1}{N^2}\frac{1}{1-\sum_{l\in s} a_l^2}\sum_{k\in s}(1-p_k)\left(\frac{\hat{e}_k}{p_k} - A\right)^2$$

où $a_l = \dfrac{1-p_l}{\sum_{k\in s}(1-p_k)}$ et $A = \sum_{k\in s} a_k \dfrac{\hat{e}_k}{p_k}$ et $\hat{e}_k = \dfrac{\hat{N}}{\hat{N}_h}\dfrac{\hat{N}_{Rh}}{\hat{N}_R}\left(y_k - \dfrac{\sum_{k\in s_{HID}\cap strh}\frac{y_k}{p_k}}{\sum_{k\in s_{HID}\cap strh}\frac{1}{p_k}}\right)$

Dans cette approche, on ignore l'aléa provenant de la structure locale observée dans VQS. De plus, cette expression ne prend pas en compte les effets de grappe de l'échantillon. Ces éléments contribuent donc à sous-estimer la variance de l'estimateur de type « petits domaines ».

### 3.7 Tests de validité du modèle

Conscient de ne retenir aucune spécificité locale du handicap autrement qu'à travers la composition par strates de la population de la zone d'étude, le groupe a voulu s'assurer que le

[5] Les détails de la démarche sont dans le rapport de stage de Valérie Albouy – ENSAE, été 2000.

modèle de comportement adopté – l'égalité des prévalences d'incapacité par strate entre le domaine étudié et le territoire national – était bien approprié. Une mauvaise hypothèse de comportement entraînerait assurément un risque important de biais.

Le moyen le plus immédiat d'éprouver la qualité de l'estimation est de comparer les résultats de la méthode des « petits domaines » avec des estimations directes réalisées sur des super-régions - afin de disposer d'un nombre d'observations HID suffisant - ou sur le département de l'Hérault, le seul à disposer d'une extension de l'enquête.

De nombreux tests ont été réalisés notamment dans l'Hérault pour juger de la proximité des estimations directes et indirectes. On a regardé si les intervalles de confiance, définis à 95 % autour des diverses estimations, se chevauchaient ou pas. Ces tests portaient sur dix variables d'intérêt de HID, choisies en raison de la diversité de leur niveau de prévalence et parce qu'elles donnaient du handicap une vision assez globale. Le choix des critères de post-stratification n'a été fixé qu'au terme d'une période d'hésitation. Dans un premier temps, la variable « type de logement », opposant l'habitat individuel au collectif, entrait dans la définition des post-strates (cf. les résultats des tests en **annexe VIII-a**). Par la suite, le critère « taille de la commune » en 3 postes a été préféré à la notion de type d'habitat (cf. les résultats dans l'Hérault en **annexe VIII-b**).

Dans ce dernier cas, les comparaisons ont concerné trois estimateurs :

- (a) *l'estimateur post-stratifié direct*, résultat de l'adaptation des comportements observés par strate dans l'échantillon des 1 479 répondants HID de l'Hérault à la structure par strate observée sur les 16 172 réponses VQS de l'Hérault, qui représente la cible à atteindre,

- (b) *l'estimateur national* calculé à partir des 16 945 réponses de l'échantillon HID national, au titre de témoin,

- (c) *l'estimateur post-stratifié indirect* qui, en suivant la méthode dite des « petits domaines », applique les comportements nationaux observés dans les 52 strates à la structure définie par VQS dans l'Hérault.

Des confrontations de (a) avec (c), il ressort que les intervalles de confiance se recouvrent le plus souvent mais que la prévalence des incapacités estimée directement dans l'Hérault est généralement inférieure à la moyenne nationale. Un effet local résiduel persiste donc dans ce département.

Une première amélioration possible réside dans la prise en compte des spécificités sociales à l'échelon local. L'introduction des données disponibles du RP99 a été réalisée, sans attendre le chiffrement de la PCS. On dispose pour cela du niveau d'études et de la position professionnelle.

### 3.8 Calages ultimes

Comme on a pu l'observer, aucune caractéristique strictement sociale (liée à la profession personnelle ou familiale, au niveau d'études, au revenu…) n'a été prise en compte à ce stade. La raison principale tient à leur absence dans l'enquête VQS qui empêche de les croiser avec

le « groupe VQS », indicateur résumé de handicap. Ceci ne pourra être réalisé qu'après un travail d'appariement de VQS avec les fichiers du recensement.

Cette dimension sociale a été introduite par un calage de l'échantillon HID - l'interview HID recueille quant à lui diverses caractéristiques sociales - sur les « marges » sociales du recensement. On ne dispose pas encore du chiffrement de la CS et donc du milieu social dans le RP, mais on a pu construire une variable sociale d'après les informations disponibles sur le niveau d'études, l'activité, le statut professionnel et pour les salariés la position professionnelle (toutes variables disponibles dans le RP et dans HID).

Au terme de ce travail (**cf. annexe IX**) il subsiste toujours un aspect local inexpliqué, dont la valeur relative varie fortement selon la variable d'intérêt. Cette remarque a contraint le groupe à rechercher le moyen d'enrichir l'hypothèse de comportement en y introduisant des aspects plus directement liés au domaine étudié.

## 4 . Tentatives visant à améliorer le modèle

Deux perfectionnements du modèle ont été envisagés. Le premier introduit des particularités locales en modifiant directement l'hypothèse de comportement du modèle initial. Mais cette correction a peu d'effets en pratique et elle ne satisfait pas aux conditions d'utilisation du modèle.

La seconde amélioration est apparue un peu plus prometteuse. Les caractéristiques du département ou de la région agissent ici en marge du modèle de comportement classique, en tant que deuxième facteur. Cette solution respecte les contraintes d'application du modèle mais elle en complique énormément l'utilisation.

Le groupe de travail a examiné successivement ces deux approches.

### 4.1 Une seule méthode de calcul quelle que soit la variable étudiée

La première méthode a été utilisée par Olivier Sautory et Ketty Attal pour estimer des taux d'activité et des taux de chômage régionaux par sexe et tranche d'âges à partir de données issues de l'enquête emploi (cf. Olivier Sautory, Ketty Attal). Elle consistait à prendre comme estimation du comportement local dans chaque strate non pas uniquement le comportement national, mais une combinaison du comportement national et du comportement local (la moyenne, pour la variable considérée, des réponses fournies par le sous-échantillon interrogé dans la zone étudiée), les coefficients attribués aux deux composantes - nationale et locale - étant inversement proportionnels à la variance de chacune d'elles.

Cette tentative a été abandonnée pour deux raisons :

1. Au niveau des zones pour lesquelles on cherche à établir des estimations (le département), l'effectif de l'échantillon des réponses à HID est de 170 en moyenne (15 400 réponses, Hérault non compris, pour 90 départements concernés). Sachant qu'on conduit les calculs sur une cinquantaine de strates on disposerait dans le meilleur des cas d'une dizaine de réponses départementales dans la strate, ce qui

implique une variance toujours considérable et, par voie de conséquence, une prise en compte tout à fait négligeable de cette seconde composante.

2. Indépendamment de ce premier motif, la définition de la combinaison des effets nationaux et locaux à partir de la variance, quelque séduisante qu'elle soit du point de vue de l'étude, se traduit par d'importantes complications lors de son application. En effet, la variance n'est évidemment pas réductible au nombre de réponses mais elle tient compte de leur dispersion ; elle diffère donc selon la variable que l'on cherche à estimer, donnant à cette solution une tournure peu opérationnelle quant on sait qu'HID traite plusieurs centaines de variables.

Cette seconde remarque souligne à nouveau un des traits de la procédure d'estimation recherchée dans le cadre de ce travail (voir § 3.5) : trouver une méthode d'utilisation simple, qui "redresse" les diverses estimations statistiques par un coefficient unique.

Ce n'était pas possible avec une procédure prenant en compte la variance de la (ou des) variable (s) d'intérêt qu'on cherche à estimer.

### 4.2 Intégrer ou pas une composante locale résiduelle

Lors des tests réalisés dans l'Hérault (voir § 3.7) les deux types d'estimateurs – post-stratifiés direct (a) et indirect (c) - présentaient une différence sensible, faisant apparaître, pour la majorité des variables d'intérêt, une prévalence de handicaps plus faible selon l'estimateur direct que selon l'estimateur "petits domaines". En conséquence, pour cette deuxième approche, on a tenté de définir un "effet résiduel" de la zone étudiée, imputable soit à une spécificité de comportement de la zone soit à des variables "structurelles" non prises en compte ou non encore disponibles. Cet effet propre à la zone s'additionne à l'effet « classique » de la strate. Une formalisation du modèle à deux facteurs est proposée en **annexe X** et l'expression des pondérations nationales associées à ce modèle en **annexe XI**. Mais son expérimentation dans le département de **l'Hérault** souligne quelques problèmes spécifiques à cette nouvelle approche (cf. les résultats et commentaires en **annexe XII**).

Pour les départements **autres que l'Hérault**, la présence d'un tel "effet résiduel" ne pouvait être mis en évidence à partir de HID, compte tenu de la faible dimension de l'échantillon départemental. Le groupe de travail s'est interrogé sur la stabilité de la mesure de l'effet départemental « résiduel » (cf. **annexe XIII)** et a proposé qu'elle soit testée par Boot Strap en tirant des sous-échantillons de l'Hérault. Ce travail a montré combien l'estimateur « combiné », avec effet résiduel mesuré sur **un seul** département, n'offre pas plus de précision qu'un estimateur direct (cf. les résultats de cette méthode en **annexe XIV**).

On a alors envisagé de s'appuyer sur une zone plus large que le département et pour ce faire, on a posé deux hypothèses.

1. On a supposé qu'une spécificité de comportement, si elle existe, se retrouverait également dans les résultats de l'enquête VQS. Dès lors, les réponses à VQS ont été traitées pour réaliser une typologie des départements. Celle-ci a conduit à la constitution de quatre classes, construites par regroupement de départements

présentant des similitudes en matière de handicap, après élimination des effets dus aux variables de stratification déjà identifiées (**cf. annexe XV**).

2. On a alors admis qu'il était possible d'attribuer à chacun des départements étudiés la spécificité de comportement relevée sur les variables HID de la classe de départements dans laquelle il avait été ainsi rangé.

Ces opérations se traduisent en pratique par une adaptation unique des poids individuels de l'échantillon national à chaque situation locale, indépendamment des variables étudiées (se reporter, en **annexe XVI,** à l'expression de la nouvelle pondération du fichier national HID associé à un estimateur local à deux facteurs, dont l'effet local serait mesuré sur une classe de départements).

Mais les travaux menés n'ont pas donné de résultat positif et s'ils avaient abouti plus favorablement, leur utilisation aurait été difficile :

- en effet, la distance moyenne entre les estimations de variables HID obtenues par exploitation directe (sur l'Hérault et sur des régions suffisamment grandes) et respectivement, soit l'estimateur strictement petits domaines, soit l'estimateur corrigé de la spécificité locale, ne fait pas apparaître d'amélioration - au sens d'une réduction de la distance observée (cf : le bilan sur les résultats dans le département de l'Hérault en **annexe XVII** et les tests complémentaires, sur des zones plus vastes que le département, en **annexe XVIII**);

- de plus, cet estimateur à deux facteurs (estimateur petits domaines classique + effet spécifique local) présente l'inconvénient de générer des poids négatifs pour un grand nombre d'observations (cf. **annexe XIX**). Ceci risque de faire apparaître, pour des croisements un peu détaillés, des résultats négatifs et donc dépourvus de sens et inexploitables. En outre la version de SAS alors en vigueur à l'INSEE ne supportait pas de pondérations négatives pour la plupart des procédures statistiques les plus courantes.

- enfin, la cohérence de l'ensemble du modèle d'estimation n'est plus assurée. Désormais, rien ne garantit que la somme des estimations locales est égale à l'estimation nationale.

Le groupe a donc considéré que cette tentative ne pouvait être mise en application pour l'instant et a retenu la forme « classique » de l'estimateur post-stratifié indirect défini initialement. Pour chacune des huit zones considérées, un jeu de pondérations locales a été calculé selon la méthode présenté au § 3, et livré aux diverses collectivités locales intéressées.

## 5 . Précautions particulières propres aux exploitations locales

L'une des contraintes que l'on s'est imposé au moment du choix de la méthode d'estimation était de disposer d'un seul système de pondérations individuelles par zone géographique sélectionnée pour la réalisation des estimations. Dès lors, chaque jeu de poids devait être utilisé de façon universelle pour l'ensemble des variables d'intérêt. Dans la pratique, c'est bien un seul système de poids par zone qui a été appliqué, quel que soit le thème traité :

déficience, incapacité et désavantage (cf. la diversité des thèmes traités par l'enquête en **annexe XX**).

Toutefois, il faut être prudent quant à l'universalité de la méthode. Certaines pratiques ou certaines réalisations locales peuvent ne pas être le reflet de comportements moyens globaux. C'est par exemple le cas lorsque l'attribution de prestations obéit à des critères régionaux sans qu'il existe une réelle harmonisation des situations à l'échelon national ou bien, lorsque l'implantation d'équipements de portée nationale dans une région, entraîne de fortes répercussions sur le handicap local.

Pour éviter cet écueil on s'abstiendra de traiter certains sujets à un niveau infra-national : c'est le cas, par exemple, des données relatives aux revenus, aux allocations et aux reconnaissances officielles.

# *ANNEXE I*

## L'architecture d'ensemble de l'enquête HID

Pour répondre aux objectifs globaux d'HID, il est apparu nécessaire d'engager **deux enquêtes parallèles** : l'une auprès des personnes vivant en **institutions** et l'autre auprès de la population des **ménages**.

La place importante qu'occupent les institutions dans l'organisation de l'enquête tient au fait qu'un grand nombre d'entre elles hébergent précisément des personnes handicapées. C'est le cas des établissements pour personnes âgées, des foyers pour handicapés adultes ou pour enfants ou adolescents handicapés, des établissements psychiatriques…

Cependant c'est à l'intérieur de logements ordinaires que réside la majeure partie de la population handicapée. A l'avenir, cette prépondérance devrait se renforcer puisque l'on assiste, parallèlement au vieillissement de la population, à un développement important de la politique de maintien à domicile.

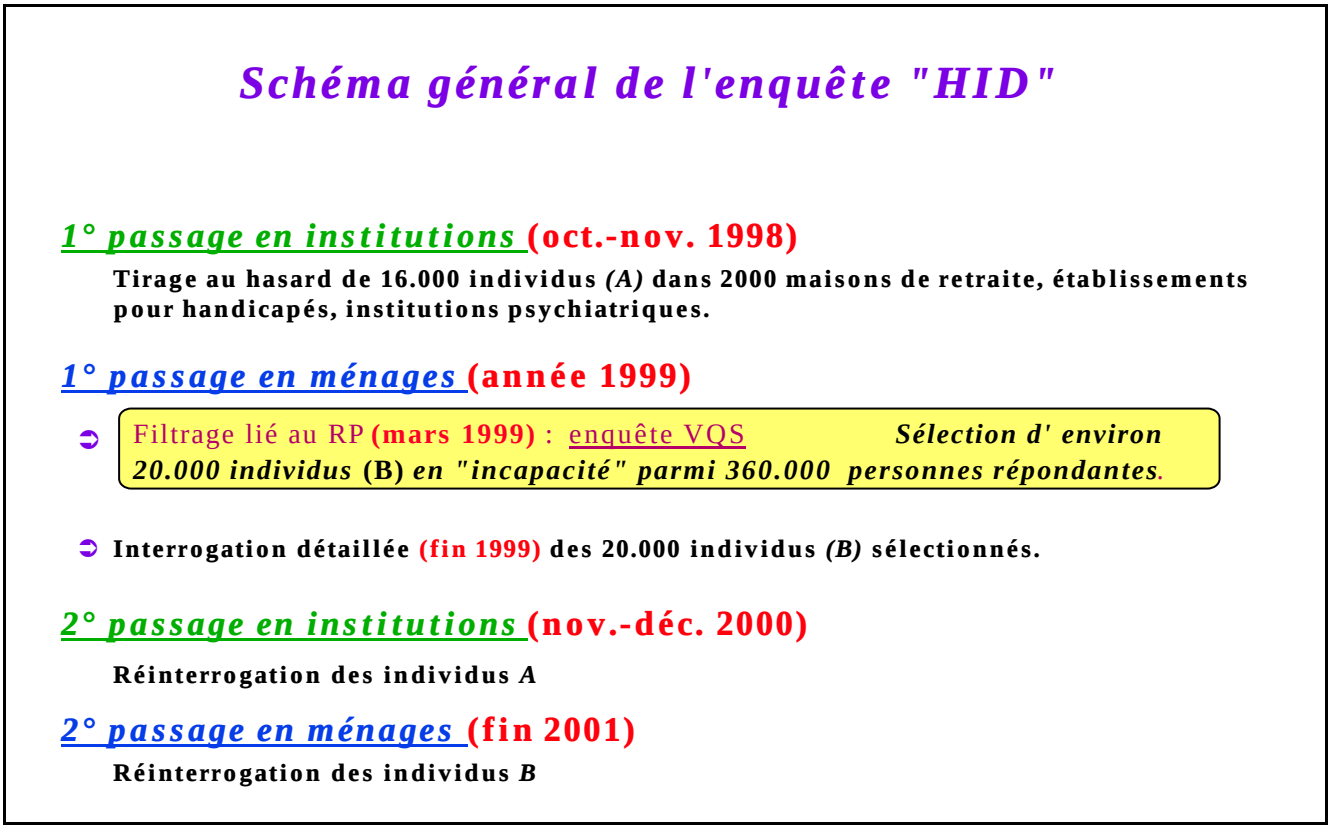
L'organisation prévue pour l'enquête comprend 3 phases :

- une phase de « filtrage » des personnes en incapacité dans la population vivant à domicile (400 000 personnes interrogées à l'occasion du recensement de mars 99, 359 000 réponses exploitables )

- une phase de description approfondie de l'incapacité, de ses origines et de ses conséquences, des aides et des besoins (20 000 personnes à domicile et 15 000 en institutions)

- un second passage, deux ans plus tard, destiné à mesurer certains flux et l'évolution des situations (auprès des individus interrogés dans la 2ème phase).

Dans la pratique, l'enquête se déroule sur quatre années, conformément au schéma ci-dessous.




Une enquête spécifique a par ailleurs été organisée pour les pensionnaires d'établissements pénitentiaires.

# *ANNEXE II*

## Tirage et pondération de l'échantillon national

La première phase de l'opération, consacrée aux ménages vivant en France métropolitaine, a recueilli 16 924 interviews individuelles auprès de 20 116 ménages et s'est déroulée du 02 novembre 1999 au 31 janvier 2000. Elle a également concerné 4 091 interviews de proches apportant leur aide aux personnes HID.

La collecte a été effectuée en face-à-face, sur micro-ordinateur portable, par les enquêteurs professionnels de l'INSEE, selon la procédure "CAPI" (Collecte Assistée Par Informatique), ce qui a permis d'intégrer de nombreuses aides pour l'enquêteur.

L'échantillon de l'enquête de fin 1999 en «ménage » a été tiré en deux temps et l'enquête menée en deux phases. La première étape ne sert qu'à compter les personnes concernées et à les sélectionner pour l'interrogation ultérieure : c'est l'étape dite de filtrage. La seconde sert à décrire les incapacités des personnes concernées, les origines ou les causes de ces incapacités et leurs conséquences éventuelles dans les principaux domaines de l'activité sociale.

**L'enquête de filtrage, dite « Vie Quotidienne et Santé »** (VQS) a été adjointe au recensement de la population de mars 1999, selon la technique classique des enquêtes sur « l'Etude de l'Histoire Familiale ». Elle a concerné une population de plus de 400 000 personnes, dont 359 000 réponses exploitables (soit un taux d'échec - refus et bulletins inexploitables - de 14 %). Le questionnaire comportait dix huit questions, sélectionnées afin de classer les individus de l'échantillon en prenant en compte plusieurs des dimensions du handicap : (1) une grille d'incapacités comportant huit questions, (2) les recours à des aides (humaines ou techniques), (3) les limitations d'activité ressenties par la personne, (4) la revendication d'un handicap par le répondant et (5) la reconnaissance sociale de la situation de handicap.

Son tirage est stratifié par zone géographique, de façon à assurer une représentativité de l'échantillon à l'échelon régional. Le territoire métropolitain a été divisé en 36 strates. A la base, une strate correspond à une région, mais certaines régions ont été éclatées pour répondre aux particularités des enquêtes VQS et EHF. Pour l'enquête VQS, huit extensions départementales ou régionales répondant aux demandes et aux financements de collectivités locales ont été réalisées. En cas d'extension VQS sur un département, le département est défini comme une strate.

Le tirage a été «aréolaire », c'est-à-dire qu'on a tiré des agents recenseurs (responsables en moyenne de zones de 600 habitants) chargés de distribuer et relever les questionnaires VQS à l'ensemble de la population vivant en ménage ordinaire dans leur zone de recensement.

Sa réalisation s'est effectuée en deux étapes. Dans un premier temps, on a procédé au tirage de **zones de délégués** au sein desquelles ont été sélectionnés par la suite des **secteurs d'agents recenseurs.**

En pratique, on a défini un nombre de zones à tirer dans chaque strate géographique alors même que le découpage géographique en zones du RP99 n'était pas terminé et donc que le nombre total de zones par strate était inconnu. On a été contraint d'utiliser pour ce tirage le fichier des districts de 1990. Celui-ci a été trié par strate géographique selon une double nomenclature, socio-économique et familiale. On a ensuite procédé à un tirage systématique en tenant compte de la taille du district. Les districts ainsi sélectionnés ont constitué des « points d'entrée », les zones de délégués retenues étant celles qui incluent ces districts. Cette procédure est assimilable à un tirage de zones proportionnel à leur taille ; elle assure une assez bonne représentativité de l'échantillon selon les caractéristiques sociales et démographiques.

Ensuite, par zone de délégué, un certain nombre de secteurs d'agent recenseur ont été sélectionnés, avec une probabilité qui diffère par type de zone (12 cas au total selon les diverses configurations possibles des 2 enquêtes EHF et VQS). Dans les secteurs d'agent recenseur VQS, les agents ne traitent que de l'enquête VQS et couvrent l'ensemble des personnes vivant en domicile ordinaire dans le secteur.

Le redressement de l'échantillon VQS sur la population a été réalisé en prenant en compte les probabilités de tirage de chaque répondant, puis en « ajustant » la composition de l'échantillon des répondants sur la population par strate.

**L'enquête détaillée «handicaps-Incapacités-Dépendance »** (HID) a été réalisée auprès d'un échantillon des répondants à VQS. Le tirage des individus HID parmi les répondants à VQS est stratifié avec des probabilités très inégales. Il a été réalisé, dans chaque zone d'enquête, de façon à sur-représenter les individus très handicapés d'après les informations disponibles dans le fichier des réponses à VQS. Pour cela on a construit un indicateur synthétique des réponses à VQS (le « groupe VQS ») comportant 6 modalités de « handicap croissant ». Le tirage a été effectué selon dix « strates HID » croisant les six groupes VQS et un ou deux groupes d'âges.

Les personnes appartenant au groupe « 6 » (le plus sévèrement handicapé) ont été tirées selon un taux de sondage élevé, celles appartenant au groupe « 1 » (n'ayant aucune difficulté dans leur vie quotidienne), le plus nombreux dans l'ensemble de la population, ont eu au contraire un faible taux de tirage. L'éventail des probabilités de tirage dans cette étape varie presque de 1 à 100.

Cette construction présente les avantages habituels des sondages stratifiés. En premier lieu, l'échantillon obtenu sur-représente fortement les personnes atteintes par un handicap, permettant ainsi d'en décrire les situations avec suffisamment de précision. Ensuite, il fournit une bonne représentation des différents niveaux de handicap, indépendamment des divers seuils de reconnaissance administrative. Enfin, en s'appuyant sur une pré-enquête tirée dans l'ensemble de la population vivant en domiciles ordinaires, l'échantillon ainsi obtenu produit des résultats représentatifs de celle-ci.

Le redressement de HID sur l'échantillon VQS a été réalisé en plusieurs étapes, par des calages successifs : calage du fichier des répondants HID sur l'échantillon de l'enquête HID, puis sur le fichier VQS, ensuite sur les effectifs globaux du RP par zone géographique et enfin sur la pyramides des âges du RP projetée à la date de l'enquête.

On trouvera ci-dessous l'effectif de la population des ménages ordinaires en France métropolitaine au recensement de mars 1999, celui de l'échantillon des répondants VQS et celui des répondants HID par sexe, groupe VQS et tranche d'âge.

**Répartition par sexe, classe d'âges et groupes VQS des deux échantillons (HID ET VQS)**

| | Echantillon HID | | Echantillon VQS | | Population RF | |
|---|---|---|---|---|---|---|
| SEXE | Effectif | % | Effectif | % | Effectif | % |
| Hommes | 7885 | 46.5 | 173137.4 | 48.2 | 28186454 | 48.7 |
| Femmes | 9060 | 53.5 | 185872.6 | 51.8 | 29645362 | 51.3 |
| Total | 16945 | 100.0 | 359010.0 | 100.0 | 57831816 | 100.0 |

| | Echantillon HID | | Echantillon VQS | | Population RF | |
|---|---|---|---|---|---|---|
| GROUPE VQS | Effectif | % | Effectif | % | Effectif | % |
| 1 | 2936 | 17.3 | 282126.9 | 78.6 | 44870564 | 77.6 |
| 2 | 1335 | 7.9 | 23820 | 6.6 | 3965130 | 6.9 |
| 3 | 2037 | 12.0 | 15705 | 4.4 | 2674887 | 4.6 |
| 4 | 1773 | 10.5 | 9453 | 2.6 | 1583648 | 2.7 |
| 5 | 3634 | 21.4 | 13502 | 3.8 | 2345905 | 4.1 |
| 6 | 5230 | 30.9 | 14403.1 | 4.0 | 2391681 | 4.1 |
| Total | 16945 | 100.0 | 359010.0 | 100.0 | 57831816 | 100.0 |

| | Echantillon HID | | Echantillon VQS | | Population RF | |
|---|---|---|---|---|---|---|
| Groupe d'âge | Effectif | % | Effectif | % | Effectif | % |
| 00 à 09 ans | 729 | 4.3 | 41969.3 | 11.7 | 6580607 | 11.4 |
| 10 à 19 | 824 | 4.9 | 49553.9 | 13.8 | 7706744 | 13.3 |
| 20 à 29 | 889 | 5.2 | 47120.3 | 13.1 | 7699033 | 13.3 |
| 30 à 39 | 1373 | 8.1 | 53794.3 | 15.0 | 8584028 | 14.8 |
| 40 à 49 | 2012 | 11.9 | 52793.4 | 14.7 | 8405679 | 14.5 |
| 50 à 59 | 2359 | 13.9 | 41596.5 | 11.6 | 6781073 | 11.7 |
| 60 à 69 | 2552 | 15.1 | 32172.0 | 9.0 | 5408915 | 9.4 |
| 70 à 79 | 4230 | 25.0 | 27718.2 | 7.7 | 4568635 | 7.9 |
| 80 à 89 | 1639 | 9.7 | 10496.4 | 2.9 | 1667703 | 2.9 |
| 90 et plus | 338 | 2.0 | 1795.8 | 0.5 | 429398 | 0.7 |
| Total | 16945 | 100.0 | 359010.0 | 100.0 | 57831816 | 100.0 |

Ainsi, les populations les plus certainement handicapées (les «groupe 6» de VQS) sont fortement sur-représentées dans l'échantillon HID ; le redressement, en prenant en compte leur probabilité de tirage élevée, ramène leur proportion dans la population à un niveau beaucoup plus faible. De même, la population âgée est naturellement plus nombreuse dans l'échantillon que dans la population générale.

# *ANNEXE III*

## Annexe bibliographique

Quelques éléments de bibliographie en matière d'estimation locale:


- INSEE - Document de travail : Méthodologie statistique n° 9807 « Estimation de données régionales à l'aide de techniques d'analyse multidimensionnelle ». K. Attal-Toubert, O. Sautory.

- Statistique Canada - Techniques d'enquête, juin 1994. Vol. 20, n°1, pp. 3-23. « Les petites régions : problèmes et solutions ». M.P. Singh, J. Gambino et H.J. Mantel.

- Les contributions des participants à la conférence de Riga : « Small area estimation ». dont :
  - Local estimations for the « Handicap, Disability and Dependence » survey. Ketty Attal Toubert.
  - The French local estimates on labour statistics,based on administrative registers and surveys : current situation and future work. Marc Christine.

- Rapport de stage de Valérie Albouy – ENSAE, été 2000.

- Projet Statistique ENSAI 2ème année – « Calcul de la variance des estimateurs sur petits domaines dans l'enquête HID par une méthode empirique ». A. Bourgeois, D. Bonnery, j. Valentino et R. Bunales, intervenant L. Qualité (2002).

# *ANNEXE IV*

## Propositions d'estimateurs pour l'enquête HID

L. Wilms (UMS)

Les estimateurs que l'on cherche à construire sont de 2 sortes :

- soit nationaux,

- soit régionaux (au sens administratif du terme).

Nous traiterons, d'abord, des estimateurs nationaux pour une moyenne puis nous adapterons les estimateurs au cas régional en faisant **une hypothèse de comportement des régions**.

Le principe de construction des estimateurs (nationaux ou régionaux) est de caler, sur des variables du recensement de 1999, l'estimateur post-stratifié bâtit à partir des post-strates de degré de handicap.

Rappelons l'échantillonnage HID :

- 1ère phase : sélection d'un échantillon dit VQS puis constitution de $H$ post-strates[6].

- 2ème phase : sélection dans chacune des post-strates d'un échantillon, l'ensemble des $H$ sous-échantillons sera nommé échantillon HID.

### I- Les estimateurs nationaux

*I-1 L'estimateur post-stratifié*

Soit une variable d'intérêt HID notée $Y$[7] dont on veut estimer la moyenne nationale, $\overline{Y}$. On a, en repérant les post-strates par l'indice $h$ :

$$\sum_{h=1..H} \frac{N_h}{N} \overline{Y}_h$$

Notons $s_{hid}$ l'échantillon HID national et $p_k$ la probabilité de sélectionner l'individu $k$ issu d'un tirage HID. L'estimateur post-stratifié, $\hat{\overline{Y}}_{pst1}$, s'écrit :

[6] Les post- strates sont construites selon un degré de handicap croissant.

[7] Par exemple : Y=1 si l'individu utilise du matériel audiovisuel adapté pour les malentendants (Question DAUDIO du questionnaire HID).

$$\hat{\bar{Y}}_{pst1} = \sum_{postrateh} \frac{N_h}{N} \frac{\sum_{k \in s_{hid} \cap postrateh} \frac{y_k}{p_k}}{\sum_{k \in s_{hid} \cap postrateh} \frac{1}{p_k}}$$

Evidemment, les proportions nationales des groupes de handicap sont inconnues. On les estime grâce à l'échantillon VQS. On obtient finalement comme estimateur, avant calage sur des variables de recensement :

$$\hat{\bar{Y}}_{pst2} = \sum_{postrateh} \frac{\hat{N}_h}{\hat{N}} \frac{\sum_{k \in s_{hid} \cap postrateh} \frac{y_k}{p_k}}{\sum_{k \in s_{hid} \cap postrateh} \frac{1}{p_k}}$$

avec

$$\frac{\hat{N}_h}{\hat{N}} = \frac{\sum_{k \in s_{VQS} \cap poststrateh} \frac{1}{p'_k}}{\sum_{k \in s_{VQS}} \frac{1}{p'_k}}$$

(on note $s_{VQS}$ l'échantillon VQS national et $p'_k$ la probabilité de sélectionner l'individu $k$ issu d'un tirage VQS).

Remarquons que l'estimateur $\hat{\bar{Y}}_{pst2}$ peut s'écrire sous la forme :

$$\sum_{k \in s_{hid}} w_k y_k$$

avec, si l'individu $k$ appartient à la post-strate $h$ :

$$w_k = \frac{\hat{N}_h}{\hat{N}} \frac{\frac{1}{p_k}}{\sum_{k \in s_{hid} \cap postrateh} \frac{1}{p_k}}$$

*I-2 L'estimateur post-stratifié redressé*

Soit $X_1$, $X_2$ ..., $X_K$ , K variables issues du recensement et également présente dans le questionnaire HID. On souhaite caler $\hat{\bar{Y}}_{pst2} = \sum_{k \in s_{hid}} w_k y_k$ sur **les moyennes nationales** de ces variables.

Il suffit pour cela de mettre en oeuvre une procédure de calage (par exemple CALMAR) calculant les nouveaux poids de calage $w_k^*$ . Ces derniers seront peu « éloignés » des anciens

poids, les $w_k$. Ainsi les bonnes propriétés de l'estimateur post-stratifié seront, du moins l'espère t-on, honorablement conservées.

## II- Les estimateurs régionaux

On se référera au § III pour le rappel du principe général de construction des estimateurs sur petits domaines.

*II-1 L'estimateur post-stratifié régional*

Soit une variable d'intérêt HID notée $Y$[8] dont on veut estimer la moyenne régionale, $\overline{Y}_R$. On a, en repérant les post-strates par l'indice $h$ :

$$\sum_{h=1..H} \frac{N_{Rh}}{N_R} \overline{Y}_{Rh}$$

Notons $s_{hidR}$ l'échantillon HID régional et $p_k$ la probabilité de sélectionner l'individu $k$. L'estimateur post-stratifié classique, $\hat{\overline{Y}}_{pst1R}$, s'écrit :

$$\hat{\overline{Y}}_{pst1R} = \sum_{postrate_h} \frac{N_{Rh}}{N_R} \frac{\sum_{k \in s_{hidR} \cap postrate_h} \frac{y_k}{p_k}}{\sum_{k \in s_{hidR} \cap postrate_h} \frac{1}{p_k}}$$

Malheureusement $s_{hidR}$ est trop petit pour espérer obtenir une précision raisonnable. On effectue donc une hypothèse liant le comportement de la région à celui d'un domaine plus important, ici, la France. L'hypothèse retenue est, quelque soit la région $R$ et quelque soit la post-strate $h$ :

$$\overline{Y}_{Rh} = \overline{Y}_h$$

On en déduit l'estimateur post-stratifié régional, $\hat{\overline{Y}}_{pst2R}$, inférant directement à partir de l'échantillon national HID :

$$\hat{\overline{Y}}_{pst2R} = \sum_{postrate_h} \frac{N_{Rh}}{N_R} \frac{\sum_{k \in s_{hid} \cap postrate_h} \frac{y_k}{p_k}}{\sum_{k \in s_{hid} \cap postrate_h} \frac{1}{p_k}}$$

Evidemment, les proportions régionales des groupes de handicap sont inconnues. On les estime grâce à l'échantillon régional VQSR. On obtient finalement comme estimateur, avant calage sur des variables de recensement :

[8] Par exemple : Y=1 si l'individu utilise du matériel audiovisuel adapté pour les malentendants (Question DAUDIO du questionnaire HID).

$$\hat{\bar{Y}}_{pst3} = \sum_{postrateh} \frac{\hat{N}_{Rh}}{\hat{N}_R} \frac{\sum_{k \in s_{hid} \cap postrateh} \frac{y_k}{p_k}}{\sum_{k \in s_{hid} \cap postrateh} \frac{1}{p_k}}$$

avec

$$\frac{\hat{N}_{Rh}}{\hat{N}_R} = \frac{\sum_{k \in s_{VQSR} \cap poststrateh} \frac{1}{p'_k}}{\sum_{k \in s_{VQSR}} \frac{1}{p'_k}}$$

Remarquons que l'estimateur $\hat{\bar{Y}}_{pst3}$ peut s'écrire sous la forme $\sum_{k \in s_{hid}} w_k y_k$

avec, si l'individu *k* appartient à la post-strate *h* :

$$w_k = \frac{\hat{N}_{Rh}}{\hat{N}_R} \frac{\frac{1}{p_k}}{\sum_{k \in s_{hid} \cap postrateh} \frac{1}{p_k}}$$

*II-2 L'estimateur post-stratifié régional redressé*

Soit $X_1$, $X_2$ ..., $X_K$ , K variables issues du recensement et également présente dans le questionnaire HID. On souhaite caler $\hat{\bar{Y}}_{pst3} = \sum_{k \in s_{hid}} w_k y_k$ sur **les moyennes régionales** de ces variables.

Il suffit pour cela de mettre en oeuvre une procédure de calage (par exemple CALMAR) calculant les nouveaux poids de calage $w_k^*$ . Ces derniers seront peu «éloignés » des anciens poids, les $w_k$. Ainsi, les bonnes propriétés de l'estimateur post-stratifié $\hat{\bar{Y}}_{pst3}$ seront, du moins l'espère t-on, honorablement conservées.

Exemple d'application :

Si l'on souhaite estimer la proportion d'individus sourds, parmi les individus de la région R, en utilisant « l'estimateur post-stratifié régional redressé », à savoir $\sum_{k \in s_{hid}} w_k^* y_k$

il faut donc poser $\begin{cases} y_k = 1 \text{ si l'individu k est sourd} \\ y_k = 0 \text{ sinon} \end{cases}$

Insistons sur le fait que l'estimateur utilise l'ensemble de l'échantillon HID et non la seule partie régionale de l'échantillon.

## III- Principe des estimations sur petits domaines

Compte tenu de la faible taille de l'échantillon HID dans chacune des régions, il est illusoire d'espérer obtenir une estimation régionale correcte relative à chacune des variables de l'enquête HID, si cette estimation est obtenue en inférant à partir de l'échantillon HID régional. Une méthode d'estimation sur petits domaines permet de construire des estimateurs régionaux inférant à partir de l' échantillon HID national.

Décrivons le mode de construction de ce type d'estimateurs.

*III.1 - Les modèles de comportement*

La construction des estimateurs sur petits domaines s'appuie sur **un modèle de comportement**, relatif à la variable d'intérêt Y, liant les petits domaines entre-eux.

Par exemple un modèle de comportement, s'appuyant sur la connaissance ou l'information auxiliaire Z, peut être :

modèle 1 : le comportement moyen d'une région R noté $\overline{Y}_R$ est identique au comportement moyen national $\overline{Y}$ .

modèle 2: le comportement moyen d'une région R est identique à celui des régions qui lui sont contiguës.

On peut affiner le modèle de comportement si l'on dispose d'une information auxiliaire Z plus précise. Par exemple, si l'on considère *H* catégories d'individus :

modèle 3 : le comportement moyen, dans une région, au sein d'une catégorie *h* est le même que le comportement moyen national pour cette même catégorie (ce qui revient à dire qu'il n'y a pas d'effet région au sein d'une même catégorie).

*III.2 - Les propriétés statistiques*

Mentionnons, dès à présent, les propriétés fondamentales d'un estimateur sur petits domaines utilisant un modèle de comportement.

- sa variance varie comme l'inverse de la taille de l'échantillon HID **élargi aux régions ayant un comportement lié à celui de la région d'intérêt.**

- si le modèle de comportement n'est pas exact, l'estimateur est biaisé et ce biais n'est pas réductible en deçà d'une borne strictement positive (i-e il est inutile d'augmenter la taille de l'échantillon HID pour espérer passer sous cette borne).

*III.3 - Exemples de construction*

Montrons, à partir des modèles de comportement 1 et 3, comment peut être construit un estimateur **d'une moyenne** sur une région donnée.

Soit R la région d'étude. On suppose que l'échantillon HID national, $s_{hid}$ est issu d'un tirage probabiliste quelconque et l'on note $p_k$ la probabilité de sélection de chaque individu k du champs de l'enquête.

**Sous le modèle 1**, plutôt que de prendre l'estimateur suivant, sans biais, et qui n'utilise pas le modèle :

$$\hat{\bar{Y}}_1 = \frac{1}{N_R} \sum_{k \in s_{hidR}} \frac{y_k}{p_k}$$

(avec $N_R$ population vraie de la région),

on peut lui préférer :

$$\hat{\bar{Y}}_{\text{mod1}} = \frac{1}{N} \sum_{k \in s_{hid}} \frac{y_k}{p_k}$$

(avec $N$ population nationale vraie)

qui utilise le modèle.

Cet estimateur présente une variance vraisemblablement plus faible car $s_{hidR} \subset s_{hid}$. Mais son biais vaut exactement :

$$E\left(\hat{\bar{Y}}_{\text{mod1}}\right) - \bar{Y} = \bar{Y}_R - \bar{Y}$$

**Sous le modèle 3**, supposons que l'on connaisse, en plus de l'information contenue dans le modèle, la structure vraie, selon la catégorie *h*, de la région R. On note $N_{Rh}$ le nombre total de personne de la région appartenant à cette catégorie et $N_R$ désigne le nombre total d'individu dans la région R.

Plutôt que de prendre l'estimateur post-stratifié classique,

$$\hat{\bar{Y}}_{pst1R} = \sum_{postrateh} \frac{N_{Rh}}{N_R} \frac{\sum_{\substack{k \in s_{hid} \cap postrateh \\ \cap régionR}} \frac{y_k}{p_k}}{\sum_{\substack{k \in s_{hid} \cap postrateh \\ \cap régionR}} \frac{1}{p_k}}$$

asymptotiquement sans biais, et qui n'utilise pas le modèle, on peut lui préférer :

$$\hat{\bar{Y}}_{pst2R} = \sum_{postrateh} \frac{N_{Rh}}{N_R} \frac{\sum_{k \in s_{hid} \cap postrateh} \frac{y_k}{p_k}}{\sum_{k \in s_{hid} \cap postrateh} \frac{1}{p_k}}$$

qui utilise le modèle. Cet estimateur présente une variance vraisemblablement plus faible car $s_{hidR} \subset s_{hid}$. Mais son biais vaut (asymptotiquement) :

$$E\left(\hat{\bar{Y}}_{\text{mod1}}\right) - \bar{Y} = \bar{Y}_R - \bar{Y}$$

# *ANNEXE V*

## Choix d'un modèle de comportement

(extrait du rapport de Valérie Albouy)

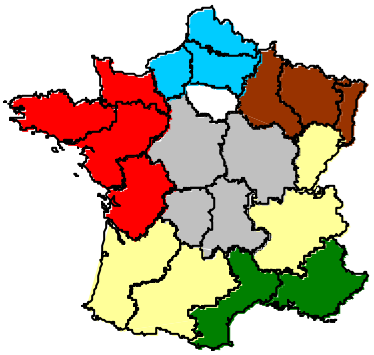
### 1. A la recherche d'un modèle de comportement national

Rappelons qu'établir un modèle de comportement consistait à rechercher pour une variable d'intérêt donnée les critères qui rendaient homogènes les comportements entre les régions.

Dans le cas de l'enquête HID, les critères testés ont été des critères socio-démographiques tels que l'âge, le sexe ou le milieu social ainsi que des niveaux de prévalence de handicap établis à partir des résultats de l'enquête VQS. Il s'agissait donc de trouver les critères qui permettaient de constituer des classes d'individus au sein desquelles **le comportement au niveau national et le comportement au niveau régional auraient été identiques.**

#### 1.1. Le choix de la méthode

Dans un premier temps, la méthode retenue a été de régresser par des **modèles de régression qualitatifs** de type LOGIT un certain nombre de variables d'intérêt issues de HID sur des caractéristiques socio-démographiques et des variables issues de VQS. La localisation géographique a été introduite sous forme d'indicatrice.



Les régions ont été regroupées pour disposer d'effectifs suffisants en terme de réponses HID. Les regroupements ont été les suivants : **Ile de France**, **Nord Ouest** (Haute Normandie + Picardie + Nord), **Ouest** (Basse Normandie + Bretagne, Pays de la Loire, Poitou-Charentes**), Sud Ouest** (Aquitaine et Midi-Pyrénées), **Centre** (Bourgogne, Centre, Limousin, Auvergne), **Sud Est** (Languedoc, PACA, Corse**), Est** (Franche Comté, Rhône Alpes), **Nord Est** (Champagne Ardennes, Lorraine, Alsace).

**L'objectif était de trouver une série de variables explicatives qui expliqueraient suffisamment les disparités géographiques pour les annuler. Concrètement, il s'agissait**

**de chercher un ensemble de variables explicatives qui rendraient les coefficients attachés aux variables géographiques non significativement différents de zéro.**

Remarquons que l'objectif d'estimation locale est de fournir des estimations à des niveaux géographiques plus fins que les groupes de régions ainsi constitués. Ce regroupement était nécessaire pour disposer d'effectifs suffisants pour construire des modèles fiables. L'hypothèse serait ensuite faite que le modèle de comportement, ainsi valable pour ces groupes de régions, serait toujours valable pour des zones plus fines. La validité de cette hypothèse pourra être testée sur l'Hérault qui dispose d'une extension HID.

### 1.2. Résultats des régressions sur cinq variables HID

Dans un premier temps, cinq variables d'intérêt ont été retenues : **DADAPT, CONFIN, C_AIDKI, RINVAL, RALLOC**.

**DADAPT** est une variable qui indique si la personne dispose de meubles ou d'équipements spécialement adaptés à son usage en raison de problèmes de santé, handicap ou infirmités.
**CONFIN** est une variable qui indique si en raison de problèmes de santé, handicap ou infirmités, la personne est confinée au lit, au fauteuil, ou à l'intérieur de son logement.
**C_AIDKI** indique s'il y a une ou des personnes qui aident régulièrement la personne interrogée à accomplir certaines tâches de la vie quotidienne en raison d'un handicap ou d'un problème de santé.
**RINVAL** indique si la personne a un taux d'incapacité ou d'invalidité reconnu.
**RALLOC** indique si la personne interrogée perçoit au moment où elle a été interrogée une allocation, pension ou un autre revenu en raison de ses problèmes de santé.

Le choix de ces indicateurs devait répondre à plusieurs objectifs. Tout d'abord, puisque l'un des buts premiers de l'enquête est d'évaluer la fréquence des différents types de difficultés liées à la santé, nous avons essayé de prendre des indicateurs révélateurs d'un 'niveau' de handicap ou d'une sévérité d'incapacité, et non de partir de la source de ce handicap, c'est-à-dire du type de déficience dont la personne est atteinte. L'objectif de l'enquête est également de fournir aux conseils généraux des fichiers de l'enquête nationale pondérés suivant la structure régionale en terme de strates retenues. L'idéal serait de fournir une pondération unique pour toutes les variables d'intérêt, c'est-à-dire de trouver un jeu de variables corrigeant des effets régionaux pour toutes les variables d'intérêt. Nous avons donc choisi de travailler au préalable sur des variables d'intérêt relativement génériques et pouvant résulter de plusieurs problèmes de santé différents pour ne pas retenir des variables explicatives propres à un problème de santé particulier. Par ailleurs, ces handicaps touchent également la population de manière très différente, ce qui assure que les populations touchées ne sont pas parfaitement identiques. Les taux de prévalence nationaux dans le fichier sont les suivants :

| prévalence de handicap observée sur l'échantillon | |
|---|---|
| DADAPT | 7,4 |
| CONFIN | 4,8 |
| C_AIDKI | 34,9 |
| RALLOC | 30,1 |
| RINVAL | 19,3 |

### 1.2.1. Résultats sur DADAPT

Une première régression effectuée avec la localisation géographique[9] comme seule variable explicative permet de vérifier qu'il existe bien des différences significatives entre les régions. Les paramètres estimés montrent que la propension à faire adapter son logement en raison de problèmes de santé est la moins forte à Paris, est la plus élevée dans l'Ouest[10]. Elle est plus forte dans le Nord-Ouest que dans le Sud et il n'y a pas de différences significatives entre le Sud[11], le Centre[12], et le Nord-Est.

Si l'on rajoute le groupe de prévalence de handicap[13] comme variable explicative, le pouvoir explicatif du modèle se trouve amélioré, mais les coefficients relatifs à la localisation restent significativement différents de zéro pour l'Ile de France, l'Ouest, et le Centre et ne diminuent pas beaucoup.

| | Ile de France | Ouest | Centre | Nord |
|---|---|---|---|---|
| coefficients avec région seule | -0,40 | 0,36 | 0,22 | 0,21 |
| Wald associé | 15,9 | 18 | 6,3 | 5,2 |
| | | | | |
| coefficient avec groupe | -0,38 | 0,34 | 0,21 | 0,14 |
| Wald associé | 13,5 | 15,4 | 5,4 | 2,2 |
| | | | | |
| coefficient avec groupe et âge | -0,32 | 0,38 | 0,27 | 0,31 |
| Wald associé | 9,4 | 19,4 | 8,9 | 11,1 |

Si l'on rajoute l'âge, non seulement certaines différences entre régions augmentent, mais la proportion de personnes ayant fait adapter leur logement dans le Nord redevient significativement différente de celle du Sud.

Ces premiers résultats soulèvent un certain nombre de questions. En effet, : **bien que l'on arrive à déterminer des variables permettant d'affiner les modèles logistiques, les coefficients attachés aux régions restent significatifs et ne diminuent pas forcément. Ceci reste vrai même si l'on inclut dans les modèles des variables issues de la pré-enquête VQS, censées déterminer avec précision des prévalences. Conditionnellement à ces variables, il reste donc un effet marginal élevé lié à la localisation géographique.**

***l'outil est-il parfaitement adapté ?***

Ce résultat peut provenir de ce que l'estimation du modèle se fait en déterminant de manière conjointes les divers coefficients. Pour le modèle, si deux variables $X$ et $Y$ sont colinéaires, estimer :

[9]La localisation géographique est introduite au niveau des groupes de régions définis page 29.

[10] Poitou Charentes, Bretagne, Pays de la Loire et Basse Normandie.

[11] PACA, Languedoc et Corse

[12] Auvergne, Bourgogne, Limousin et Centre

[13]Rappelons que ce groupe a été déterminé à partir des réponses à l'enquête Vie Quotidienne et Santé et doit classer les individus selon leur probabilité ou sévérité de handicap croissante.

$$\ln \frac{p_{X,Y}}{1-p_{X,Y}} = \hat{a} + \hat{b}X$$

est équivalent à estimer :

$$\ln \frac{p_{X,Y}}{1-p_{X,Y}} = \hat{a} + \hat{b}Y$$

ce qui ne permet pas de conclure sur l'effet 'résiduel' lié aux régions.

Certes, si la colinéarité entre les variables $X$ et $Y$ est parfaite, l'identification du modèle est impossible car la matrice de variances covariances n'est pas inversible. Mais admettons qu'il existe un lien fort entre le fait d'appartenir à une région et l'âge. **Ne sera-t-il pas équivalent pour le modèle d'estimer l'effet marginal de la région quand on raisonne à âge donné que d'estimer l'effet marginal de l'âge quand on raisonne à région donnée?**

***Ces résultats peuvent-ils provenir, même partiellement des comportements de non réponse ?***

Rappelons que le taux d'échec de l'enquête HID est de l'ordre de 20%. Le refus de répondre explique près de 60% de ces échecs.

La modélisation des comportements de non-réponse en fonction du groupe, du sexe, de l'âge et la région montre que, conditionnellement à ces trois premières variables, la probabilité de ne pas répondre à l'enquête va être sensiblement plus élevée en Ile de France et dans le Sud Ouest que dans les autres régions. La probabilité de souffrir d'un handicap pour un répondant et un non répondant n'est probablement pas homogène. Dès lors, les comportements de non-réponse vont tronquer les proportions de handicap observées, et les disparités en terme de non-réponse vont induire des disparités en terme de biais entre l'observation et la réalité entre les régions.

Certes, on pensait que les personnes qui souffraient d'un handicap étaient plus sensibilisées au sujet de l'enquête et avaient répondu plus facilement. Dès lors, la non réponse devrait entraîner une surestimation de la proportion réelle par la proportion observée, puisque les non répondants auraient plus de chances de ne pas souffrir du handicap. Ceci signifierait que les disparités réelles entre l'Ile de France et l'Ouest de la France sont en réalité plus fortes que ne le montrent les probabilités observées sur l'échantillon. Toutefois, il se peut que les non réponses aient des origines très différentes d'une région à l'autre. Si les proportions de handicap parmi les non répondants varient fortement d'une région à l'autre, cela va induire des disparités entre régions dans les troncatures des probabilités observées sur les répondants.

Toutefois, les modélisations des proportions de handicap font varier les estimations de proportion du simple au double d'une région à l'autre. S'il est troublant de constater que les régions se positionnent au sein des modèles modélisant la non réponse dans le même ordre qu'au sein des modèles déterminant les prévalences d'un handicap donné, s'il est étonnant de voir que le positionnement est le même pour des variables d'intérêt ayant des prévalence très éloignées, **on peut difficilement maintenir que la troncature des probabilités observées par la non réponse explique des prévalences variant du simple au double.**

***Ou les modèles sont-ils tout simplement incomplets ?***

Ces différences entre régions peuvent tout simplement provenir de comportements de réponse à VQS ou à HID différents. La réponse à la question de savoir si l'on a fait adapter son logement en raison de problèmes de santé contient une part de subjectivité. Il se peut aussi que les réponses au questionnaire VQS, distribué avec le questionnaire du recensement et sans le contrôle d'un enquêteur, aient pu entraîner des appréciations différentes suivant les régions des groupes de prévalence. Par contre, il est probable que les réponses au questionnaire beaucoup plus complet HID, faites en présence d'un enquêteur, renvoient à une réalité du handicap beaucoup plus homogène suivant les régions. Dès lors, la probabilité de déclarer souffrir d'un handicap conditionnellement à l'appartenance à une sévérité de handicap présumée à partir des réponses VQS peut varier d'une région à l'autre.

Enfin, la difficulté à trouver un comportement national peut provenir de ce que certains facteurs influençant l'état de santé –et par suite ses conséquences sur la vie quotidienne- sont difficiles à prendre en compte dans les modèles : c'est le cas du climat –l'humidité peut jouer sur l'arthrose et donc sur la mobilité des personnes-, ou des habitudes alimentaires. La perception des limitations d'activité peut également être très différente d'une région à l'autre. De manière plus générale, les conséquences des problèmes de santé sur la vie sociale se trouvent à la croisée d'une situation personnelle et d'un environnement social. Peu de descriptifs de l'environnement social ont été introduits dans les modèles. Par exemple, seules les personnes vivant à domicile faisaient l'objet de l'enquête. Les personnes vivant en institution ont fait l'objet d'une autre enquête sur laquelle nous n'avons pas travaillé. Or la décision de rester ou non à domicile dépend certainement, outre de l'état de santé, de facteurs tels que la qualité de l'offre en institution, la possibilité d'un accompagnement à domicile, ou de facteurs culturels.

### 1.2.2. Test d'égalité global des coefficients de la régression

#### *A. Avec groupe sexe et âge*

Pour contourner les incertitudes concernant le comportement de la régression logistique en cas de colinéarité entre les variables, nous avons estimé un modèle par région et testé (par un test du rapport de vraisemblance) l'égalité des coefficients des régressions. Ce test a été fait avec les variables explicatives suivantes : le groupe, le sexe et l'âge (tranches de vingt ans[14]). Les estimations conduites sur les huit régions permettent d'obtenir $\hat{b}_1,..\hat{b}_8$.

**Sous l'hypothèse** $\hat{b}_1 =...= \hat{b}_8$**,** $-2(\ln \hat{b} - \sum_{i=1}^{8} \ln \hat{b}_i)$ **doit suivre un** $c^2_{77}$**.**

(où 77 représente le nombre de contraintes qu'impose le test, soit 11-dimension des vecteurs $\hat{b}_i$-multiplié par 7, et $\hat{b}$ le paramètre estimé sur le modèle complet).

Le calcul de $-2(\ln \hat{b} - \sum_{i=1}^{8} \ln \hat{b}_i)$ donne 140, alors que le quartile à 0.05 d'un $c^2_{77}$ est 98. **L'hypothèse d'égalité des coefficients est donc rejetée.**

#### *B. Avec d'autres variables VQS*

[14] Faire intervenir l'âge en tranches décennales n'apporte qu'un gain faible aux régressions.

L'expérience a été reconduite en ajoutant dans la liste des variables explicatives un certain nombre de réponses à l'enquête VQS. Jusqu'alors, l'information contenue dans ces réponses était reprise de manière synthétique dans la variable GROUPE, sensée représenter la position de l'individu sur une échelle de handicap à partir de ses réponses VQS. Toutefois, l'information était suffisamment diluée pour que l'appartenance à un groupe puisse signifier des types de handicap très différents.

La sélection des variables VQS les plus pertinentes pour les cinq variables d'intérêt retenues s'est faite dans la régression logistique par l'équivalent d'une procédure backward, hormis le fait que la dimension explicative relative à une variable était enlevée globalement (alors que la procédure backward peut laisser certaines modalités d'une variable explicative et en enlever d'autres). Finalement, les variables VQS retenues ont été les suivantes **: une variable indiquant si la personne interrogée a un besoin très supérieur à ce que devrait requérir son âge d'une aide d'une autre personne dans la vie quotidienne, une variable indiquant si la personne a fait une demande de reconnaissance de handicap ou d'invalidité, une variable indiquant si la personne a des problèmes pour s'habiller ou se déshabiller, et une indiquant si elle a des difficultés à remplir un questionnaire simple**. Outre le sexe et l'âge, un **indicateur d'un niveau d'étude** a de plus été rajouté.

Sous l'hypothèse d'égalité des coefficients estimés par région,

$$-2(\ln \hat{b} - \sum_{i=1}^{8} \ln \hat{b}_i) \text{ doit suivre un } c^2_{154}.$$

Le calcul de $-2(\ln \hat{b} - \sum_{i=1}^{8} \ln \hat{b}_i)$ donne 218 alors que le quartile à 5% d'un $c^2_{154}$ est de 184. Là encore, **l'hypothèse d'égalité des coefficients est rejetée.**

### 1.2.3. Résultats sur les quatre autres variables HID

Les expériences menées sur la variables ADAPT ont été reconduites sur les variables CONFIN, C_AIDKI, RINVAL, RALLOC[15].

Les résultats sont similaires : les coefficients liés aux régions dans les régressions logistiques sont peu sensibles à l'introduction dans les modèles de variables par ailleurs pertinentes en terme de pouvoir explicatif. On n'observe aucune diminution nette globale de ceux-ci, et toutes les tentatives pour trouver un ensemble de variables explicatives rendant ces coefficients liés aux régions proches de zéro ont échoué. Les différences de proportion de handicap, conditionnellement aux variables explicatives introduites, restent, d'après les modèles, relativement importantes puisqu'elles varient du simple au double.

[15] **CONFIN** est une variable qui indique si en raison de problèmes de santé, handicap ou infirmités, la personne est confinée au lit, au fauteuil, ou à l'intérieur de son logement.
**C_AIDKI** indique s'il y a une ou des personnes qui aident régulièrement la personne interrogée à accomplir certaines tâches de la vie quotidienne en raison d'un handicap ou d'un problème de santé.
**RINVAL** indique si la personne a un taux d'incapacité ou d'invalidité reconnu.
**RALLOC** indique si la personne interrogée perçoit au moment où elle a été interrogée une allocation, pension ou un autre revenu en raison de ses problèmes de santé.

### 1.3. Conclusion

On ne parvient pas à définir des caractéristiques qui définissent des groupes de personnes au sein desquels le risque de souffrir d'une incapacité ou d'un handicap est homogène sur toute la France. Par contre, les régions[16] se positionnent souvent de la même manière quand on passe d'une variable expliquée à une autre, alors même que les prévalences de ces variables, et les types de handicap qu'elles décrivent sont très différents. Cela porte à croire qu'il serait possible de constituer des groupes de régions homogènes pour la plupart des variables expliquées. Un deuxième axe de travail a donc été **la constitution de classes de régions.** Cette solution aurait l'avantage de ne pas biaiser les estimateurs en appliquant un modèle de comportement qui ne correspondrait pas au comportement local.

## 2. Constitution de groupes de régions

### 2.1. Par test d'égalité du vecteur des coefficients de régression logistique

Devant la difficulté à trouver des variables annulant les disparités régionales, une première solution a été de constituer des groupes de régions en effectuant des régressions sur des sous ensembles de régions. La méthode était de prendre les régions comme référence une à une et de recommencer les régressions en ne gardant que les régions qui n'avaient pas de coefficient significativement différent de zéro au tour précédent.

Ces tests ont été réalisés sur les cinq variables d'intérêt citées plus haut : DADAPT, CONFIN, C_AIDKI, RINVAL, RALLOC[11]. Le vecteur des variables explicatives était volontairement le plus large possible étant donné que certaines variables explicatives jouent pour certaines des variables d'intérêt et ne jouent pas pour d'autres. Il comprenait le groupe VQS, l'état matrimonial, l'âge (en tranche de vingt ans), le sexe, le type de logement, le milieu social, le diplôme (en quatre modalités : moins de quinze ans, sans diplôme, primaire, secondaire et supérieur), la tranche d'unité urbaine habitée (en trois modalités : rural, petite ville, grande ville), le nombre de personnes dans le ménage (une, deux, trois, plus de quatre), et la variable géographique.

Dans certains cas, étant donnée la colinéarité de beaucoup de ces variables, il a été difficile de séparer les effets marginaux liés à chaque variable (par exemple, diplôme 0 - moins de quinze ans- et première tranche : d'âge moins de vingt ans) et le modèle n'a pas pu converger. Toutefois, certains groupes de régions se dessinent :

l'Ile de France et le Sud Ouest,
le Nord, le Sud Ouest, le Sud et l'Est
l'Ouest, le Centre et le Nord Est
le Sud Ouest, le Nord et le Sud
le Centre, l'Est et le Nord Est
le Sud, le Sud Ouest et le Nord
l'Est, le Sud Ouest et le Centre
le Nord Est, le Centre et l'Ouest

[16] Non pas au sens administratif du terme mais au sens de l'échelon géographique où nous avons travaillé.

Les résultats sont synthétisés dans le tableau suivant :

| | dadapt | confin | c_aidki | rinval | ralloc |
|---|---|---|---|---|---|
| **l'Ile de France et le Sud Ouest,** | NC | NC | 0 | 0 | 0 |
| **le Nord Ouest, le Sud Ouest, le Sud et l'Est** | 1 | 0 | 0 | 2 | 0 |
| **l'Ouest, le Centre et le Nord Est** | 0 | 2 | 1 | 0 | 0 |
| **le Sud Ouest, le Nord Ouest et le Sud** | 0 | 0 | 0 | 0 | 1 |
| **le Centre, l'Est et le Nord Est** | 0 | 1 | 0 | 0 | 0 |
| **le Sud, le Sud Ouest et le Nord Ouest** | 1 | 0 | 0 | 1 | 0 |
| **l'Est, le Sud Ouest et le Centre** | 0 | 1 | 1 | 1 | 0 |
| **le Nord Est, le Centre et l'Ouest** | 2 | 2 | 2 | 2 | 2 |

NC = non convergent
0 = toutes les régions ont des coefficients non significatifs
1 = toutes les régions ont des wald <7
2 = une ou plusieurs régions ont des wald >7
On voit qu'il est particulièrement difficile de trouver une zone comprenant le Nord Est qui soit homogène du point de vue du comportement. Par contre, les Sud Est et le Centre se fondent relativement bien dans des zones plus grandes.
L'inconvénient de cette méthode est que les regroupements ne se font pas forcément de la même manière lorsqu'on change la région de référence dans la régression logistique.

### 2.2. Par classification hiérarchique ascendante sur ADAPT

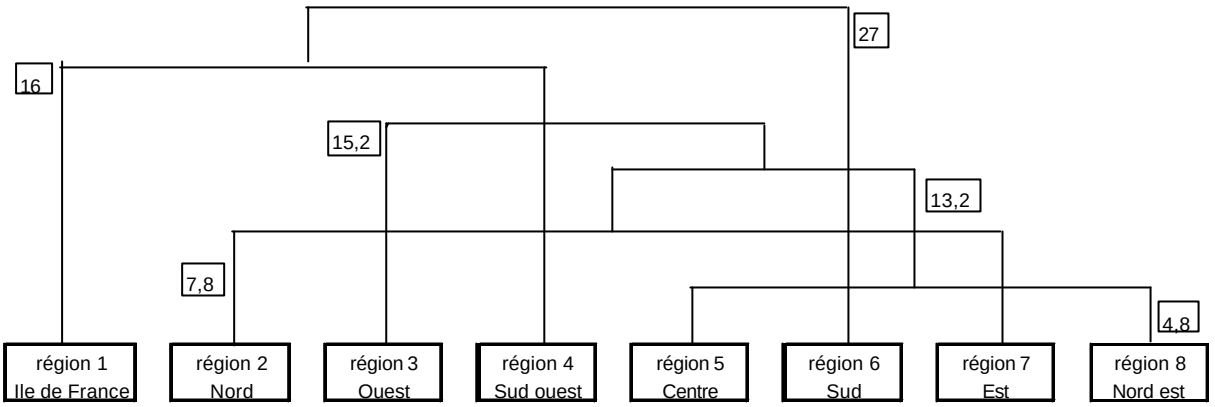
Pour consolider ces résultats, une classification ascendante a ensuite été réalisée constituant des classes de régions. Au départ, les régressions sont effectuées région par région, puis à chaque étape, les deux régions les plus proches sont rassemblées. Le modèle est à chaque fois ré-estimé pour cette nouvelle 'super-région' constituée. Le critère de distance utilisé est toujours le rapport de vraisemblance.

#### 2.2.1. avec groupe sexe et âge

Une première version a été effectuée avec les variables groupe, sexe et âge. Rappelons que l'hypothèse d'égalité des coefficients estimés région par région avait été rejetée. L'hypothèse d'identité des modèles sur la région A et sur la région B est acceptée tant que :

$$-2\left[\ln \hat{b}_{A+B} - (\ln \hat{b}_A + \ln \hat{b}_B)\right]$$ n'excède pas le quartile à 5% d'un $c_{11}^2$, soit 20

L'arbre obtenu est le suivant :




Les chiffres encadrés correspondent à deux fois la perte en terme de log vraisemblance au niveau d'agrégation où on se trouve.
Le dernier niveau d'agrégation est refusé puisque la perte en terme de log vraisemblance excède la limite du quartile à 5% du $c_{11}^{2}$.

**On voit finalement que l'on obtient trois classes homogènes :**

- **une première regroupant l'Ile de France et le Sud ouest**
- **une deuxième regroupant l'Ouest, le Centre, l'Est, le Nord et le Nord est**
- **une troisième avec le Sud tout seul.**

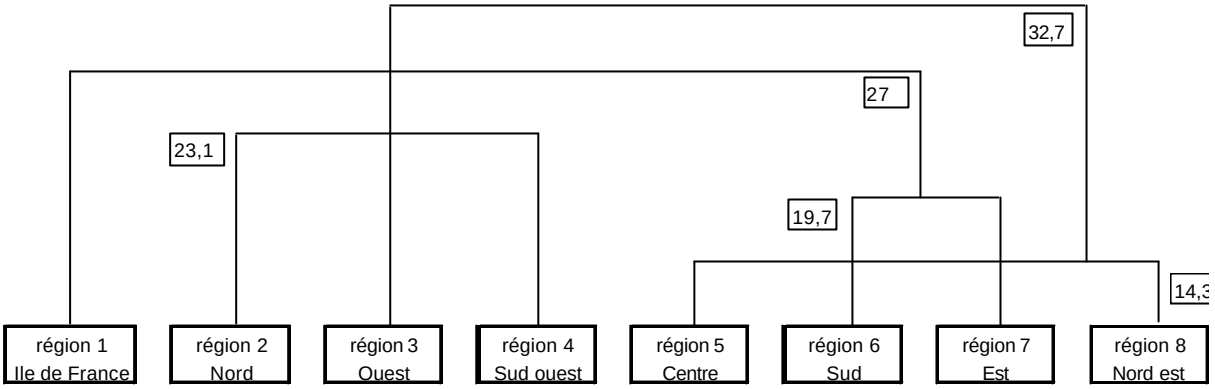
De tels groupes permettent une estimation d'une précision acceptable. En effet, le nombre de réponses à l'enquête HID est de 3633 dans la première classe constituée, 10422 dans la deuxième, 2890 dans la troisième. **Ainsi, s'il s'avère que l'on ne peut pas trouver de modèle de comportement au niveau national, on pourra appliquer la méthode d'estimation sur petit domaine sur ces trois classes de régions. Il s'agira, au lieu de repondérer l'échantillon national de manière à le rendre conforme en terme de structure par groupe, sexe et âge à la zone d'étude de repondérer la super-région à laquelle appartient la zone d'étude**. L'inconvénient de cette méthode serait la nécessité de fournir plusieurs fichiers différents aux utilisateurs ultérieurs. De plus, il est évident que si le gain en biais peut être appréciable, la perte en terme de variance est considérable.

### 2.2.2. avec d'autres variables VQS que groupe

Le même travail a été refait **en incorporant dans les modèles de régression les variables issues de VQS** qui paraissent globalement les plus pertinentes pour les cinq variables d'intérêt étudiées. **L'objectif est de tester la robustesse des classes constituées lorsqu'on introduit un modèle plus complet**.

Rappelons que ces variables sont, outre une variable indicatrice du niveau d'études, une variable indiquant si la personne interrogée a un besoin très supérieur à ce que devrait requérir son âge d'une aide d'une autre personne dans la vie quotidienne, une variable indiquant si la personne a fait une demande de reconnaissance de handicap ou d'invalidité, une variable indiquant si la personne a des problèmes pour s'habiller ou se déshabiller, et une indiquant si elle a des difficultés à remplir un questionnaire simple comme un chèque.

L'arbre qui se forme avec le critère de minimisation de la perte de log vraisemblance est le suivant :




Le dernier niveau d'agrégation est refusé car la perte de log vraisemblance excède le niveau du quartile à 5% d'un chideux à 22 degrés de liberté. Finalement, quatre classes homogènes se forment :

- une première avec l'Ile de France, le Sud et l'Est,
- une deuxième avec le Nord et le Sud ouest
- une troisième avec l'Ouest
- une quatrième avec le Centre et le Nord est.

Cette classification est loin d'être optimale. En effet, les répondants HID se répartissent comme suit :

pour la première région, plus de 7000 répondants
pour la deuxième, 3300 répondants HID
pour la troisième, moins de 2500,
pour la quatrième, 4119.

L'inconvénient par rapport à la classification précédente est que l'on est obligé de garder quatre classes de 'super-régions', perdant plus en terme de variance sans être assuré pour autant que le biais induit par le modèle de comportement sera moindre.

Dès lors, nous avons voulu voir jusqu'où la première classification en trois régions était acceptable avec les variables VQS. Si les deux premières fusions, entre le Centre et le Nord d'une part, et entre le Nord et l'Est d'autre part, passent le test de la log vraisemblance, l'agrégation au niveau supérieur ne passe pas. En fait, il est seulement possible d'obtenir quatre classes de régions qui sont les suivantes :

- une première regroupant l'Ile de France et le Sud ouest
- une deuxième regroupant l'Ouest, l'Est et le Nord,
- une troisième avec le Centre et le Nord est,
- une quatrième avec le Sud tout seul.

**Par rapport à la première classification obtenue, on voit que la deuxième classe s'est scindée en deux classes distinctes. Les effectifs HID sont toutefois mieux répartis dans cette partition en quatre classes que dans celle obtenue en minimisant les pertes de log vraisemblances.**

### 2.2.3. Par classification hiérarchique ascendante sur les autres variables HID

La solution de la constitution de classes homogènes du point de vue du comportement pourrait n'être retenue que s'il s'avérait que ces classes étaient les mêmes pour toutes les variables d'intérêt. Le contraire supposerait que l'on fournisse aux utilisateurs des fichiers avec des pondérations différentes suivant les variables de travail.

Le travail de classification a donc été fait pour quatre autres variables que DADAPT : RALLOC, AHANDI, RINVAL et C_AIDKI[17].Les classes obtenues pour les différentes variables sont les suivantes :

- pour RALLOC :
  - une première classe avec l'Ile de France,
  - une deuxième avec l'Ouest, le Centre et le Nord Est,
  - une troisième avec le Nord Ouest, le Sud Ouest, le Sud et l'Est.

- pour AHANDI :
  - une première classe avec l'Ile de France et le Nord Est,
  - une deuxième avec l'Ouest, le Sud Ouest et l'Est,
  - une troisième avec le Nord Ouest et le Centre,
  - une quatrième avec le Sud.

- pour C_AIDKI :
  - une première classe avec l'Ile de France et le Sud Ouest,
  - une deuxième avec le Nord Ouest et le Nord Est,
  - une troisième avec l'Ouest,
  - une quatrième avec le Centre, le Sud et l'Est.

- pour RINVAL :
  - une première classe avec l'Ile de France, le Sud Ouest, le Nord Ouest et le Sud.
  - une deuxième avec l'Ouest, le Centre et l'Est
  - une troisième avec le Nord Ouest.

Rappelons par ailleurs que les classes obtenues pour DADAPT étaient les suivantes :

- une première classe avec l'Ile de France et le Sud Ouest,
- une deuxième avec l'Ouest, le Centre, l'Est, le Nord Ouest et le Nord Est,
- une troisième avec le Sud.

On voit donc que les classes qui se forment par le critère de la perte minimale de la log vraisemblance ne sont pas spontanément les mêmes quelle que soit la variable d'intérêt. Toutefois, l'examen des proximités en terme de vraisemblance montre que certains regroupements sont possibles pour plusieurs variables.

---

[17] **C_AIDKI** indique s'il y a une ou des personnes qui aident régulièrement la personne interrogée à accomplir certaines tâches de la vie quotidienne en raison d'un handicap ou d'un problème de santé.
**RINVAL** indique si la personne a un taux d'incapacité ou d'invalidité reconnu.
**RALLOC** indique si la personne interrogée perçoit au moment où elle a été interrogée une allocation, pension ou un autre revenu en raison de ses problèmes de santé.
**AHANDI** indique si la personne déclare rencontrer dans la vie de tous les jours des difficultés physiques, sensorielles, intellectuelles ou mentales.

Une première classification en trois classes a été testée sur les cinq variables. Les trois classes ont été formées de manière discrétionnaire de la façon suivante :

- une première classe avec l'Ile de France et le Sud Ouest,
- une deuxième avec le Sud et le Nord Ouest,
- une troisième avec l'Ouest, le Centre, l'Est et le Nord Est.

Il s'agit de voir si ces trois classes sont homogènes du point de vue de chacune des variables.

pour RALLOC :
- l'homogénéité est acceptable pour la première classe,
- l'homogénéité est acceptable pour la deuxième classe,
- l'homogénéité est acceptable pour la troisième classe.

pour AHANDI :
- la première classe n'est pas homogène, la différence de log-vraisemblance vaut 11 et la limite d'acceptation est 9,8.
- la deuxième classe n'est pas homogène (11,8)
- l'homogénéité n'est pas acceptable pour la troisième classe (28,1 pour une limite d'acceptation du test à 95% à 23,7).

pour C_AIDKI :
- l'homogénéité est acceptable pour la première classe,
- l'homogénéité est acceptable pour la deuxième classe,
- l'homogénéité n'est pas acceptable pour la troisième classe (32,9 pour une limite d'acceptation à 23,7).

pour RINVAL :
- l'homogénéité est acceptable pour la première classe,
- l'homogénéité est acceptable pour la deuxième classe,
- l'homogénéité est acceptable pour la troisième classe.

pour DADAPT :
- l'homogénéité est acceptable pour la première classe,
- la deuxième classe n'est pas homogène (11,6)
- l'homogénéité est acceptable pour la troisième classe.

Par ailleurs, plusieurs variantes de regroupements ont été testées : aucune ne passe le test d'homogénéité pour les cinq variables. Toutefois, l'homogénéité semble globalement plus grande du point de vue des cinq variables en faisant les regroupements suivants :

- une première classe avec l'Ile de France et le Sud Ouest,
- une deuxième avec le Sud et le Nord Ouest et l'Est,
- une troisième avec l'Ouest, le Centre et le Nord Est.

La deuxième classe devient homogène du point de vue de DADAPT, et le test d'égalité des coefficients est refusé de peu pour AHANDI dans la troisième classe (écart des log-vraisemblance de 18,7 pour une limite d'acceptation du test à 95% de 17).

### 2.3. Conclusion

Si l'on choisit un ensemble de variables de stratifications, on peut décomposer la proportion de personnes handicapées dans une région de la manière suivante :

$$p_R = \sum_h \frac{N_{Rh}}{N_R} p_{Rh}$$

où $h$ indice les strates, $N_{Rh}$ est l'effectif de la strate $h$ dans la région $R$, $N_R$ est l'effectif régional total de la région $R$ et $p_{Rh}$ est la proportion de handicap dans la région $R$ et la strate $h$.

Si l'on décompose de la même manière la proportion nationale, on obtient :

$$p = \sum_h \frac{N_h}{N} p_h$$

On voit donc qu'il y a **deux sources de divergence e**ntre la proportion nationale et la proportion régionale :

**au niveau de l'importance relative des strates : c'est ce qu'on appelle une différence de structure**

**au niveau des proportion de handicap au sein de chaque strate : ce sont alors des différences de comportement.**

La méthode d'estimations sur petits domaines reposait sur l'idée que l'on pouvait trouver une stratification pertinente qui permettrait de ne plus avoir de spécificité en terme de comportement au sein des strates. **Que ce soit par rapport à l'ensemble du territoire ou par rapport à des zones moins étendues, nous avons vu qu'il était difficile de trouver une stratification qui gommait les spécificités régionales en terme de comportement au sein des strates, que ces spécificités soient par rapport à un comportement national ou par rapport à un regroupement de régions.**

L'application d'un comportement au sein de strates régionales risque donc d'introduire un biais dans l'estimation Toutefois**, le travail au niveau individuel, s'il a mis en évidence des spécificités locales dans le comportement des individus, ne permettait pas de mesurer l'incidence de ces spécificités**. **Le travail au niveau macro, c'est-à-dire au niveau des strates permet de clarifier ce point**. Il a consisté à essayer de séparer les différences structurelles des différences comportementales pour chaque région. Le but était de trouver les variables de structure qui minimisaient les différences de comportement au sein des strates.

# *ANNEXE VI*

## Définition des post-strates

Pour éviter l'émiettement de la population de l'échantillon HID dans des « strates de comportement homogène » du fait du croisement de critères trop nombreux, on a tenté de réduire le nombre de post-strates.

En effet, si on cherche à définir des comportements nationaux à travers plusieurs variables explicatives, on risque de rencontrer des effectifs trop faibles dans un grand nombre de strates en raison de la multiplication des croisements. La réduction du nombre de croisements permet aux populations interrogées d'être assez nombreuses au niveau national pour pouvoir dégager des comportements stables par strate.

Par exemple, d'après le modèle de comportement le plus fréquent, qui croise le sexe, le groupe VQS, l'âge, la tranche d'unité urbaine et (ou) le type de logement [18], on définit environ 2 x 6 x 5 x 3 x 2 = 360 strates [19] pour 17 000 répondants HID. L'idée est donc de procéder à un regroupement raisonné des strates qui présentent soit un nombre d'observations trop faible, soit des comportements moyens observés assez proches.

Ces regroupements sont testés par des modèles de type LOGIT reliant quelques variables d'intérêt issues de HID à des caractéristiques socio-démographiques et à une variable mesurant le degré de handicap, tirée de l'enquête VQS.

Une variable d'intérêt a été privilégiée : le recours éventuel à une aide régulière en raison d'un problème de handicap (variable aidki). En effet, cette variable couvre des formes variées d'incapacités et son taux de prévalence dans le fichier HID (non pondéré) est assez important (34,8 %). Pour chacun des six " groupes de handicap " construits à partir des réponses à VQS, 60 situations sont envisageables selon le sexe, le groupe d'âge, la tranche d'unité urbaine ou le type de logement (individuel ou collectif). On teste l'hypothèse de la nullité des coefficients: cette hypothèse étant rejetée lorsque la statistique de Wald dépasse le seuil de 4 pour une significativité à 5 %.

Assez fréquemment des regroupements de strates voisines s'imposent chez les individus de moins de 60 ans parce que le comportement étudié est souvent absent de la strate : cette situation est révélée par des valeurs très fortes de l'écart-type estimé. Ce problème est parfois combiné à une population de référence trop peu nombreuse : ce cas est fréquemment rencontré dans les croisements entre commune rurale et habitat collectif. Toujours parmi les personnes de moins de 60 ans, le test de nullité des coefficients est souvent vérifié en ce qui

[18] En l'absence d'informations croisées avec la variable « groupe VQS », les caractéristiques purement sociales ne pourront être introduites qu'ultérieurement, par un calage sur les marges du RP.

[19] 180 ou 120 strates, selon les cas, si on ne retient que l'une ou l'autre des 2 variables, tranche d'unité urbaine et type de logement.

concerne les subdivisions par tranche d'unité urbaine et type de logement. De façon générale, ces deux variables étant étroitement liées, le choix a été fait de privilégier la variable " tranche d'unité urbaine " dans la mesure où elle permet une meilleure discrimination des comportements.

Pour les groupes VQS de 1 à 5, seule subsiste la distinction selon le sexe parmi les populations âgées de moins de 60 ans. Au-delà de 60 ans, les distinctions selon le sexe et le groupe d'âge sont retenues. Dans le groupe 6 les croisements entre sexe et tranche d'âges se conjuguent avec le critère de taille de la commune (en 3 postes) entre 20 et 80 ans.

En résumé, les cinq premiers groupes de handicaps totalisent 30 strates. Le dernier groupe qui couvre les handicaps les plus sévères est subdivisé en 22 strates : le sexe et la tranche d'unité urbaine pour trois des cinq grands groupes d'âges (20-40 ans, 40-60 ans, 60-80 ans), et uniquement le sexe pour les deux autres ( moins de 20 ans et 80 ans ou plus). On est donc passé des 360 strates initiales à un regroupement en 52 strates qui constituent les modalités d'une nouvelle variable.

La répartition de l'échantillon HID selon ces modalités s'effectue comme suit pour les groupes 1 à 5 :

| GROUPE VQS | GROUPE D'AGES | TAILLE DE LA COMMUNE | SEXE | NOMBRE DE POST-STRATES | EFFECTIF DANS L'ECHANTILLON HID NATIONAL |
|---|---|---|---|---|---|
| **Groupe n°1** | moins de 60 ans | | Homme | (6) | 625 |
| | | | Femme | | 672 |
| | 60-80 ans | | Homme | | 617 |
| | | | Femme | | 772 |
| | 80 ans ou plus | | Homme | | 107 |
| | | | Femme | | 143 |
| **Groupe n°2** | | Chacun de ces groupes se décompose de la même façon que le groupe n°1 | | (4 x 6 = 24) | 1335 |
| **Groupe n°3** | | | | | 2037 |
| **Groupe n°4** | | | | | 1773 |
| **Groupe n°5** | | | | | 3634 |

tandis que le groupe 6 fait l'objet d'un éclatement plus important :

| GROUPE VQS | GROUPE D'AGES | TAILLE DE LA COMMUNE | SEXE | NOMBRE DE POST-STRATES | EFFECTIF DANS L'ECHANTILLON HID NATIONAL |
|---|---|---|---|---|---|
| **Groupe n°6** | moins de 20 ans | | Homme | (2) | 328 |
| | | | Femme | | 203 |
| | 20-40 ans | c.rurales | Homme | (6) | 108 |
| | | | Femme | | 75 |
| | | c. urbaine<100 000 | Homme | | 176 |
| | | | Femme | | 119 |
| | | c. urbaine>=100 000 | Homme | | 226 |
| | | | Femme | | 160 |
| | 40-60 ans | c.rurales | Homme | (6) | 306 |
| | | | Femme | | 174 |
| | | c. urbaine<100 000 | Homme | | 354 |
| | | | Femme | | 290 |
| | | c. urbaine>=100 000 | Homme | | 419 |
| | | | Femme | | 392 |
| | 60-80 ans | c.rurales | Homme | (6) | 252 |
| | | | Femme | | 207 |
| | | c. urbaine<100 000 | Homme | | 271 |
| | | | Femme | | 243 |
| | | c. urbaine>=100 000 | Homme | | 314 |
| | | | Femme | | 314 |
| | 80 ans ou plus | | Homme | (2) | 122 |
| | | | Femme | | 177 |
| **TOTAL** | | | | **52** | 5 230 |

# *ANNEXE VII*

## Estimation de la variance de l'estimateur régional

*(extrait du rapport de Valérie Albouy)*

### ***Une première approximation***

**Dans une première approche, on peut considérer que l'estimation des effectifs régionaux des post-strates et celle de l'effectif global n'interviennent qu'au second ordre dans la variance** de l'estimateur régional post-stratifié. En effet, on sait que :

$$V(\hat{N}_{Rh}) << V(\hat{\bar{Y}}_h)$$

et

$$V(\hat{N}_R) << V(\hat{\bar{Y}}_h)$$

On peut donc considérer que $\hat{N}_{Rh}$ et $\hat{N}_R$ sont constants par rapport à la variable aléatoire $\hat{\bar{Y}}_h$.

Dès lors, on commence par évaluer la variance de

$$\boxed{\hat{\bar{Y}}_{R1} = \sum_{h=1..H} \frac{N_{Rh}}{N_R} \hat{\bar{Y}}_h.}$$

où $N_{Rh}$ et $N_R$ sont les effectifs régionaux 'vrais' de la post-strate *h* et de l'ensemble de la région et où $\hat{\bar{Y}}_h$ est toujours l'estimateur de la moyenne de la variable d'intérêt sur la post-strate *h.*

#### 1. Rappel de l'expression de la variance pour l'estimateur d'Horvitz-Thompson

**De manière générale**, rappelons que si $\hat{\bar{Z}}_{HT}$ est l'estimateur de Horvitz-Thompson de $\frac{1}{N}\sum_{k\in U} z_k$, soit donc $\hat{\bar{Z}}_{HT} = \frac{1}{N}\sum_{k\in s} \frac{z_k}{p_k}$, sa variance peut s'écrire quel que soit le plan de sondage sous la forme :

$$V(\hat{\bar{Z}}_{HT}) = \frac{1}{N^2} \sum_{k\in U}\sum_{l\in U} \Delta_{kl} \frac{z_k}{p_k}\frac{z_l}{p_l} \qquad (1)$$

avec $\Delta_{kl} = p_{kl} - p_k p_l$ *si* $k \neq l$

$\Delta_{kl} = p_k(1-p_k)$, *si* $k=l$.

On calcule une estimation de la variance en appliquant cette formule sur les individus présents dans l'échantillon, soit, en calculant

$$\tilde{V}(\hat{\bar{Z}}_{HT}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{p_{kl}}\frac{z_k}{p_k}\frac{z_l}{p_l}$$

où par convention $p_{kk} = p_k$.

**Si de plus l'échantillon tiré est de taille fixe *n*** (ce qui est le cas dans l'enquête HID), la formule de la variance peut également s'écrire :

$$V(\hat{\bar{Z}}_{HT}) = -\frac{1}{2N^2}\sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}}\Delta_{kl}\left(\frac{z_k}{p_k} - \frac{z_l}{p_l}\right)^2 \qquad (2)$$

Dès lors, on calcule une estimation de la variance en appliquant cette formule sur les individus présents dans l'échantillon, soit, en calculant :

$$\tilde{V}(\hat{\bar{Z}}_{HT}) = -\frac{1}{2N^2}\sum_{k\in s}\sum_{\substack{l\in s\\ l\neq k}}\frac{\Delta_{kl}}{p_{kl}}\left(\frac{z_k}{p_k} - \frac{z_l}{p_l}\right)^2$$

## 2. Application au cas de l'estimateur post-stratifié avec modèle de comportement

*cas d'un estimateur post-stratifié*

On peut montrer que **pour obtenir la variance d'un estimateur post-stratifié**, il faut appliquer les formules décrites plus haut non pas à la variable d'intérêt elle-même **mais à l'écart** (noté e) **entre cette variable et sa moyenne sur la strate.**
Ainsi, un estimateur de la forme,

$$\hat{\bar{Z}} = \sum_{h=1}^{H}\frac{N_h}{N}\hat{\bar{Z}}_h$$

où $$\hat{\bar{Z}}_h = \frac{\sum_{k\in s\cap strh}\frac{z_k}{p_k}}{\sum_{k\in s\cap strh}\frac{1}{p_k}}$$

aura pour variance,

dans le cas général : $V(\hat{\bar{Z}}_{HT}) = \frac{1}{N^2}\sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}}\Delta_{kl}\frac{e_k}{p_k}\frac{e_l}{p_l}$,

et dans le cas d'un plan de sondage de taille fixe : $V(\hat{\bar{Z}}_{HT}) = -\frac{1}{2N^2}\sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}}\Delta_{kl}\left(\frac{e_k}{p_k} - \frac{e_l}{p_l}\right)^2$

avec $e_k$, écart de l'individu *k* par rapport à la moyenne de la variable sur sa strate.

On peut donc calculer **un estimateur de la variance** avec les formules :

$$\text{dans le cas général : } \tilde{V}(\hat{\bar{Z}}_{HT}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{p_{kl}}\frac{\hat{e}_k}{p_k}\frac{\hat{e}_l}{p_l}$$

$$\text{et dans le cas d'un plan de sondage de taille fixe : } \tilde{V}(\hat{\bar{Z}}_{HT}) = -\frac{1}{2N^2}\sum_{k\in s}\sum_{\substack{l\in s\\ l\neq k}}\frac{\Delta_{kl}}{p_{kl}}\left(\frac{\hat{e}_k}{p_k}-\frac{\hat{e}_l}{p_l}\right)^2$$

*si on rajoute une hypothèse de comportement*

**Dans le cas de l'estimateur petits domaines**, la formule appliquée est un peu différente de la formule d'un estimateur post-stratifié, puisque l'on prend la structure régionale à laquelle on applique un **comportement national.**

L'estimateur s'écrit :

$$\hat{\bar{Y}}_{R1} = \sum_{h=1..H}\frac{N_{Rh}}{N_R}\hat{\bar{Y}}_h .$$

On peut toutefois **le réécrire sous la forme d'un estimateur post-stratifié**. En effet,

$$\hat{\bar{Y}}_{R1} = \sum_{h=1..H}\frac{N_{Rh}}{N_R}\frac{\sum_{k\in s_{HID}\cap strh}\frac{y_k}{p_k}}{\sum_{k\in s_{HID}\cap strh}\frac{1}{p_k}} = \sum_{h=1..H}\frac{N_h}{N}\frac{\sum_{k\in s_{HID}\cap strh}\frac{N}{N_h}\frac{N_{Rh}}{N_R}\frac{y_k}{p_k}}{\sum_{k\in s_{HID}\cap strh}\frac{1}{p_k}}$$

On peut donc réécrire :

$$\hat{\bar{Y}}_{R1} = \sum_{h=1..H}\frac{N_h}{N}\frac{\sum_{k\in s_{HID}\cap strh}\frac{z_k}{p_k}}{\sum_{k\in s_{HID}\cap strh}\frac{1}{p_k}}$$

avec $z_k = \frac{N}{N_h}\frac{N_{Rh}}{N_R}y_k$ si l'individu *k* appartient à la post-strate *h* de la région *R.*

**On retombe donc sur l'expression d'un estimateur post-stratifié :**

$$\hat{\bar{Y}}_{R1} = \sum_{h=1..H}\frac{N_h}{N}\hat{\bar{Z}}_h$$

où $\hat{\bar{Z}}_h$ est la moyenne observée sur la strate h de la variable *z.*

On peut dès lors calculer la variance de l'estimateur à partir des formules d'estimation de variance pour estimateur post statifié. Comme on se trouve dans le cas d'un plan de sondage de taille fixe, on préfèrera la seconde formule de la variance, soit,

$$V(\hat{\bar{Z}}_{HT}) = -\frac{1}{2N^2}\sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}}\Delta_{kl}\left(\frac{e_k}{p_k}-\frac{e_l}{p_l}\right)^2$$

où $e_k = z_k - \overline{Z}_h$ si l'individu appartient à la strate *h*, soit

$$e_k = \frac{N}{N_h}\frac{N_{Rh}}{N_R}\left( y_k - \frac{\sum_{k \in strh} z_k}{N_h} \right)$$

**Un estimateur de la variance sera obtenu à partir de la formule :**

$$\tilde{\tilde{V}}(\hat{\overline{Z}}_{HT}) = -\frac{1}{2N^2}\sum_{k \in s}\sum_{\substack{l \in s \\ l \neq k}} \frac{\Delta_{kl}}{p_{kl}}\left(\frac{\hat{e}_k}{p_k} - \frac{\hat{e}_l}{p_l}\right)^2$$

où $$\hat{e}_k = \frac{\hat{N}}{\hat{N}_h}\frac{\hat{N}_{Rh}}{\hat{N}_R}\left( y_k - \frac{\sum_{k \in s_{HID} \cap strh} \frac{y_k}{p_k}}{\sum_{k \in s_{HID} \cap strh} \frac{1}{p_k}} \right)$$

et où $\hat{N}_{Rh}$, $\hat{N}$, $\hat{N}_h$ et $\hat{N}_R$ sont estimés à partir de l'échantillon VQS.

**Cette première approximation bute toutefois sur le problème du calcul des $p_{kl}$. En effet, il est particulièrement difficile de calculer les probabilités conjointes d'appartenance à un échantillon dès que le plan de sondage est un peu compliqué**.

### 3. Une approximation de l'estimateur de la variance de l'estimateur 'petits domaines'

L'estimateur de la variance n'étant pas directement calculable, on va approcher sa valeur en utilisant une formule proposée par J.C. DEVILLE[20] pour les sondages à probabilités inégales de taille fixe. Celui-ci propose d'approximer la formule

$$\tilde{\tilde{V}}(\hat{\overline{Z}}_{HT}) = -\frac{1}{2N^2}\sum_{k \in s}\sum_{\substack{l \in s \\ l \neq k}} \frac{\Delta_{kl}}{p_{kl}}\left(\frac{y_k}{p_k} - \frac{y_l}{p_l}\right)^2$$

par la formule suivante :

$$\hat{V}_2(\overline{Z}_{HT}) = \frac{1}{N^2}\frac{1}{1 - \sum_{l \in s} a_l^2}\sum_{k \in s}(1 - p_k)\left(\frac{y_k}{p_k} - A\right)^2 \qquad (3)$$

où $$a_l = \frac{1 - p_l}{\sum_{k \in s}(1 - p_k)}$$

et $$A = \sum_{k \in s} a_k \frac{y_k}{p_k}$$

[20]note interne UMS

**Dans le cas de l'estimateur petits domaines, on obtient donc comme formule d'approximation de l'estimateur de la variance :**

$$\hat{V}_2(\hat{\bar{Z}}_{HT}) = \frac{1}{N^2}\frac{1}{1-\sum_{l\in s} a_l^2}\sum_{k\in s}(1-p_k)\left(\frac{\hat{e}_k}{p_k} - A\right)^2$$

où $a_l = \dfrac{1-p_l}{\sum_{k\in s}(1-p_k)}$

et $A = \sum_{k\in s} a_k \dfrac{\hat{e}_k}{p_k}$

et $\hat{e}_k = \dfrac{\hat{N}}{\hat{N}_h}\dfrac{\hat{N}_{Rh}}{\hat{N}_R}\left(y_k - \dfrac{\sum_{k\in s_{HID}\cap strh}\frac{y_k}{p_k}}{\sum_{k\in s_{HID}\cap strh}\frac{1}{p_k}}\right)$

Toutefois, de l'avis même de son auteur, cette formule ne prend pas en compte des effets de grappe de l'échantillon et n'est pas bien adaptée au tirage en deux temps de l'échantillon. Selon lui, ce calcul sous-estimerait la variance réelle de l'estimateur 'petits domaines'.

Il serait toutefois possible de calculer une seconde approximation de la variance, qui, elle, sur-estimerait la variance, en prenant comme unité dans l'échantillon la grappe. Ce travail n'a pas été fait pour l'instant.

## ***Une seconde approximation***

Pour être plus rigoureux dans le calcul de variance, il faudrait également tenir compte du fait que les effectifs des strates sont aléatoires puisque calculés à partir de l'échantillon VQS.

Si l'on adopte les notations suivantes :

$s_{VQS}$ représente l'échantillon VQS national

h', pour *h' variant de 1 à H,* indice les strates d'individus formées dans l'échantillon VQS à partir de leur degré de prévalence

$s_{VQS\cap strh'}$ représente les individus de l'échantillon VQS appartenant à la strate *h'*

$s_{HID}$ représente l'échantillon HID

*h* indice les *H* post-strates construites sur l'échantillon HID une fois déterminé le modèle de comportement.

$s_{HID\cap postrh}$ représente les individus de l'échantillon HID appartenant à la post-strate *h*

$s_{VQS\cap postrh}$ représente les individus de l'échantillon VQS appartenant à la post-strate *h* (construite après détermination du modèle du comportement).

En fait, en conditionnant par l'échantillon VQS, la variance de l'estimateur se décompose en deux termes :

$$V\left(\hat{\bar{Y}}_{R1}\right) = \underset{s_{VQS}}{E}\ \underset{s_{HID}/s_{VQS}}{V}\left(\hat{\bar{Y}}_{R1}\right) + \underset{s_{VQS}}{V}\ \underset{s_{HID}/s_{VQS}}{E}\left(\hat{\bar{Y}}_{R1}\right)$$

Une approximation du second terme a déjà été calculée en partie plus haut (en tous les cas, une borne inf de l'approximation). Par contre, aucune approximation du premier terme n'a pour l'instant été calculée.

# *ANNEXE VIII-a*

## Premiers tests du modèle de comportement

*(stratification en 52 modalités : sexe*tranche d'âges*groupe VQS*type de logement)*

Signification des 10 variables HID sur lesquelles ont porté les tests:

- **aidki** : cette variable s'intéresse à la présence d'une aide régulière pour accomplir certaines tâches de la vie quotidienne en raison d'un handicap ou d'un problème de santé.
- **confin** : indique si en raison de problèmes de santé, handicap ou infirmités, la personne est confinée au lit, au fauteuil, ou à l'intérieur de son logement.
- **dadapt** : indique si la personne dispose de meubles ou d'équipements spécialement adaptés à son usage en raison de problèmes de santé, handicap ou infirmités.
- **ralloc** : si la personne interrogée perçoit au moment de l'interview une allocation, pension ou un autre revenu en raison de ses problèmes de santé.
- **rinval** : indique si on a reconnu à la personne un taux d'incapacité ou d'invalidité.
- **handi** : s'intéresse de savoir si la personne rencontre dans la vie de tous les jours des difficultés physiques, sensorielles, intellectuelles ou mentales.
- **mob** : signale si la personne est confinée au lit ou a besoin d'aide pour la toilette, l'habillage ou pour sortir.
- **defi** : indique si la personne souffre d'au moins une déficience.
- **cotor** : indique si la personne a déposé un dossier devant la COTOREP.
- **expr** : indique si la personne, âgée de plus de 6 ans, ne sait pas lire, écrire ou compter.

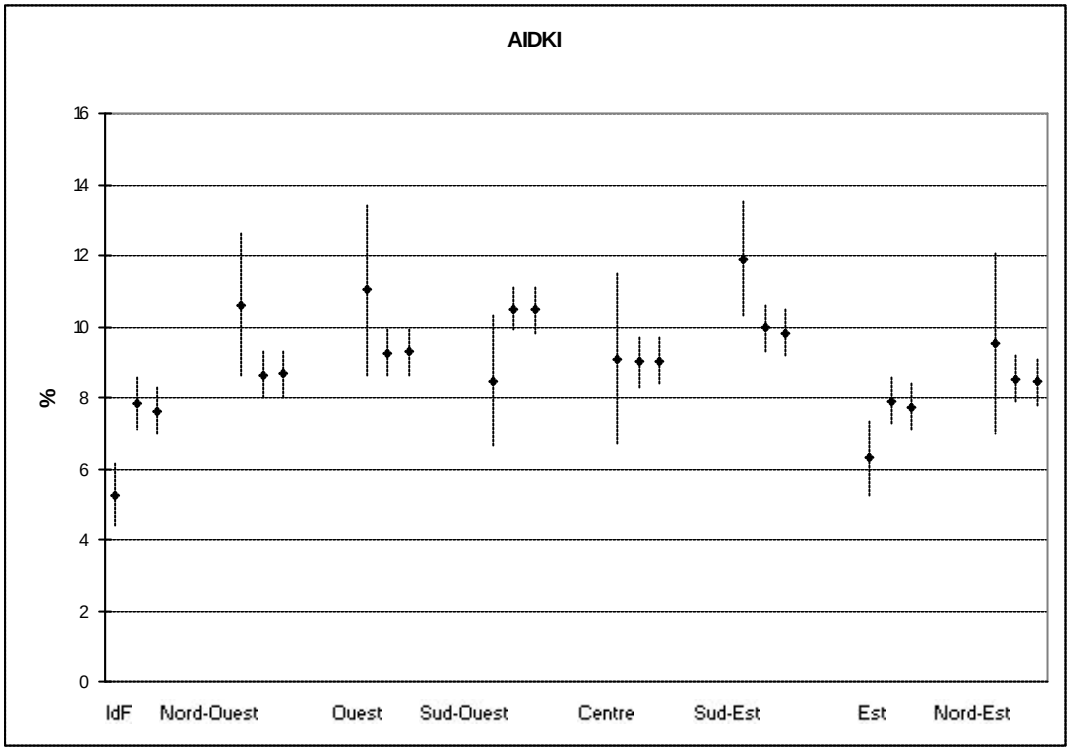
### I . Résultats sur des zones plus vastes que le départements

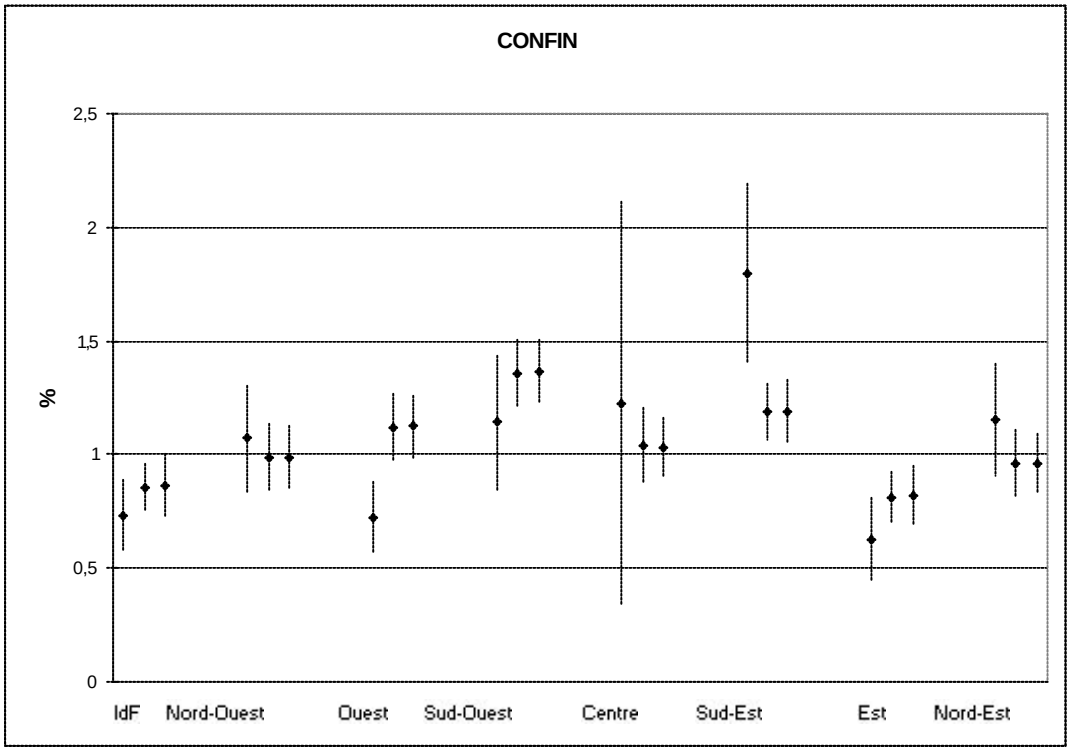
Pour chacune des 10 variables d'intérêt du fichier HID, choisies pour la diversité de leur taux de prévalence, on a représenté, sur chacune des 8 zones géographiques, l'intervalle de confiance à 95 % de trois estimateurs (de gauche à droite) :

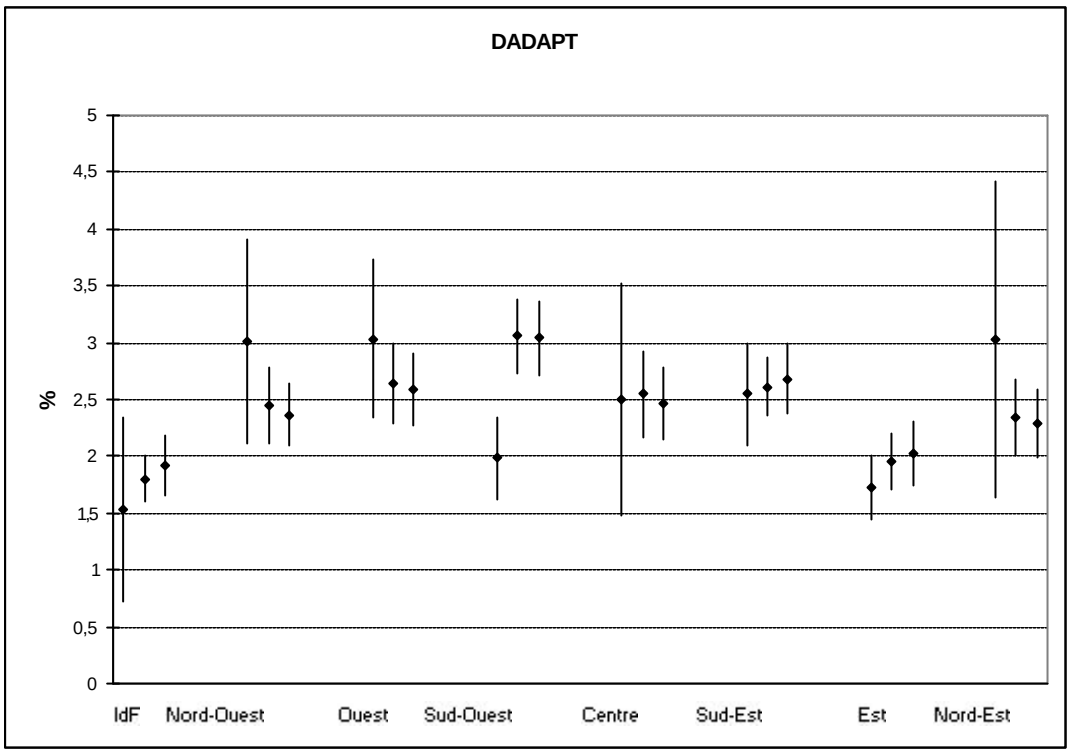
- l'estimateur direct ( basé sur les données locales)
- l'estimateur post-stratifié indirect (à partir de l'échantillon national) correspondant à :
  - 120 strates définies par : sexe*tranche d'âges*groupe VQS*type de logement
  - un regroupement de ces cases en 52 strates.

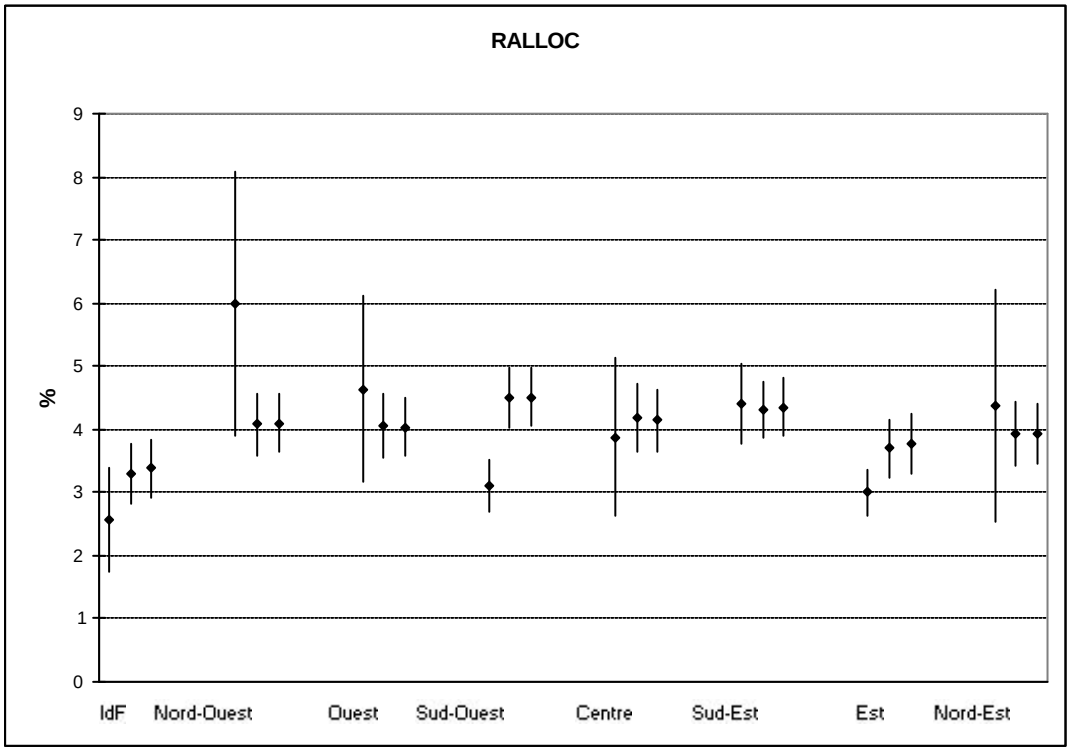
Les 8 zones géographiques sont les suivantes :

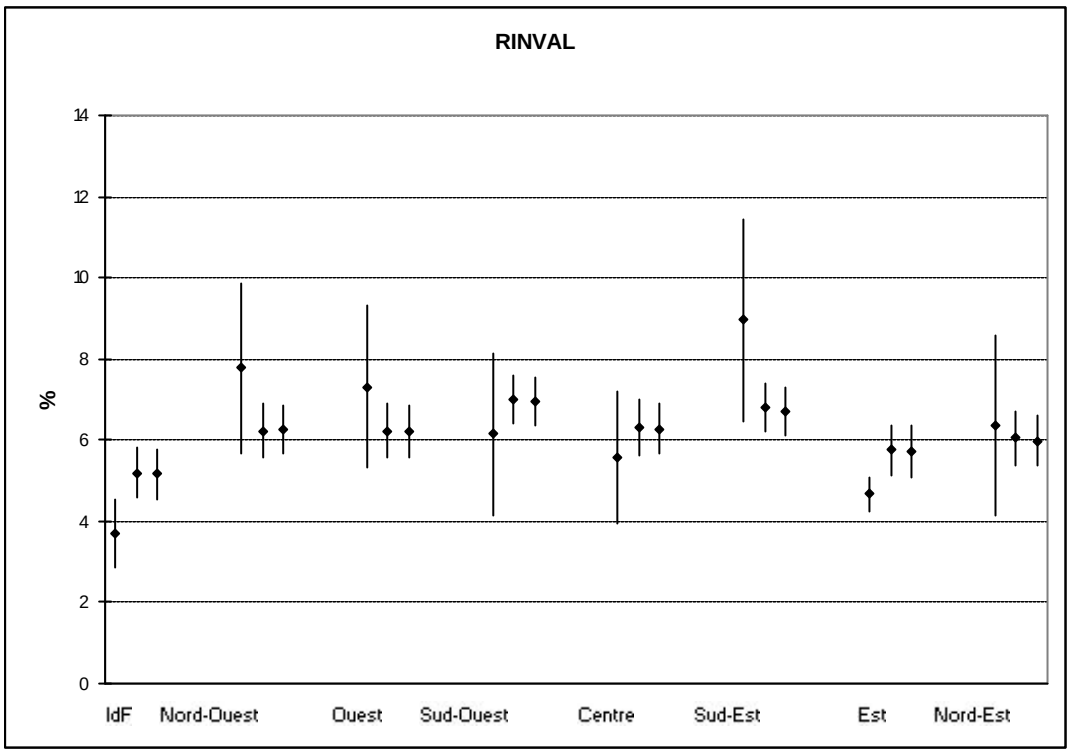
**Ile de France**, **Nord-Ouest** (Nord et Normandie), **Ouest** (Bretagne, Pays de la Loire et Poitou-Charentes), **Sud-Ouest** (Aquitaine et Midi-Pyrénées), **Centre** (Bourgogne, Centre, Limousin et Auvergne), **Sud-Est** (Languedoc, PACA et Corse), **Est** (Franche Comté et Rhône Alpes), et **Nord-Est** (Champagne Ardennes, Lorraine et Alsace).

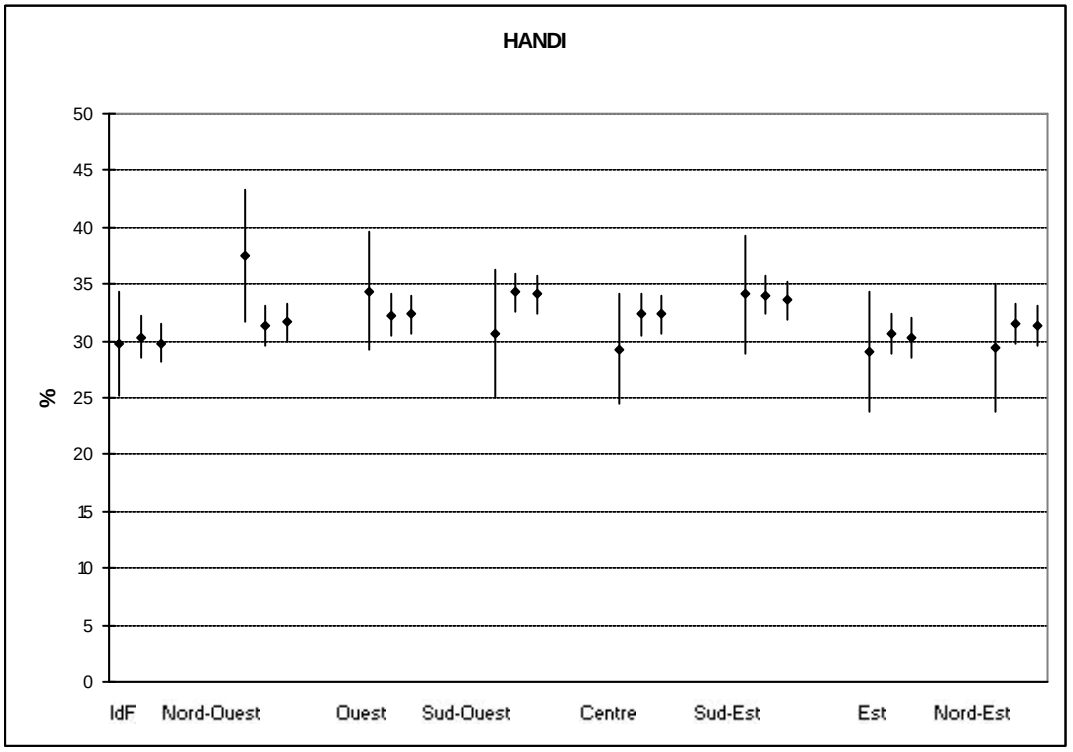

AIDKI
%
16
14
12
10
8
6
4
2
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est

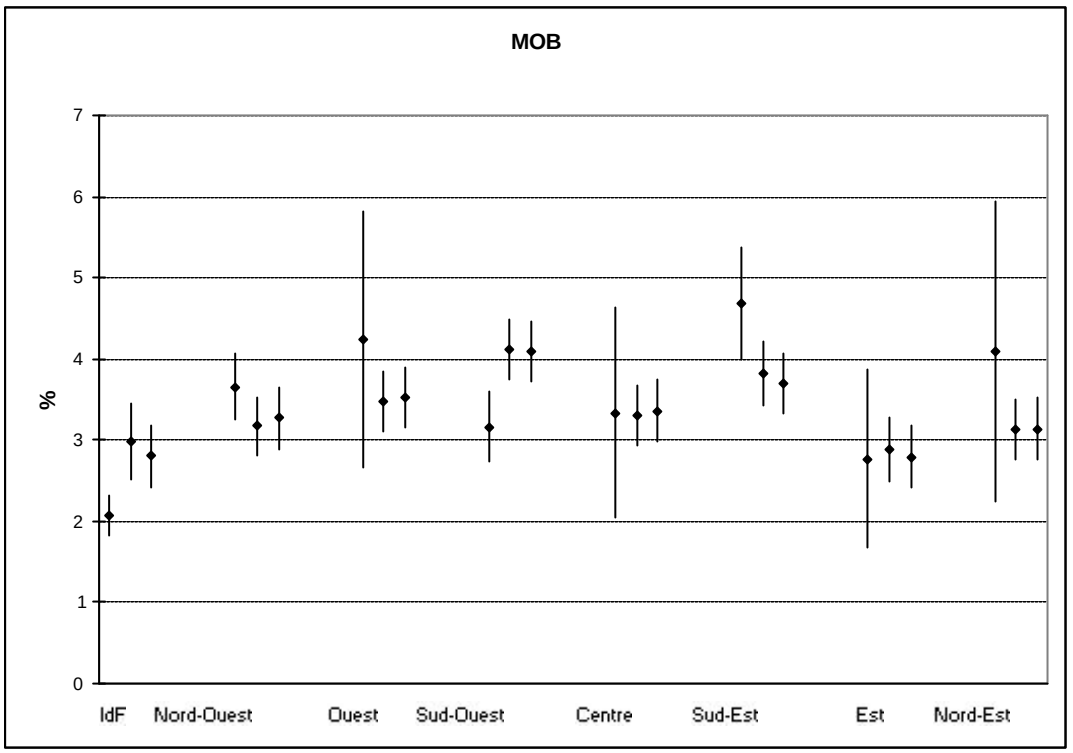

CONFIN
%
2,5
2
1,5
1
0,5
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est

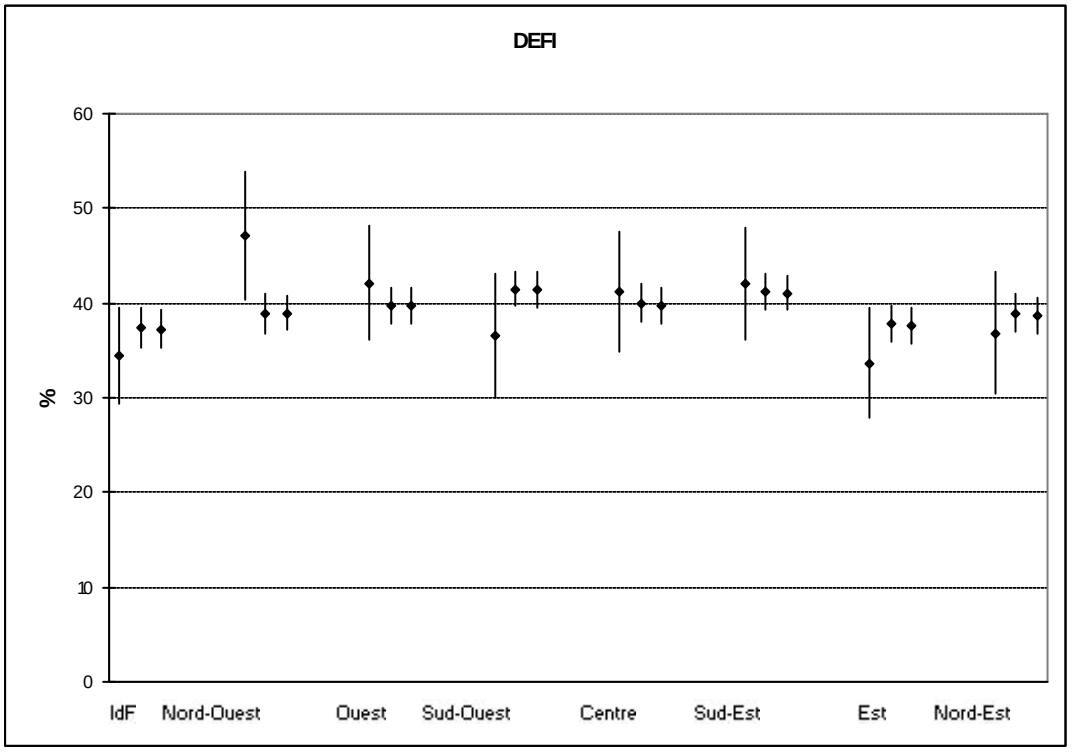

DADAPT
%
5
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est

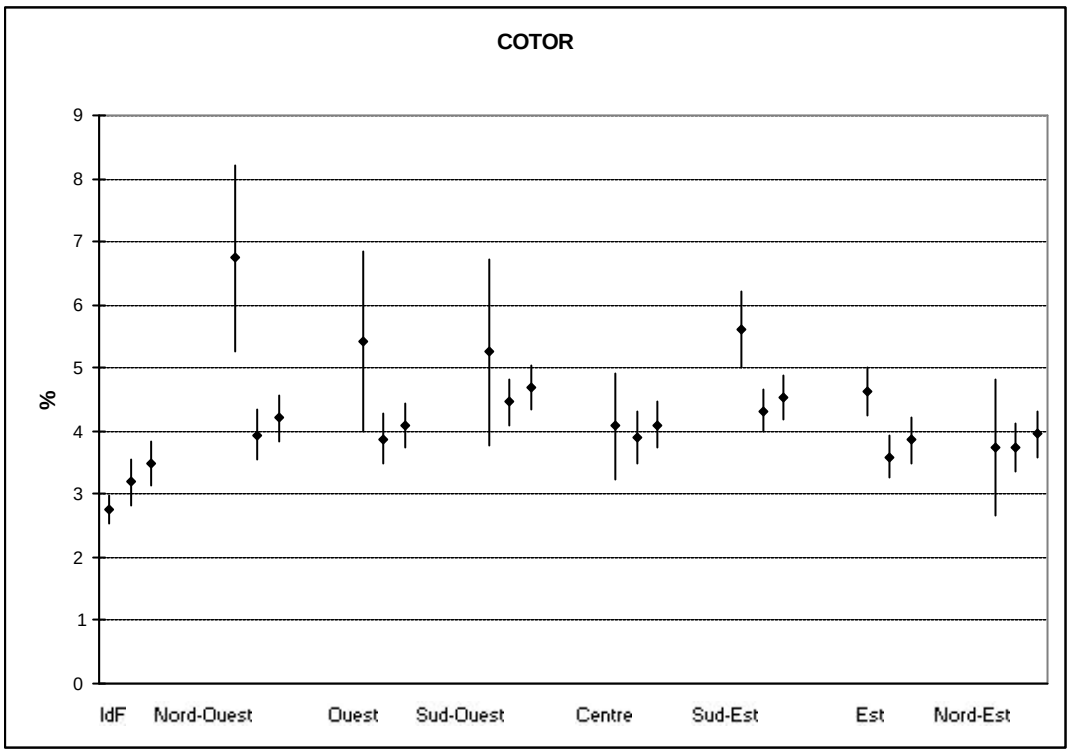

RALLOC
%
9
8
7
6
5
4
3
2
1
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est

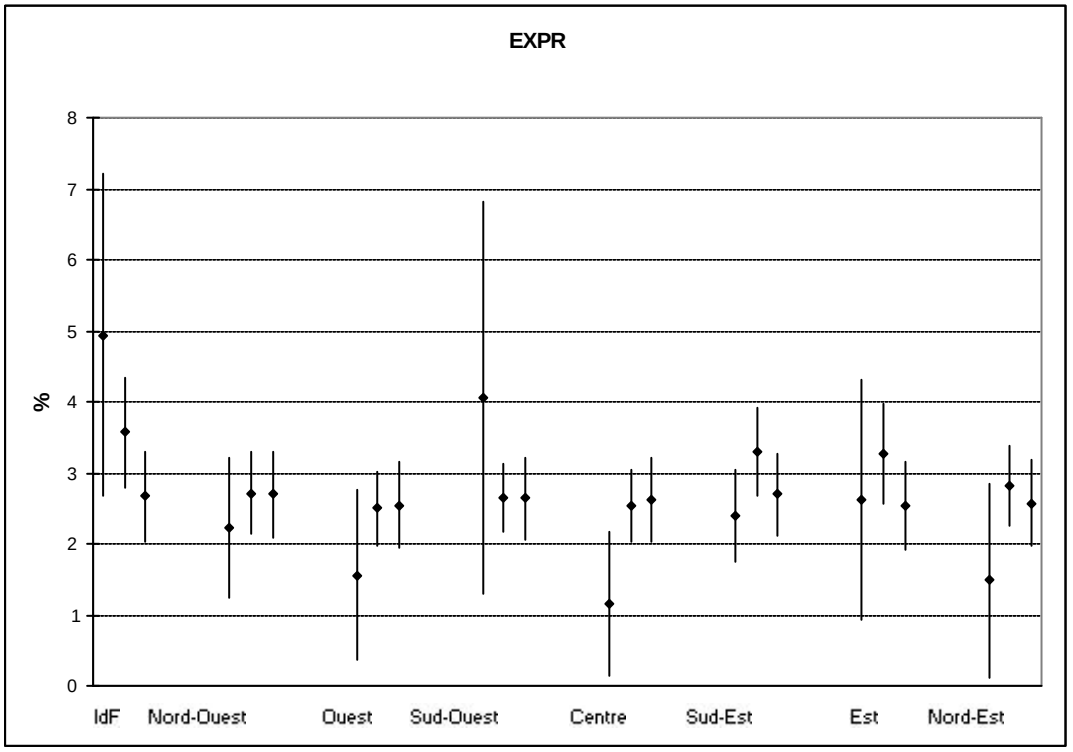

RINVAL
%
14
12
10
8
6
4
2
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est


HANDI
%
50
45
40
35
30
25
20
15
10
5
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est


MOB
%
7
6
5
4
3
2
1
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est


DEFI
%
60
50
40
30
20
10
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est

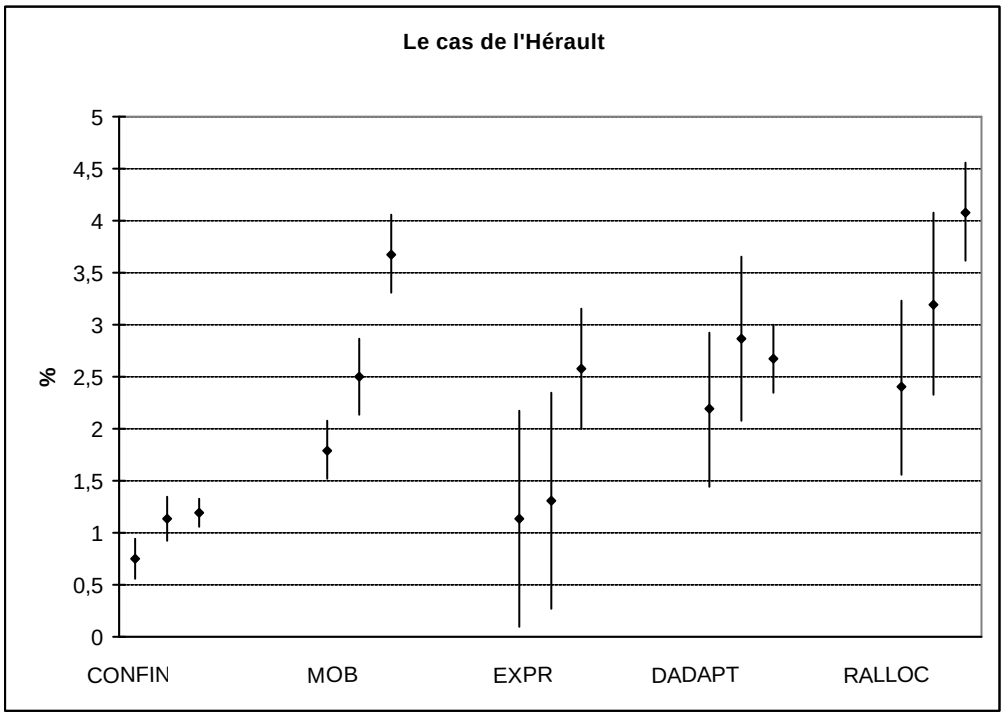

COTOR
%
9
8
7
6
5
4
3
2
1
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est

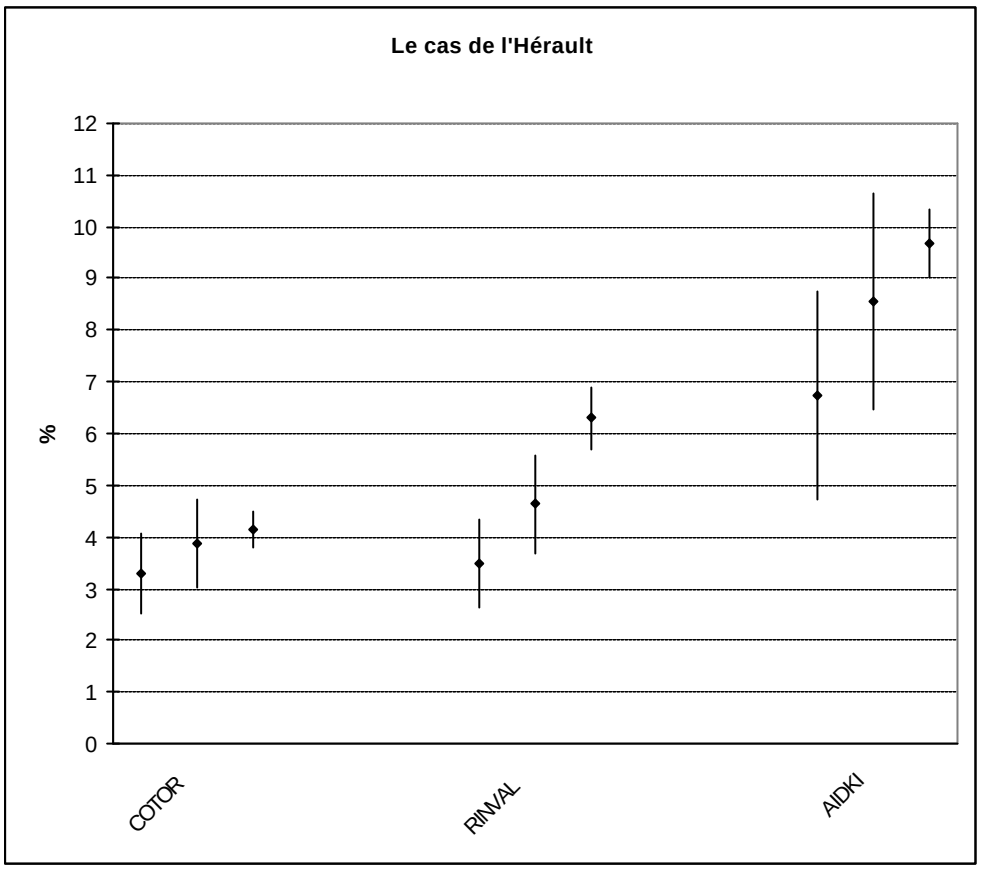

EXPR
%
8
7
6
5
4
3
2
1
0
IdF
Nord-Ouest
Ouest
Sud-Ouest
Centre
Sud-Est
Est
Nord-Est

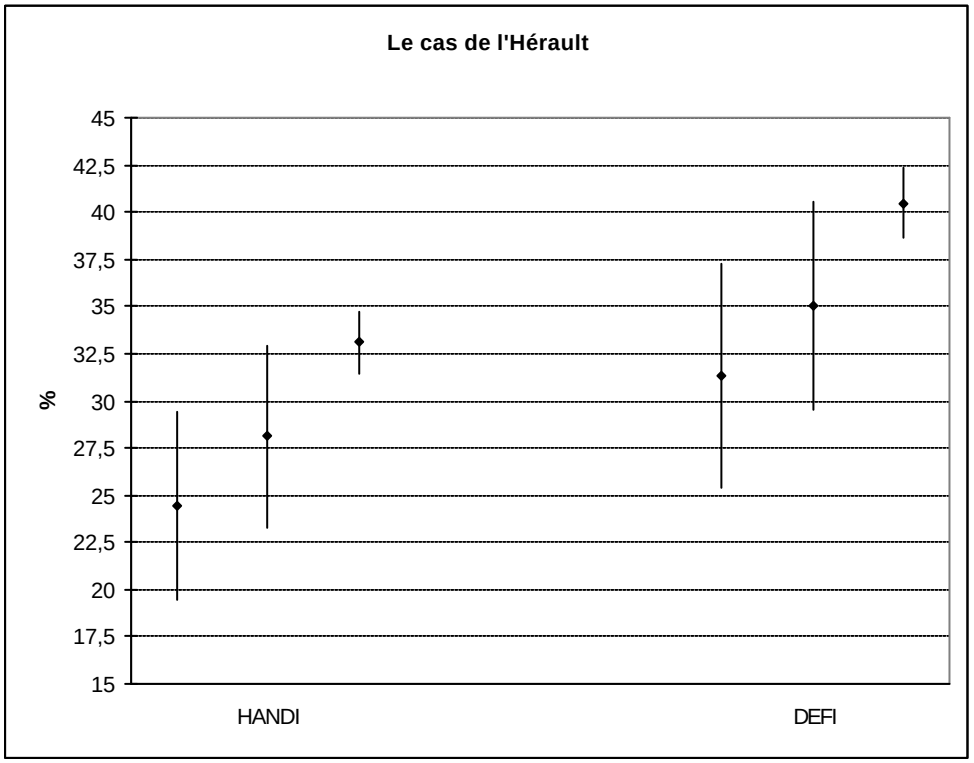
## II . Résultats dans le département de l'Hérault

Valeur estimée et intervalle de confiance à 95% de 3 estimateurs portant sur 10 variables HID dans le département de l'Hérault, avec de gauche à droite :

- l'estimateur direct calculé à partir des 1 475 réponses de l'échantillon HID,
- l'estimateur post-stratifié direct qui applique les comportements observés par strate dans l'échantillon HID de l'Hérault à la structure par strate observée sur des 16 172 réponses VQS de l'Hérault (la stratification est un regroupement en 52 cases des croisements sexe*tranche d'âges*groupe VQS*type de logement),
- l'estimateur post-stratifié indirect qui adapte les comportements nationaux observés dans ces mêmes strates à la structure VQS de l'Hérault.




Le cas de l'Hérault
%
12
11
10
9
8
7
6
5
4
3
2
1
0
COTOR
RINVAL
AIDKI


Le cas de l'Hérault
%
45
42,5
40
37,5
35
32,5
30
27,5
25
22,5
20
17,5
15
HANDI
DEFI

Différences dans la répartition par sexe et âge entre les fichiers HID et VQS et la source RP99

le tableau ci-dessous souligne ces différences :

**COMPARAISON DES MARGES PAR SEXE ET TRANCHE D'AGE ENTRE HID VQS ET RP DANS L'HERAULT**

| SEXE | TRANCHE D'AGE | HID | | | VQS | | | RP | |
|---|---|---|---|---|---|---|---|---|---|
| | | _FREQ_ | HMAR34 | % | _FREQ_ | VMAR34 | % | RMAR34 | % |
| ENSEMBLE | ENSEMBLE | 1475 | 875 097 | 100,00 | 16172 | 876 058 | 100,00 | 876 643 | 100,00 |
| ENSEMBLE | moins de 20 ans | 111 | 173 323 | 19,81 | 3609 | 195 450 | 22,31 | 214 642 | 24,48 |
| ENSEMBLE | 20-39 ans | 175 | 227 865 | 26,04 | 4677 | 254 515 | 29,05 | 248 070 | 28,30 |
| ENSEMBLE | 40-59 ans | 374 | 301 978 | 34,51 | 4191 | 223 836 | 25,55 | 221 744 | 25,29 |
| ENSEMBLE | 60-79 ans | 612 | 147 743 | 16,88 | 3021 | 165 065 | 18,84 | 161 249 | 18,39 |
| ENSEMBLE | 80 ans ou plus | 203 | 24 188 | 2,76 | 674 | 37 192 | 4,25 | 30 938 | 3,53 |
| 1 | ENSEMBLE | 682 | 400 610 | 45,78 | 7757 | 419 366 | 47,87 | 419 406 | 47,84 |
| 2 | ENSEMBLE | 793 | 474 487 | 54,22 | 8415 | 456 692 | 52,13 | 457 237 | 52,16 |
| 1 | moins de 20 ans | 52 | 88 500 | 10,11 | 1836 | 99 665 | 11,38 | 109 538 | 12,50 |
| 1 | 20-39 ans | 90 | 118 346 | 13,52 | 2314 | 125 088 | 14,28 | 119 356 | 13,62 |
| 1 | 40-59 ans | 196 | 110 496 | 12,63 | 2041 | 109 232 | 12,47 | 106 519 | 12,15 |
| 1 | 60-79 ans | 282 | 75 960 | 8,68 | 1342 | 73 008 | 8,33 | 72 891 | 8,31 |
| 1 | 80 ans ou plus | 62 | 7 309 | 0,84 | 224 | 12 373 | 1,41 | 11 102 | 1,27 |
| 2 | moins de 20 ans | 59 | 84 823 | 9,69 | 1773 | 95 786 | 10,93 | 105 104 | 11,99 |
| 2 | 20-39 ans | 85 | 109 519 | 12,52 | 2363 | 129 427 | 14,77 | 128 714 | 14,68 |
| 2 | 40-59 ans | 178 | 191 482 | 21,88 | 2150 | 114 604 | 13,08 | 115 225 | 13,14 |
| 2 | 60-79 ans | 330 | 71 783 | 8,20 | 1679 | 92 057 | 10,51 | 88 358 | 10,08 |
| 2 | 80 ans ou plus | 141 | 16 879 | 1,93 | 450 | 24 819 | 2,83 | 19 836 | 2,26 |

Remarques :

Le fichier VQS a été calé au niveau national sur les effectifs en ménage au RP99 répartis par région ( ou par département dans le cas où, comme dans l'Hérault, le département a fait l'objet d'une extension de l'enquête VQS). La relativement bonne représentation de la répartition de la population par sexe et âge dans l'Hérault tient à l'importance de l'échantillon VQS dans ce département (plus de 16 000 répondants).

L'échantillon HID national a, quant à lui, fait l'objet de plusieurs calages sur les effectifs au RP99 répartis notamment :
- par zones géographiques (les mêmes que VQS),
- par sexe et tranche d'âges, etc...

Toutefois les calages d'HID sur les marges du RP ne peuvent assurer une conformité des effectifs (et des structures) entre ces deux sources, aux croisements par exemple du sexe et de l'âge avec le département. La taille de l'échantillon départemental ( 1 475 répondants) ne garantit apparemment pas une bonne représentativité (on note plus particulièrement une forte sur-représentation des femmes de 40 à 60 ans dans l'échantillon HID). Cette remarque conduit à privilégier entre les résultats des deux estimateurs directs celui qui découle de la post-stratification VQS.

# *ANNEXE VIII-b*

## Test du modèle de comportement dans le département de l'Hérault

*(stratification en 52 modalités : sexe*tranche d'âges*groupe VQS*tranche d'unité urbaine)*

Le test porte sur dix variables de l'enquête HID. Pour chacune de ces variables choisies en raison de la diversité de leur taux de prévalence, l'estimateur post-stratifié de type petits domaines a été calculé dans l'Hérault – avec 52 post-strates - et les valeurs ont été comparées à celles de l'estimateur direct.

Rappel de la signification des 10 variables :

- **aidki** : cette variable s'intéresse à la présence d'une aide régulière pour accomplir certaines tâches de la vie quotidienne en raison d'un handicap ou d'un problème de santé.
- **confin** : indique si en raison de problèmes de santé, handicap ou infirmités, la personne est confinée au lit, au fauteuil, ou à l'intérieur de son logement.
- **dadapt** : indique si la personne dispose de meubles ou d'équipements spécialement adaptés à son usage en raison de problèmes de santé, handicap ou infirmités.
- **ralloc** : si la personne interrogée perçoit au moment de l'interview une allocation, pension ou un autre revenu en raison de ses problèmes de santé.
- **rinval** : indique si on a reconnu à la personne un taux d'incapacité ou d'invalidité.
- **handi** : s'intéresse de savoir si la personne rencontre dans la vie de tous les jours des difficultés physiques, sensorielles, intellectuelles ou mentales.
- **mob** : signale si la personne est confinée au lit ou a besoin d'aide pour la toilette, l'habillage ou pour sortir.
- **defi** : indique si la personne souffre d'au moins une déficience.
- **cotor** : indique si la personne a déposé un dossier devant la COTOREP.
- **expr** : indique si la personne, âgée de plus de 6 ans, ne sait pas lire, écrire ou compter.

Plus que d'apprécier la proximité des valeurs estimées directement et indirectement, il s'agit surtout de voir si les intervalles de confiance (à 95 %) définis autour de ces valeurs se chevauchent ou pas. Le rappel de la position de l'estimation nationale donne une idée de la distance parcourue lorsqu'on applique aux comportements nationaux une structure locale.

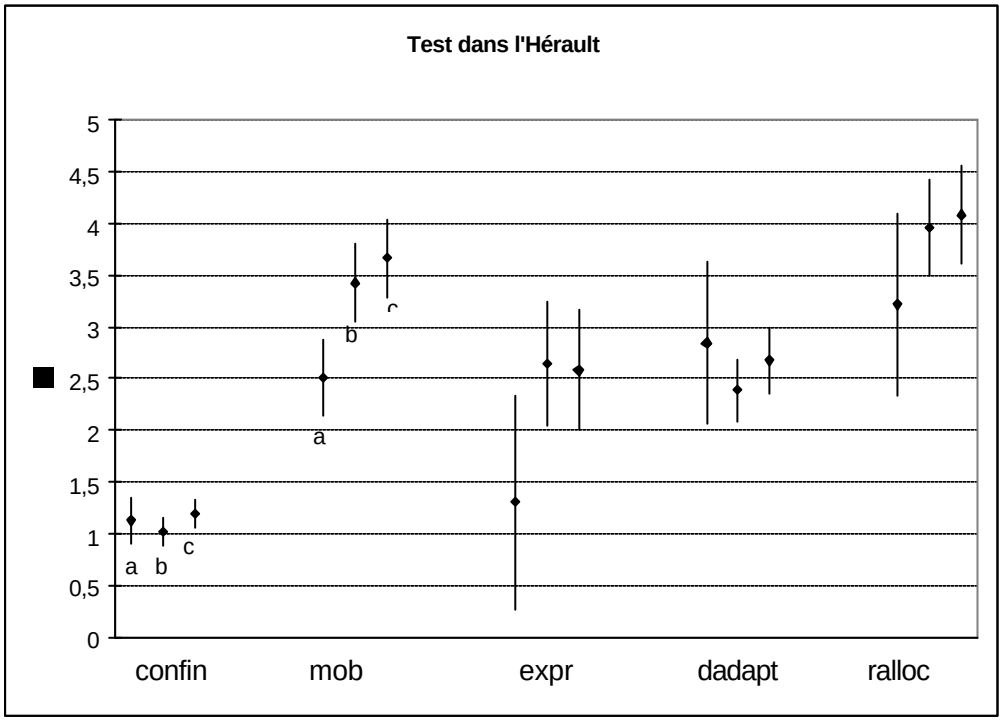
**Valeur estimée et intervalle de confiance à 95%**
**de 3 estimateurs portant sur 10 variables HID**
**dans le département de l'Hérault**

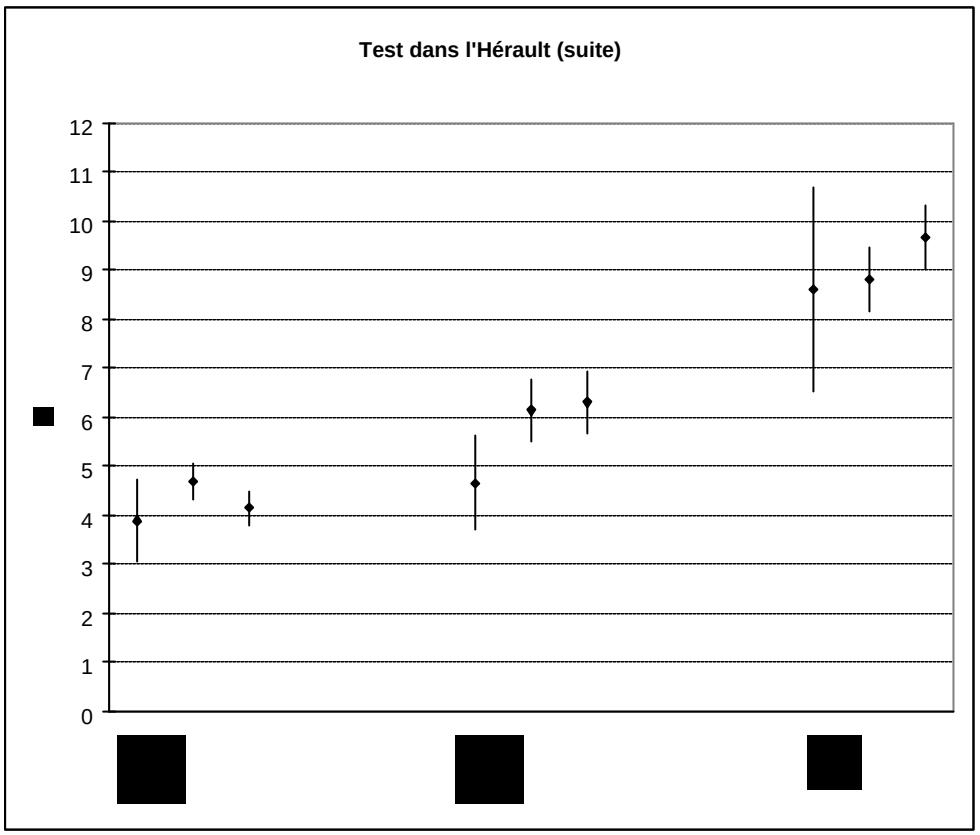
avec de gauche à droite :

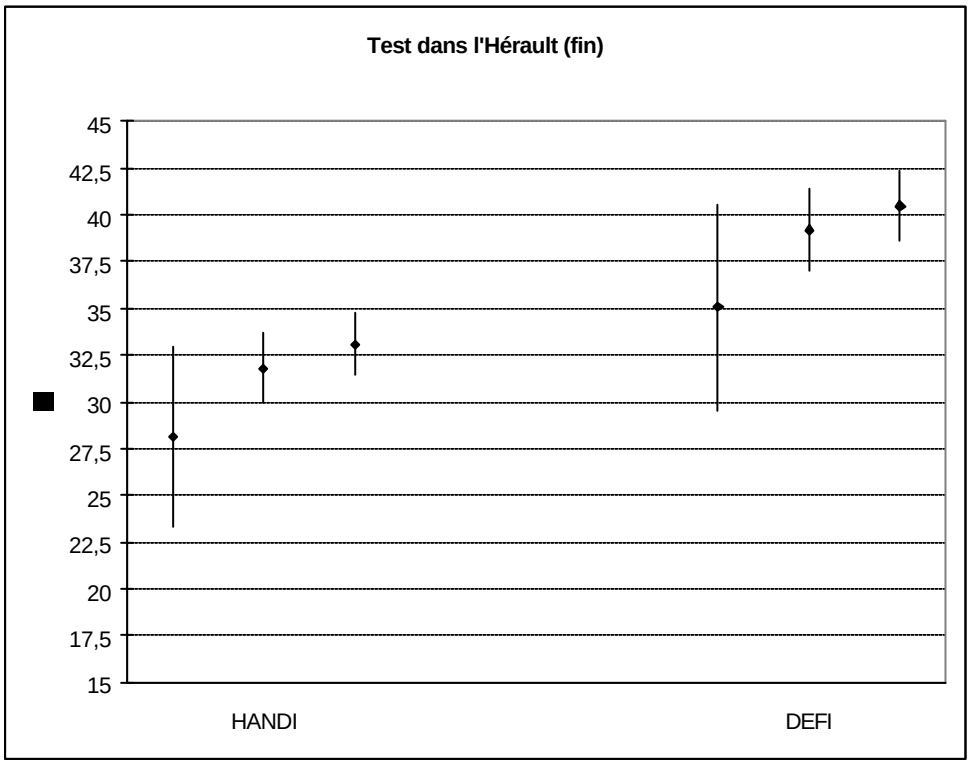
- (a) l'estimateur post-stratifié direct qui résulte de l'adaptation des comportements observés par strate dans l'échantillon des 1 479 répondants HID de l'Hérault à la structure par strate observée sur les 16 172 réponses VQS de l'Hérault (la stratification est établie en 52 cases, à partir des croisements sexe*tranche d'âges*groupe VQS*tranche d'unité urbaine),

- (b) l'estimateur national calculé à partir des 16 945 réponses de l'échantillon HID national,

- (c) l'estimateur post-stratifié indirect qui applique les comportements nationaux observés dans les 52 strates à la structure VQS de l'Hérault.




Test dans l'Hérault (suite)
12
11
10
9
8
7
6
5
4
3
2
1
0


Test dans l'Hérault (fin)
45
42,5
40
37,5
35
32,5
30
27,5
25
22,5
20
17,5
15
HANDI
DEFI

Les valeurs les plus à gauche (a) constituent notre cible :

- la valeur estimée directement est à deux exceptions près (dadapt et confin) la plus faible des trois,
- l'intervalle de confiance est le plus large car il est estimé à partir d'un échantillon départemental peu important.

La comparaison de a et c indique que dans l'Hérault, la prévalence du handicap est le plus souvent inférieure à la moyenne nationale.

Par ailleurs, la structure socio-démographique de l'Hérault tend à accroître la prévalence du handicap (confrontation de c et b). La prévalence du handicap en «c » est supérieure à celle de «b » pour 8 variables sur 10 (seules expr - illettrisme - et cotor - reconnaissance par la COTOREP - connaissent une situation inversée mais ces deux variables sont probablement les moins liées au facteur âge). Ceci vient probablement de ce qu'on compte dans la population vivant en domicile ordinaire de l'Hérault une proportion plus importante de personnes âgées que sur le territoire national comme le souligne le tableau suivant :

**COMPARAISON DES STRUCTURES PAR AGE DES MENAGES**
**DANS L'HERAULT (source VOS) ET SUR LE TERRITOIRE METROPOLITAIN (source HID)**

| TRANCHE D'AGE | VOS - HERAULT | | | HID - TERRITOIRE METROPOLITAIN | | |
|---|---|---|---|---|---|---|
| | _FREQ_ | POPVQS34 | % | _FREQ_ | POPHIDNAT | % |
| ENSEMBLE | 16172 | 876 058 | 100,00 | 16945 | 57 431 807 | 100,00 |
| moins de 20 ans | 3609 | 195 450 | 22,31 | 1553 | 14 770 369 | 25,72 |
| 20-39 ans | 4677 | 254 515 | 29,05 | 2262 | 16 136 700 | 28,10 |
| 40-59 ans | 4191 | 223 836 | 25,55 | 4371 | 14 939 177 | 26,01 |
| 60-79 ans | 3021 | 165 065 | 18,84 | 6782 | 9 797 994 | 17,06 |
| 80 ans ou plus | 674 | 37 192 | 4,25 | 1977 | 1 787 567 | 3,11 |

Ainsi, on arrive à la situation assez paradoxale qui est la suivante : dans 7 cas sur 10 l'estimation nationale (b) est plus proche de notre cible (a) que ne l'est l'estimation de type « petits domaines » (c). Cette dernière associe des taux de handicap relativement élevé (comportement national) à une population plutôt âgée (celle de l'Hérault). Par contre dans le cas de l'estimation nationale, ces mêmes prévalences nationales (beaucoup trop élevées) voient leurs effets atténués par une population concernée plus jeune et donc moins touchée par le handicap.

La persistance d'un effet local résiduel important conduit à rechercher d'autres critères. Une première piste consiste à mieux appréhender l'aspect social en corrigeant la structure sociale telle qu'elle ressort de HID par celle tirée du RP99.

# *ANNEXE IX*

## Calage sur les marges socio-démographiques du RP 99 dans l'Hérault

Jusqu'à présent, le rôle joué par le milieu social n'était pas directement pris en compte dans l'estimation indirecte. Aucune question relative au milieu social de l'individu n'étant posée dans l'enquête VQS, cette variable ne pouvait pas être croisée avec d'autres variables essentielles comme le groupe VQS (indiquant le degré de gravité du handicap). La prise en compte de la dimension sociale s'effectue donc dans un deuxième temps, par calage sur les marges du RP.
Cependant, la codification de la catégorie sociale n'étant pas encore disponible dans le RP99, on a été amené à construire une variable (socm), représentative du milieu social de la personne de référence du ménage, à l'aide du type d'activité, du statut, de la position professionnelle ou du niveau d'étude de cette personne de référence.

On voit ci-dessous que la répartition selon la variable socm, de la population nationale vivant en ménage d'après HID diffère quelque peu de celle observée au RP99 :

| VARIABLE CREE SOCM | REPARTITION DE LA POPULATION DES MENAGES SELON LE TYPE D'ACTIVITE, LE STATUT, LA POSITION PROFESSIONNELLE OU LE NIVEAU D'ETUDE DE LA PERSONNE DE REFERENCE DU MENAGE AU RP99 | | | | | | HID | | |
|---|---|---|---|---|---|---|---|---|---|
| | TACTM | | STATM | | POSPM | % | % | RP-HID | (RP-HID) *100/RP |
| | **11 = actifs ayant un emploi** | 36 970 904 | **10 = salarié** | 31 219 737 | | | | | |
| **01 - manoeuvre, OS** | | | | | **1B = manœuvre, OS** | 4 177 391 | **7,3** | 6,7 | 0,6 | 8,2 |
| **02 - OQ** | | | | | **1C = OQ ou très qualifié** | 7 916 496 | **13,8** | 14,2 | -0,4 | -2,6 |
| **03 - employés** | | | | | **1D+1E = employés** | 5 327 312 | **9,3** | 11,5 | **-2,2** | **-23,5** |
| **04 - techniciens** | | | | | **1F+1G+1H = technicien** | 5 457 064 | **9,5** | 9,3 | 0,2 | 2,5 |
| **05 - instit, catg B** | | | | | **1I = instit., infirmier, cat. B** | 1 998 176 | **3,5** | 3,4 | 0,1 | 2,6 |
| **06 - ingénieurs** | | | | | **1J = ingénieur** | 3 847 886 | **6,7** | 7,7 | **-1,0** | **-14,5** |
| **07 - catg A** | | | | | **1K = catg. A** | 2 495 412 | **4,4** | 2,0 | **2,4** | **54,1** |
| | | | | | **NIVETM** | | | | | |
| **08 - n/s prim+coll** | | | **21,22,23 =** | 5 751 167 | **primaire + collège** | 3 148 640 | **5,5** | 5,9 | -0,4 | -7,2 |
| **09 - n/s lycée+sup** | | | **non salarié** | | **lycée + supérieur** | 2 602 527 | **4,5** | 5,4 | **-0,9** | **-18,7** |
| **10 - anc actif prim** | **21 = ancien actif** | 13 064 931 | | 13 064 931 | **primaire** | 7 502 151 | **13,1** | 12,7 | 0,4 | 3,1 |
| **11 - anc actif coll** | | | | | **collège** | 3 162 206 | **5,5** | 5,7 | -0,2 | -3,2 |
| **12 - anc actif lyc+sup** | | | | | **lycée + supérieur** | 2 400 574 | **4,2** | 4,4 | -0,2 | -4,9 |
| **13 - aut inact prim** | **12,13,22,23 = chômeur + mil. du contingent + élève, étudiant + personnes âgées de - de 15 ans = autres inactifs** | 7 190 373 | | 7 190 373 | **primaire** | 2 083 872 | **3,6** | 3,0 | 0,6 | 17,6 |
| **14 - aut inact coll** | | | | | **collège** | 2 842 910 | **5,0** | 4,3 | **0,7** | **13,4** |
| **15 - aut inact lyc+sup** | | | | | **lycée + supérieur** | 2 263 591 | **4,0** | 3,8 | 0,2 | 3,9 |
| | **TOTAL** | **57 226 208** | | **57 226 208** | | **57 226 208** | **100,0** | **100,0** | 0,0 | 0,0 |

Les écarts importants constatés sur certains postes font penser qu'entre ces deux sources, les grilles de classement ne sont peut-être pas équivalentes et qu'il y a sans doute des améliorations à apporter.

Ce calage concerne le fichier national HID «repondéré » afin de l'adapter à la situation de l'Hérault (après correction, au sein de chaque strate définie par la nouvelle variable V52, des poids nationaux par le rapport de la population de la strate dans l'Hérault sur le national).
Ce calage s'effectue sur les marges à la fois démographiques (sexe et tranche d'âges) et sociales (la taille de la commune en 3 modalités et une approche du milieu social de la personne de référence du ménage en 15 postes).

Le calage pose problème car 39 individus HID ont un niveau d'étude du chef de ménage à valeur manquante. Deux solutions sont envisagées :
- soit faire un seul calage sur les quatre variables et on obtient en sortie la structure désirée, sans valeur manquante mais on perd 39 observations HID (ligne b du tableau suivant);
- soit faire deux calages successifs, le premier sur le milieu social et le second sur les trois autres variables : on obtient alors en sortie sur l'ensemble des observations HID, des pondérations qui reconstituent la population de l'Hérault selon une structure un peu différente des marges initiales (ligne c).

De toute façon les estimations de prévalence de handicap dans l'Hérault qui découlent de ces deux méthodes sont très proches et ne différent que de très peu des estimations indirectes établies en ligne a.

Les résultats sont regroupés dans les deux tableaux suivants :

**Effets du calge sur marges (sexe, tragev, tu3 et socm)**
**sur l'estimation de la prévalence du handicap dans l'Hérault.**

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL |
|---|---|---|---|---|---|
| estimateur post-stratifié indirect (a) | 1,19 | 9,66 | 2,68 | 4,08 | 6,29 |
| estimateur post-stratifié indirect après calage simple (b) | 1,13 | 9,59 | 2,70 | 4,96 | 7,14 |
| estimateur post-stratifié indirect après calage double © | 1,14 | 9,53 | 2,73 | 4,89 | 7,05 |
| estimateur post-stratifié direct (d) | 1,12 | 8,59 | 2,84 | 3,22 | 4,65 |
| estimateur post-stratifié direct après calage simple (e) | 1,07 | 8,18 | 2,55 | 3,85 | 5,44 |
| estimateur direct avec calage/marges VQS (f) | 1,15 | 8,81 | 2,92 | 3,18 | 4,69 |
| estimateur direct avec calage/marges VQS+RP (g) | 1,11 | 8,42 | 2,64 | 4,02 | 5,77 |
| estimateur post-stratifié indirect - direct (a-d) | 0,07 | 1,07 | -0,17 | 0,87 | 1,64 |
| le même écart après calage simple (b-e) | 0,06 | 1,41 | 0,15 | 1,11 | 1,70 |
| estimateur post-stratifié indirect calé - direct calé (b-f) | -0,03 | 0,78 | -0,22 | 1,78 | 2,45 |
| estimateur post-stratifié indirect calé - direct calé (b-g) | 0,01 | 1,16 | 0,07 | 0,94 | 1,37 |
| écart en % = (b-g)*100/g | 1,2 | 13,8 | 2,5 | 23,4 | 23,8 |

**Effets du calge sur marges sur la prévalence du handicap dans l'Hérault**
(suite)

| | HANDI | MOB | DEFI | COTOR | EXPR |
|---|---|---|---|---|---|
| estimateur post-stratifié indirect (a) | 33,09 | 3,66 | 40,50 | 4,14 | 2,58 |
| estimateur post-stratifié indirect après calage simple (b) | 33,62 | 3,73 | 41,22 | 5,74 | 2,32 |
| estimateur post-stratifié indirect après calage double © | 33,53 | 3,72 | 41,10 | 5,64 | 2,32 |
| estimateur post-stratifié direct (d) | 28,11 | 2,51 | 35,06 | 3,89 | 1,30 |
| estimateur post-stratifié direct après calage simple (e) | 27,78 | 2,47 | 36,56 | 5,68 | 1,97 |
| estimateur direct avec calage/marges VQS (f) | 26,94 | 2,71 | 33,90 | 4,38 | 1,29 |
| estimateur direct avec calage/marges VQS+RP (g) | 27,71 | 2,56 | 36,27 | 5,93 | 1,98 |
| estimateur post-stratifié indirect - direct (a-d) | 4,98 | 1,15 | 5,44 | 0,25 | 1,27 |
| le même écart après calage simple (b-e) | 5,83 | 1,26 | 4,66 | 0,05 | 0,36 |
| estimateur post-stratifié indirect calé - direct calé (b-f) | 6,68 | 1,02 | 7,32 | 1,35 | 1,03 |
| estimateur post-stratifié indirect calé - direct calé (b-g) | 5,91 | 1,17 | 4,95 | -0,19 | 0,34 |
| écart en % = (b-g)*100/g | 21,3 | 45,6 | 13,7 | -3,2 | 17,2 |

Du côté de l'estimateur direct, on a tenté :

- de caler sur les 4 marges du RP une version post-stratifiée, suivant la variable V52 de VQS (ligne e du tableau). Cependant, la post-stratification n'est pas satisfaisante dans la mesure où le fichier relatif à l'Hérault est assez restreint et où de nombreux croisements sont vides.
- C'est pourquoi, on a choisi de caler directement un fichier brut (ligne f) sur les 4 marges VQS (sexe, âge, taille de commune et groupe VQS),
- puis de recaler le résultat obtenu sur les 4 marges RP99 (sexe, âge, taille de commune et milieu social) (résultat en ligne g).

Pour apprécier les effets des calages, il faut comparer les écarts entre les deux estimateurs (direct et indirect) avant calage (a-d) et après calage (b-g). Selon la variable d'intérêt, le calage rapproche ou éloigne ces deux estimations mais, globalement on ne peut pas dire que son effet soit déterminant.

Il subsiste toujours un aspect local inexpliqué, dont la valeur relative varie fortement selon la variable d'intérêt.

# *ANNEXE X*

## Estimateur de types petits domaines sous un modèle à 1 ou 2 facteurs

(Laurent Wilms – UMS)

On donne, dans ce qui suit, une autre approche de la construction des estimateurs post-stratifié indirects utilisés dans l'enquête-ménages HID. Cette approche est basée sur une modélisation du comportement individuel au sein de chaque croisement d'un domaine (région ou département) et de l'une des 52 strates de handicap finalement retenues.

Une première application, au département de l'Hérault, d'un estimateur à un facteur, nous a permis de mieux comprendre son comportement. Nous proposons de l'améliorer en testant une nouvelle classe d'estimateurs à 2 facteurs tentant de corriger son biais régional.

**Notation :**

On note *k*, l'indice d'un individu, *U* le champ de l'enquête, à savoir les individus vivant en ménages (hors collectivité) en métropole.
*h*, l'indice se rapportant à l'une des H strates de handicap (*H*=52)
*d*, l'indice se rapportant au domaine géographique (région ou département)

On s'intéresse à l'estimation d'un total $t_d$ sur le domaine *d* en une variable *y :*

$$t_d = \sum_{k \in U \cap d} y_k$$

### I- Estimateur petits domaines sous un modèle à un facteur

Soit un individu k appartenant à la sous-population *(h,d)*, c'est à dire appartenant au groupe de handicap *h* et au domaine géographique *d*.

On pose comme modèle de comportement individuel :

$$\forall k \in (h,d) : y_k = a_h + e_k$$

où :
$a_h$ est une constante (ou facteur) de comportement individuel et ne dépend que du groupe de handicap de l'individu *k* considéré.

$e_k$ est l'erreur individuel de modélisation.

Une régression linéaire de type MCO opérée sur U, montre que les $e_k$ seront minimales (au sens des MCO) pour :

$a_h = \bar{y}_h$

avec $\bar{y}_h = \frac{1}{N_h} \sum_{k \in U \cap h} y_k$ , comportement moyen national dans le groupe *h*

Ainsi, on dispose d'un prédicteur du comportement de l'individu *k*, d'un groupe *h* donné :

$$\tilde{y}_k = \bar{y}_h$$

D'après Sarndäl, Swenson and Wretman (Model Assisted Survey Sampling), il devient alors possible de construire un prédicteur, puis un estimateur de $t_d$.

Le prédicteur est donné par :

$$\boxed{\tilde{t}_d = \sum_{k \in U \cap d} \tilde{y}_k}$$

qui peut se réécrire en :

$$\boxed{\tilde{t}_d = \sum_{k \in U \cap d} N_{hd} \bar{y}_h}$$

Les $N_{hd}$ et $\bar{y}_h$ étant des quantités inconnues, on les estime à partir de l'échantillon VQS du domaine *d* (pour les $N_{hd}$) et de l'échantillon HID national pour les $\bar{y}_h$. On retrouve alors exactement l'estimateur poststratifié indirect (noté $\hat{y}_{d1}$)

## II- Estimateur petits domaines sous un modèle à deux facteurs

Soit un individu *k* appartenant à la sous-population *(h,d)*, c'est à dire appartenant au groupe de handicap *h* et au domaine géographique *d*.

On pose comme modèle de comportement individuel :

$$\boxed{\forall k \in (h,d) : y_k = a_h + b_d + e_k}$$

On introduit donc un nouveau facteur $b_d$, où $b_d$ est une constante de comportement individuel qui ne dépend que du domaine (région ou département) d'appartenance de l'individu *k* considéré.

$e_k$ est l'erreur individuel de modélisation.

Une régression linéaire de type MCO opérée sur *U*, montre que les $e_k$ seront minimales (au sens des MCO) pour :

$$\boxed{\begin{aligned} a_h &= \bar{y}_h \\ b_d &= \bar{y}_d - \bar{y} \end{aligned}}$$

avec $\bar{y}_h = \frac{1}{N_h} \sum_{k \in U \cap h} y_k$ , comportement moyen national dans le groupe *h*

avec $\bar{y}_d = \frac{1}{N_d} \sum_{k \in U \cap d} y_k$ , comportement moyen national dans le domaine *d*

avec $\bar{y} = \frac{1}{N} \sum_{k \in U} y_k$ , comportement moyen national.

Ainsi, on dispose d'un prédicteur du comportement de l'individu *k*, d'un groupe *h* donné par:

$$\tilde{y}_k = \bar{y}_h + \bar{y}_d - \bar{y}$$

Le prédicteur est donc :

$$\tilde{t}_d = \sum_{k \in U \cap d} \tilde{y}_k$$

qui peut se réécrire, dans ce cas-ci :

$$\boxed{\tilde{t}_d = (\sum_{k \in U \cap d} N_{hd} \bar{y}_h) + N_d (\bar{y}_d - \bar{y})}$$

Les $N_{hd}$, $\bar{y}_h$, $\bar{y}_d$ et $\bar{y}$ (comportement national) étant des quantités inconnues, on les estime à partir de l'échantillon VQS du domaine *d* (pour les $N_{hd}$) et de l'échantillon HID national pour les $\bar{y}_h, \bar{y}_d$ et $\bar{y}$. On obtient alors un estimateur composé de $\hat{y}_{d1}$ et d'un terme correctif additif noté $(\hat{\bar{y}}_d - \hat{\bar{y}})$ sensé corriger le biais de $\hat{y}_{d1}$. Trois propositions de définition de ce terme correctif ont été formulées. On aboutit donc à 3 estimateurs à deux facteurs.

## *ANNEXE XI*

### Les nouvelles pondérations du fichier national correspondant à l'estimateur post-stratifié départemental à deux facteurs

Soit une variable d'intérêt Y, l'estimateur post-stratifié de type «petits domaines » à deux facteurs (effet de la strate + effet départemental) de la **proportion** $\bar{y}$ dans le département d s'écrit :

$$\hat{\bar{y}}_{d2} = \hat{\bar{y}}_{d1} + \left( \hat{\bar{y}}_{d} - \hat{\bar{y}} \right),$$

où $\hat{\bar{y}}_{d1}$ est l'estimateur post-stratifié indirect à un facteur que l'on connaît, qui dépend du comportement moyen national observé dans chacune des **52 strates h**, dont l'expression est :

$$\hat{\bar{y}}_{d1} = \sum_{h=1}^{52} \frac{\hat{N}_{h,d}}{\hat{N}_{d}} \quad \hat{\bar{y}}_{h} \qquad \text{avec} \qquad \hat{\bar{Y}}_{h} = \frac{1}{\hat{N}_{h}} \sum_{k \in h, HID} \frac{1}{p_{k}} \quad y_{k} \text{ tandis que}$$

$\frac{\hat{N}_{h,d}}{\hat{N}_{d}}$ correspond à la structure par strate dans le département d, tirée de l'enquête VQS.

Selon cette même combinaison de facteurs, l'estimation du **total** d'un handicap dans le département d devient :

$$\hat{y}_{d2} = \hat{N}_{d} \cdot \hat{\bar{y}}_{d2} = \hat{N}_{d} \cdot \hat{\bar{y}}_{d1} + \hat{N}_{d} \left( \hat{\bar{y}}_{d} - \hat{\bar{y}} \right),$$

et donc :

$$\hat{y}_{d2} = \sum_{h=1}^{52} \hat{N}_{h,d} \, \hat{\bar{y}}_{h} + \hat{N}_{d} \left( \hat{\bar{y}}_{d} - \hat{\bar{y}} \right),$$

$\hat{N}_{h,d}$ et $\hat{N}_d$ étant évalués d'après l'enquête VQS.

Cette prévalence peut être estimée, comme on va le voir, quel que soit le choix de la variable d'intérêt Y, à partir du **fichier national repondéré**. On se propose de déterminer l'expression nouvelle de ces poids correspondant aux trois termes des équations ci-dessus.

Précaution d'écriture : $\hat{N}$ est incliné quand la source est HID et il se redresse $\hat{N}$ quand l'information vient de VQS.

1. Développement du premier terme

$$\hat{Y}_{d1} = \hat{N}_d \cdot \hat{\bar{y}}_{d1} = \sum_{h=1}^{52} \hat{N}_{h,d} \quad \hat{\bar{y}}_h$$

$$= \sum_{h=1}^{52} \hat{N}_{h,d} \cdot \frac{1}{\hat{N}_h} \cdot \sum_{k \in h, HID} \frac{1}{p_k} y_k = \sum_{k \in HID} \frac{\hat{N}_{h,d}}{\hat{N}_h} \frac{1}{p_k} y_k$$

Appelons $w_k$ la nouvelle pondération du fichier national correpondant à ce premier terme :

$$wk = \frac{\hat{N}_{h,d}}{\hat{N}_h} \cdot \frac{1}{p_k} .$$

Le poids initial des individus dans chaque strate est multiplié par la population de la strate départementale (source VQS) rapportée à la population de la strate nationale (source HID). C'est le $w_k$ que l'on connaissait déjà dans le cadre d'un estimateur post-stratifié à un facteur.

2. Développement du deuxième terme

Ici, plusieurs variantes sont possibles :

***2.1.Effet départemental simple.***

$$\hat{N}_d \cdot \hat{\bar{y}}_d = \frac{\hat{N}_d}{\hat{N}_d} \cdot \sum_{k \in d, HID} \frac{1}{p_k} y_k = \sum_{k \in d, HID} \frac{\hat{N}_d}{\hat{N}_d \quad p_k} y_k .$$

Ainsi, la pondération correspondant au deuxième terme $\tilde{w}_k$ est égale :

- au poids de l'individu $\frac{1}{p_k}$ multiplié par le rapport des populations départementales (sources VQS / HID) si l'individu k appartient au département d,

- ou à zéro sinon.

Mais, on peut vouloir souhaiter que le comportement moyen dans le département d ne soit pas estimé directement à partir de l'échantillon HID départemental (trop restreint) mais après un redressement de celui-ci sur VQS (par post-stratification).

### *2.2. Effet départemenal redressé.*

La moyenne départementale se calculant sur un nombre assez restreint d'observations, il est donc préférable de la redresser sur peu de strates, par exemple sur **10 postes (s = 1, ..., 10)** correspond aux 6 groupes VQS croisés avec l'âge.
Sachant que la moyenne départementale redressée s'écrit,

$$\hat{\bar{y}}_{d_r} = \sum_{s=1}^{10} \frac{\hat{N}_{s,d}}{\hat{N}_d} \hat{\bar{y}}_{s,d}$$

(somme, pondérée par le poids relatif de la strate dans VQS, des comportements moyens par strate observés dans HID).

et que le comportement moyen dans la strate se définit comme suit,

$$\bar{y}_{s,d} = \frac{1}{\hat{N}_{s,d}} \sum_{k \in s,d} \frac{1}{p_k} y_k$$

alors, $\hat{N}_d . \hat{\bar{y}}_{d_r}$ devient :

$$\hat{N}_d . \hat{\bar{y}}_{d_r} = \sum_{k \in d, HID} \frac{\hat{N}_{s,d}}{\hat{N}_{s,d}} \cdot \frac{1}{p_k} y_k$$

$\tilde{w}_k$ est égal :

- au poids de l'individu $\frac{1}{p_k}$, corrigé du rapport des populations de la strate dans le département (sources VQS / HID) si l'individu k appartient au département d et à la strate s,

- ou à zéro sinon.

### *2.3. Effet départemental normalisé.*

On ne retient comme effet départemental que ce qui est dû au comportement ; ce qui conduit à caler les comportements locaux sur la structure nationale (pour tenir compte des critiques de Pierre Mormiche et de Lionel Qualité).

La moyenne départementale normalisée s'écrit alors :

$$\hat{\bar{y}}_{d_n} = \sum_{s=1}^{10} \frac{\hat{N}_s}{\hat{N}} \hat{\bar{y}}_{s,d}$$

et le deuxième terme devient :

$\hat{N}_d . \hat{\bar{y}}_{d_n} = \sum_{k \in d, HID} \frac{\hat{N}_d}{\hat{N}} \frac{\hat{N}_s}{\hat{N}_{s,d}} \cdot \frac{1}{p_k} y_k$ si k appartient au département d et à la strate s,

$\hat{N}_d . \hat{\bar{y}}_{d_n} = 0$ sinon.

3. Développement du troisième terme

$$-\hat{N}_d \cdot \hat{\bar{y}} = -\sum_{k \in HID} \frac{\hat{N}_d}{\hat{N}} \cdot \frac{1}{p_k} y_k$$

$\tilde{\tilde{w}}_k$ est égal au poids de l'individu $\frac{1}{p_k}$, multiplié par (- 1) et corrigé du rapport des populations départementales sur les effectifs nationaux (sources VQS / HID).

4. Expressions de la nouvelle pondération

$$\hat{Y}_{d2} = \sum_{k \in HID} \left( w_k + \tilde{w}_k + \tilde{\tilde{w}}_k \right) \; y_k$$

Soit en développant l'expression (et en tenant compte des trois variantes possibles du deuxième terme) :

$$\hat{Y}_{d21} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \; p_k} + \frac{\hat{N}_d}{\hat{N}_d \; p_k} \; 1_d - \frac{\hat{N}_d}{\hat{N} \; p_k} \right) y_k$$

$$\hat{Y}_{d22} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \; p_k} + \frac{\hat{N}_{s,d}}{\hat{N}_{s,d} \; p_k} \; 1_d - \frac{\hat{N}_d}{\hat{N} \; p_k} \right) y_k$$

$$\hat{Y}_{d23} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \; p_k} + \frac{\hat{N}_d \; \hat{N}_s}{\hat{N} \; \hat{N}_{s,d} \; p_k} \; 1_d - \frac{\hat{N}_d}{\hat{N} \; p_k} \right) y_k$$

où toutes les populations du numérateur se définissent dans VQS ( à l'exception de $\hat{N}_s$ dans le deuxième terme de $\hat{Y}_{d23}$ ) et celles du dénominateur dans HID.

On remarque que le deuxième terme est égal à zéro si l'individu k n'appartient pas au département d.

Remarque : il n'est pas exclu que certains individus dans l'échantillon national se retrouvent avec des poids négatifs.

# *ANNEXE XII*

## Résultats des quatre estimateurs de type « petits domaines »

## dans le département de l'Hérault

**I- Rappel des formules des différents estimateurs utilisés :**

Quatre estimateurs indirects ont été testés. Le premier n'envisage de particularité locale qu'à travers la structure socio-démographique du département, le comportement dans chacune des 52 strates définies auparavant, étant considéré comme géographiquement universel.

Estimateur post-stratifié indirect, à un seul facteur, sans effet départemental :

$$\hat{Y}_{d1} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \quad p_k} \right) \quad y_k$$

Règle d'écriture : $\hat{N}$ est incliné quand il s'agit de la source HID et il se redresse $\hat{\mathrm{N}}$ quand l'information vient de VQS.

Les deux estimateurs suivants tiennent compte, en plus du rôle joué par les structures locales, d'un effet résiduel départemental mesuré imparfaitement par l'écart entre les comportements moyens nationaux et locaux (moyenne simple ou redressée).

Estimateur post-stratifié indirect avec effet départemental simple :

$$\hat{Y}_{d21} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \quad p_k} + \frac{\hat{N}_d}{\hat{N}_d \quad p_k} \quad 1_d - \frac{\hat{N}_d}{\hat{N} \quad p_k} \right) \quad y_k$$

Estimateur post-stratifié indirect avec effet départemental redressé :

$$\hat{Y}_{d22} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \quad p_k} + \frac{\hat{N}_{s,d}}{\hat{N}_{s,d} \quad p_k} \quad 1_d - \frac{\hat{N}_d}{\hat{N} \quad p_k} \right) \quad y_k$$

Cependant, l'écart $\left( \hat{\bar{y}}_d - \hat{\bar{y}} \right)$ prend en compte non seulement des différences de comportement mais aussi des différences de structure. Or ce facteur ne devrait apprécier que des écarts de comportement. C'est pourquoi le dernier estimateur élimine ce qui pourrait relever d'une différence de structure de population en calant le comportement local sur la structure nationale.

Estimateur post-stratifié indirect avec effet départemental normalisé :

$$\hat{Y}_{d23} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \quad p_k} + \frac{\hat{N}_d \quad \hat{N}_s}{\hat{N} \quad \hat{N}_{s,d} \quad p_k} 1_d - \frac{\hat{N}_d}{\hat{N} \quad p_k} \right) y_k$$

## II- Les résultats :

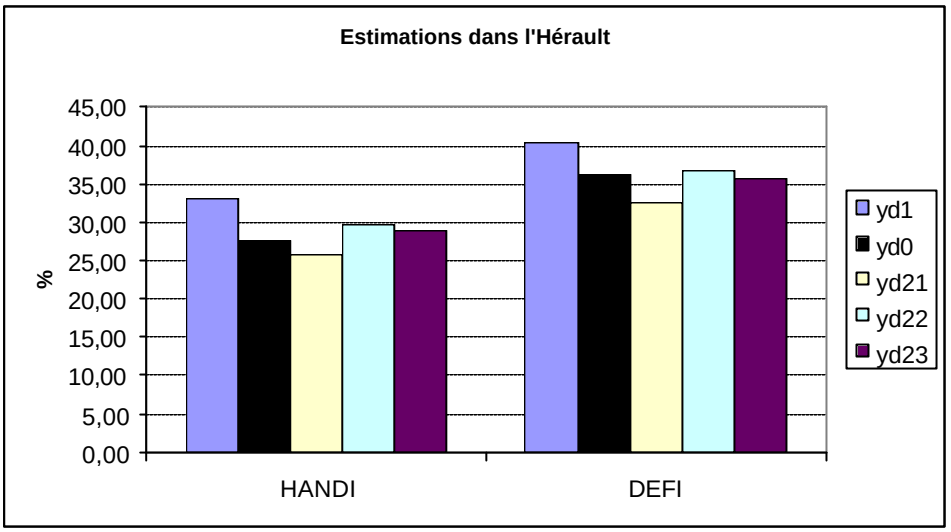
Les quatre estimations locales réalisées dans l'Hérault sont bien sûr rapprochées de l'estimation faite directement à partir du fichier départemental de l'Hérault. Plusieurs estimateurs directs sont proposés dans le tableau suivant. Leurs valeurs n'étant pas très éloignées les unes des autres, on se fixera comme cible l'estimateur direct avec calage sur les structures socio-démographiques du département observées d'abord dans VQS puis au RP99 (noté $\hat{Y}_{d0}$ ).

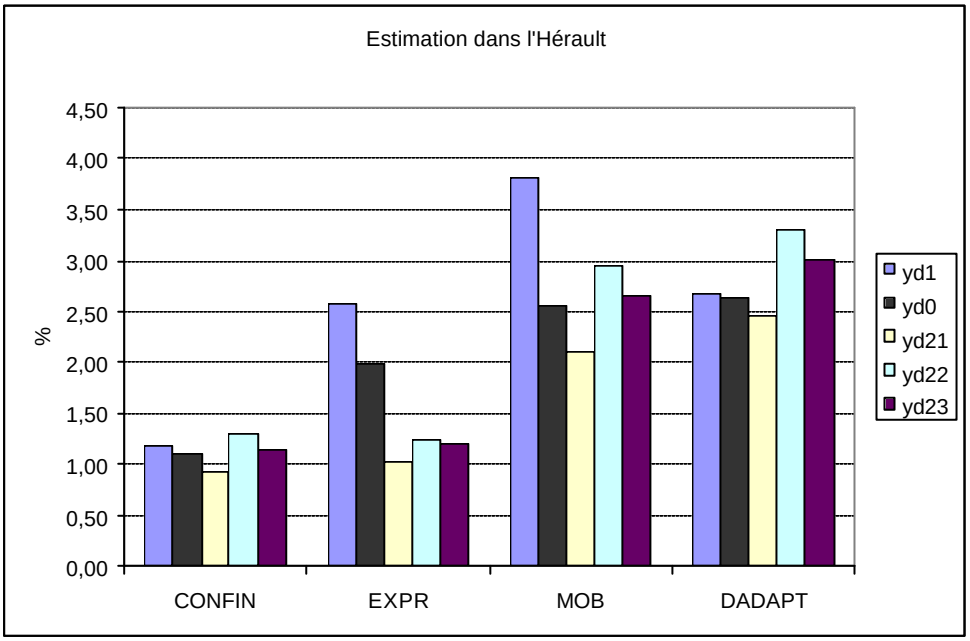
Les résultats sont portés dans les deux tableaux ci-joints :

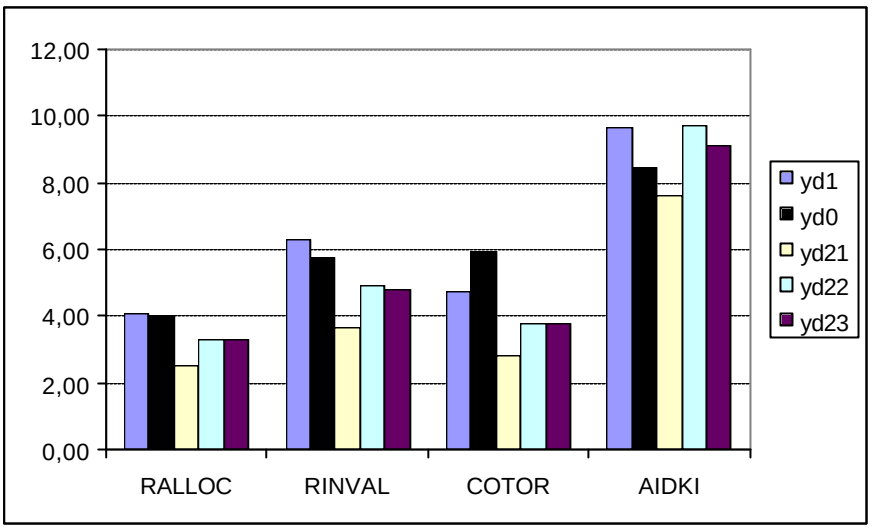
| **ESTIMATEURS DANS L'HERAULT** | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL |
|---|---|---|---|---|---|
| estimateur post-stratifié direct | 1,12 | 8,59 | 2,84 | 3,22 | 4,65 |
| estimateur post-stratifié direct après calage simple | 1,07 | 8,18 | 2,55 | 3,85 | 5,44 |
| estimateur direct avec calage/marges VQS | 1,15 | 8,81 | 2,92 | 3,18 | 4,69 |
| **estimateur direct avec calage/marges VQS+RP (yd0)** | **1,11** | **8,42** | **2,64** | **4,02** | **5,77** |
| estimateur post-stratifié indirect sans effet départemental **(yd1)** | 1,19 | 9,66 | 2,68 | 4,08 | 6,29 |
| estimateur post-stratifié indirect combiné avec effet départemental simple **(yd21)** | 0,93 | 7,60 | 2,47 | 2,53 | 3,64 |
| estimateur post-stratifié indirect combiné avec effet départemental redressé **(yd22)** | 1,30 | 9,72 | 3,31 | 3,30 | 4,93 |
| estimateur post-stratifié indirect combiné avec effet départemental normalisé **(yd23)** | 1,14 | 9,12 | 3,01 | 3,29 | 4,82 |

| **ESTIMATEURS DANS L'HERAULT** | HANDI | MOB | DEFI | COTOR | EXPR |
|---|---|---|---|---|---|
| estimateur post-stratifié direct | 28,11 | 2,51 | 35,06 | 3,89 | 1,30 |
| estimateur post-stratifié direct après calage simple | 27,78 | 2,47 | 36,56 | 5,68 | 1,97 |
| estimateur direct avec calage/marges VQS | 26,94 | 2,71 | 33,90 | 4,38 | 1,29 |
| **estimateur direct avec calage/marges VQS+RP (yd0)** | **27,71** | **2,56** | **36,27** | **5,93** | **1,98** |
| estimateur post-stratifié indirect sans effet départemental **(yd1)** | 33,09 | 3,82 | 40,50 | 4,73 | 2,57 |
| estimateur post-stratifié indirect combiné avec effet départemental simple **(yd21)** | 25,77 | 2,11 | 32,65 | 2,84 | 1,03 |
| estimateur post-stratifié indirect combiné avec effet départemental redressé **(yd22)** | 29,75 | 2,95 | 36,67 | 3,80 | 1,23 |
| estimateur post-stratifié indirect combiné avec effet départemental normalisé **(yd23)** | 28,84 | 2,66 | 35,70 | 3,77 | 1,20 |

mais les graphiques rendent les comparaisons plus parlantes :






On remarque :

- Comme on le savait déjà, yd1 est, à une exception près (cotor), plus élevé que yd0. Globalement sur l'ensemble des strates, la prévalence du handicap est en moyenne plus faible dans l'Hérault que sur le territoire métropolitain.

- La correction apportée par yd21, [$\hat{\bar{y}}_{d21} = \hat{\bar{y}}_{d1} + \left(\hat{\bar{y}}_{d} - \hat{\bar{y}}\right)$], est trop forte. Pour les 10 variables étudiées la nouvelle estimation est toujours inférieure à l'évaluation directe yd0. Cette estimation pose problème car **la prévalence du handicap** $\left(\hat{\bar{y}}_{d}\right)$ **est sous évaluée dans l'échantillon HID de l'Hérault**, comme le montre la comparaison avec VQS des structures de population selon la variable strate (la strate est un croisement du groupe VQS avec l'âge, elle mesure un handicap croissant de 1 à 10).

/************ VQS HERAULT **********************  ************** HID HERAULT ************/

| STRATE | Freq | Percent | Cumulative Frequency | Cumulative Percent | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|---|---|---|
| 1 | 638481 | 72.9 | 638481 | 72.9 | 704630.9 | 80.5 | 704630.9 | 80.5 |
| 2 | 37562.31 | 4.3 | 676043.3 | 77.2 | 26900.78 | 3.1 | 731531.7 | 83.6 |
| 3 | 45966.69 | 5.2 | 722010 | 82.4 | 38382.19 | 4.4 | 769913.8 | 88.0 |
| 4 | 13333.21 | 1.5 | 735343.2 | 83.9 | 9117.782 | 1.0 | 779031.6 | 89.0 |
| 5 | 39008.44 | 4.5 | 774351.6 | 88.4 | 26735.88 | 3.1 | 805767.5 | 92.1 |
| 6 | 27719.78 | 3.2 | 802071.4 | 91.6 | 18305.3 | 2.1 | 824072.8 | 94.2 |
| 7 | 16870.84 | 1.9 | 818942.2 | 93.5 | 11453.85 | 1.3 | 835526.7 | 95.5 |
| 8 | 21327.94 | 2.4 | 840270.2 | 95.9 | 14802.92 | 1.7 | 850329.6 | 97.2 |
| 9 | 25661.51 | 2.9 | 865931.7 | 98.8 | 19017.47 | 2.2 | 869347.1 | 99.3 |
| 10 | 10126.33 | 1.2 | 876058 | 100.0 | 5749.694 | 0.7 | 875096.8 | 100.0 |

- Ainsi, l'utilisation pour le calcul de la prévalence du handicap dans l'Hérault $\left(\hat{\bar{y}}d\right)$ d'un estimateur post-stratifié sur les 10 catégories de la variable strate dans VQS, conduit en toute logique à rehausser l'estimation indirecte : c'est ce qui se produit avec yd22, dont l'évaluation repasse au dessus de yd0 dans 6 cas sur dix (confin, aidki, dadapt, handi, mob et defi). Il est curieux de constater que les trois variables - ralloc, rinval et cotor - liées à la reconnaissance du handicap se comportent différemment ainsi que la variable expr qui est peut-être de nature différente (elle détecte des problèmes pouvant avoir d'autres causes que le handicap).

- Mais on peut s'interroger sur la nature de cet estimateur combiné en remarquant, après Pierre Mormiche et Lionel Qualité, que l'écart $\left(\hat{\bar{y}}d - \hat{\bar{y}}\right)$ prend en compte non seulement des différences de comportement mais aussi des différences de structure. Le dernier estimateur yd23 répond à cette critique puisqu'il recadre les comportements moyens du département sur une structure nationale (toujours en 10 catégories, eu égard à la faiblesse de l'échantillon local). Cet estimateur améliore légèrement la situation des variables classiques de handicap (du même type que confin, aidki, dadapt, handi, mob et defi) et n'a pratiquement pas d'effet sur les variables plus institutionnelles (ralloc, rinval et cotor).

- Les estimateurs combinés génèrent des poids négatifs. Les nouvelles pondérations sont à somme presque nulle sur les départements autres que l'Hérault, ce qui ne veut pas dire que les observations relevées dans ces départements n'ont aucune contribution sur le total du handicap estimé dans l'Hérault (mais cette contribution est fortement réduite). Les tableaux suivants illustrent cette remarque : il s'agit des résultats d'une proc summary et d'une proc univariate sur les pondérations nouvelles associées à yd21 ( yd22 et yd23 donnent des résultats approchants).

Contribution des départements enquêtés à l'estimation du handicap dans l'Hérault

| Provenance | _TYPE_ | _FREQ_ | CONFIN | AIDKI | DADAPT | RALLOC |
|---|---|---|---|---|---|---|
| Ensemble | 0 | 16945 | 8118.14 | 66563.20 | 21627.67 | 22122.31 |
| Hérault | 1 | 1475 | 6588.52 | 59165.57 | 19164.05 | 21053.50 |
| Autres dep | 1 | 15470 | 1529.62 | 7397.63 | 2463.62 | 1068.81 |

| Provenance | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR |
|---|---|---|---|---|---|---|
| Ensemble | 31884.44 | 225765.79 | 18492.62 | 286036.17 | 24862.57 | 9033.11 |
| Hérault | 30512.93 | 214494.22 | 15174.98 | 274835.48 | 24275.28 | 9525.89 |
| Autres dep | 1371.51 | 11271.57 | 3317.64 | 11200.69 | 587.29 | -492.78 |

Eventail des nouvelles pondérations associées à l’estimateur yd21

---------- ENSEMBLE DU FICHIER National ----------

Univariate Procedure

Variable=W2134 (pondération associée à l’estimateur yd21)

Moments

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| N | 16945 | Sum Wgts | 16945 | 100% Max | 7591.071 | 99% | 835.4791 |
| Mean | 51.70009 | Sum | 876058 | 75% Q3 | 2.967577 | 95% | 98.4814 |
| Std Dev | 436.7109 | Variance | 190716.4 | 50% Med | 0.48042 | 90% | 13.22536 |
| Skewness | 10.63077 | Kurtosis | 117.6129 | 25% Q1 | -0.76146 | 10% | -3.66846 |
| USS | 3.2768E9 | CSS | 3.2315E9 | 0% Min | -52.6395 | 5% | -12.106 |
| CV | 844.7005 | Std Mean | 3.354851 | | | 1% | -33.9747 |
| T:Mean=0 | 15.41055 | Pr>\|T\| | 0.0001 | Range | 7643.71 | | |
| Num ^= 0 | 16945 | Num > 0 | 10451 | Q3-Q1 | 3.729035 | | |
| M(Sign) | 1978.5 | Pr>=\|M\| | 0.0001 | Mode | 103.5726 | | |
| Sgn Rank | 21419910 | Pr>=\|S\| | 0.0001 | | | | |

Quantiles(Def=5)

Extremes

| Lowest | Obs | Highest | Obs |
|---|---|---|---|
| -52.6395( | 222) | 6261.64( | 1301) |
| -47.2888( | 376) | 6586.929( | 1378) |
| -46.059( | 230) | 6644.138( | 1235) |
| -45.4194( | 599) | 7013.511( | 1381) |
| -44.2774( | 623) | 7591.071( | 539) |

---------- POIDS DES OBSERVATIONS DES DEPARTEMENTS AUTRES QUE L’HERAULT ----------

Univariate Procedure

Variable=W2134 (pondération associée à l’estimateur yd21)

Moments

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| N | 15470 | Sum Wgts | 15470 | 100% Max | 236.7086 | 99% | 15.70431 |
| Mean | -0.0038 | Sum | -58.7459 | 75% Q3 | 2.030826 | 95% | 6.701287 |
| Std Dev | 12.13263 | Variance | 147.2006 | 50% Med | 0.222471 | 90% | 4.625616 |
| Skewness | 7.459355 | Kurtosis | 112.3299 | 25% Q1 | -0.93759 | 10% | -4.18379 |
| USS | 2277047 | CSS | 2277047 | 0% Min | -52.6395 | 5% | -13.0265 |
| CV | -319498 | Std Mean | 0.097546 | | | 1% | -34.2686 |
| T:Mean=0 | -0.03893 | Pr>\|T\| | 0.9689 | Range | 289.3481 | | |
| Num ^= 0 | 15470 | Num > 0 | 8976 | Q3-Q1 | 2.968419 | | |
| M(Sign) | 1241 | Pr>=\|M\| | 0.0001 | Mode | 5.103114 | | |
| Sgn Rank | 9474873 | Pr>=\|S\| | 0.0001 | | | | |

Quantiles(Def=5)

Extremes

| Lowest | Obs | Highest | Obs |
|---|---|---|---|
| -52.6395( | 222) | 206.4122( | 1215) |
| -47.2888( | 376) | 208.6006( | 1188) |
| -46.059( | 230) | 218.5978( | 1201) |
| -45.4194( | 534) | 221.8207( | 1193) |
| -44.2774( | 558) | 236.7086( | 1211) |

```
--------------  POIDS DES OBSERVATIONS DU DEPARTEMENT DE L'HERAULT  ---------------

                                  Univariate Procedure

Variable=W2134 (pondération associée à l'estimateur yd21)

                Moments                                          Quantiles(Def=5)

 N              1475  Sum Wgts       1475           100% Max  7591.071       99%  5773.522
 Mean       593.9775  Sum         876116.7            75% Q3   181.3781       95%  4520.774
 Std Dev    1366.919  Variance     1868467            50% Med  105.3099       90%  3181.178
 Skewness   2.736303  Kurtosis    6.165738            25% Q1    66.7737       10%  49.49627
 USS        3.2745E9  CSS         2.7541E9             0% Min  28.09595        5%   42.2982
 CV         230.1298  Std Mean    35.59154                                     1%  33.54284
 T:Mean=0   16.68873  Pr>|T|        0.0001           Range     7562.975
 Num ^= 0       1475  Num > 0         1475           Q3-Q1     114.6044
 M(Sign)       737.5  Pr>=|M|       0.0001           Mode      103.5726
 Sgn Rank     544275  Pr>=|S|       0.0001

                                     Extremes

                                Lowest    Obs     Highest    Obs
                              28.09595(    1297)  6261.64(     153)
                               28.9079(    1205) 6586.929(     160)
                              29.01445(    1276) 6644.138(      88)
                              30.14154(    1356) 7013.511(     163)
                              30.76922(     883) 7591.071(      11)
```

La présence de poids négatifs n'est pas en soi un problème (sauf dans l'utilisation des procédures SAS). En revanche il est assez ennuyeux de constater qu'avec l'utilisation d'estimateurs combinés, la mesure du handicap dans l'Hérault provient essentiellement de l'échantillon local. De plus, j'ai la vague intuition qu'avec l'estimateur yd23, la somme des estimations locales n'est plus égale à l'estimation nationale.

Autres points à ne pas perdre de vue :

- L'étape finale de la construction de l'estimateur départemental est un redressement par calage sur les marges socio-démographiques du RP99 (sexe, tranche d'âges, taille de la commune et position sociale de la personne de référence du ménage). Ce calage est-il possible avec des poids négatifs ?
- Il ne faut pas perdre de vue la nécessité de définir des intervalles de confiance.

# *ANNEXE XIII*

## Réflexions du groupe de travail à propos de la stabilité de la mesure de l'effet départemental « résiduel »

### 1. Présentation résumée d'un modèle combiné à deux facteurs

#### *a) Présentation*

Partant du constat que la quasi totalité des résultats observés sur l'Hérault indique la persistance d'un effet local propre, Laurent Wilms propose la méthode ci-après pour "ajouter" au traditionnel effet socio-démographique révélé par le découpage en strates, un effet spécifique au département. D'un modèle à un facteur, on passe alors à un modèle dont l'estimateur est la combinaison de deux facteurs.

On peut résumer l'approche de la façon suivante :

1. Modèle de type petits domaines à un facteur :

Soit une strate h (par exemple, l'une des modalités de la variable de stratification V52 que l'on s'est créé) et le département d. Pour tout individu k appartenant à la strate h dans le département d,

$\forall$ k $\in$ (h,d), on a $y_k = \bar{y}_h$, le comportement de l'individu k est égal au comportement moyen des individus dans la strate h.

Ce comportement $\bar{y}_h$ est indépendant du département d :

$$\bar{y}_h = \frac{1}{N_h} \sum_{k \in h} y_k$$

Et l'estimateur post-stratifié à un facteur devient :

$$\hat{\bar{y}}_{d1} = \sum_{h=1}^{52} \frac{N_{h,d}}{N_d} \hat{\bar{y}}_h ,$$

où $\frac{N_{h,d}}{N_d}$ est la structure par strate dans le département d, tiré de l'enquête VQS.

2. Modèle de type petits domaines à deux facteurs :

$\forall$ k $\in$ (h,d), le modèle devient :

$y_k = a_h + b_d$, c'est à dire que le comportement d'un individu dépend de la strate dans laquelle il se situe ($a_h$) et d'un effet dû au département ($b_d$).

Son comportement peut être estimé par :

$$\hat{y}_k = \bar{y}_h + (\bar{y}_d - \bar{y})$$

Le comportement de l'individu k est égal au comportement moyen des individus dans la strate h, corrigé de l'écart entre moyenne départementale et nationale. Ce deuxième facteur ne dépend plus de la strate h.

Et l'estimateur post-stratifié à deux facteurs peut s'écrire :

$$\hat{\bar{y}}_{d2} = \hat{\bar{y}}_{d1} + (\bar{y}_d - \bar{y}).$$

Cet estimateur peut être calculé à partir du fichier national, en modifiant les pondérations de façon unique pour l'ensemble des variables d'intérêt HID. Laurent Wilms se propose, après consultation de Jean-Claude Deville, d'exposer par écrit la méthode et de préciser l'expression des nouvelles pondérations.

***b) Discussion***

Bien que le modèle à deux facteurs paraisse mieux adapté à la situation observée dans l'Hérault, on va devoir le tester et en premier lieu se demander si la différence $(\bar{y}_d - \bar{y})$ peut être évaluée de façon significative dans un département qui, contrairement à l'Hérault, n'aurait pas d'extension de l'enquête HID, c'est à dire qui ne disposerait en moyenne que d'environ 150 observations HID.

Si la réponse est positive, alors on pourra adopter l'estimateur combiné avec la différence $(\bar{y}_d - \bar{y})$ évaluée au niveau du département (tout en sachant que cette éventualité est peu probable car elle conduirait à revaloriser l'estimation directe).

Dans le cas contraire, il faudrait remplacer $(\bar{y}_d - \bar{y})$ par $(\bar{y}_r - \bar{y})$, où r représente une zone géographique à "comportements résiduels homogènes", plus vaste que le département mais pas nécessairement contigüe, et qui reste à définir.

## 2. Les tests envisagés pour valider le modèle à deux facteurs, avec effet départemental

L'ajout de l'effet départemental dans l'Hérault est probablement bénéfique étant donné que l'on dispose d'environ 1 500 individus HID dans ce département. Cependant, l'objectif est davantage de se servir des données de l'Hérault pour construire les tests qui permettront de valider la méthode sur d'autres départements sans extension HID. Ainsi, François Clanché propose de tirer aléatoirement, dans l'échantillon de l'Hérault, des sous-échantillons de taille équivalente à celle des échantillons des autres départements et d'observer la stabilité de $(\bar{y}_d - \bar{y})$. Selon les précisions de Pierre Mormiche sur la méthode d'échantillonnage, cela revient à sélectionner quelques enquêteurs (environ 4) dans HID Hérault et à répéter cette opération une dizaine de fois. Il s'agit d'une application de la méthode de "boot-strap", qui sera menée par Frédérique Tardieux.

Doit-on calculer la moyenne départementale $(\bar{y}_d)$ brute ou après calage de l'échantillon de l'Hérault sur les marges tirées du recensement ? Pour Laurent Wilms le risque du calage sur les marges du RP est d'entraîner des déformations de structure, tandis qu'une post-stratification générerait de nombreuses cases vides (1 500 individus pour une cinquantaine de cases). Le choix serait d'une post-stratification plus agrégée, d'une dizaine de strates, telle que le croisement en 10 postes du degré de gravité du handicap (groupe VQS) avec l'âge, croisement qui a servi de stratification au tirage d'HID dans VQS.

Des tests identiques devront être menés sur d'autres zones que l'Hérault, des regroupements de départements en région (sur l'Ile de France par exemple).

Enfin, si les écarts $(\bar{y}_d - \bar{y})$ ne sont pas significatifs lorsque l'on dispose d'environ 150 observations départementales, alors il faudra regrouper les départements en régions (la région étant prise au sens large de zone géographique pas nécessairement contigüe).

# *ANNEXE XIV*

## Contrôle de la stabilité départementale de la mesure de l'effet résiduel dans le modèle de comportement à deux facteurs

### Application de la méthode de « Boot-Strap »

La recherche d'un estimateur local indirect a conduit à définir un modèle de comportement à deux facteurs. Dans ce modèle, le comportement individuel dépend non seulement de la strate dans laquelle se situe l'individu mais aussi d'autres effets, dont il est difficile de préciser l'origine, mais qui sont propres à chaque département.

La question qui se pose est de savoir si l'écart entre les comportements moyens observés entre le département et le territoire national constitue une mesure fiable de ce deuxième facteur. Autrement dit, l'échantillon HID d'un département quelconque est-il suffisamment important pour fournir une mesure stable du comportement moyen local ?

Grâce à l'extension de l'enquête HID dans l'Hérault, on a pu reconstituer différents échantillons possibles de l'enquête pour lesquels la part revenant à l'Hérault est comparable à celle des autres départements. Plus précisément, des 29 enquêteurs sélectionnés dans l'Hérault, 4 ont été retenus pour se ramener à la situation d'un département «normal ». Sur les 23 751 échantillons correspondant à l'ensemble des combinaisons possibles de 4 enquêteurs parmi les 29 de l'Hérault, on a procédé à des estimations de prévalence de handicap au moyen :

- d'estimateurs directs, calculés à partir de l'échantillon local :
  **y0** (simple) et **yd0** (post-stratifié sur les 52 strates définies à partir de VQS).

- d'estimateurs indirects, calculés à partir de l'échantillon national :
  l'estimateur indirect post-stratifié «classique » **yd1** et 3 estimateurs indirects combinés **yd21** (avec effet local mesuré par y0), **yd22** (l'effet local = yd0) et **yd23**.

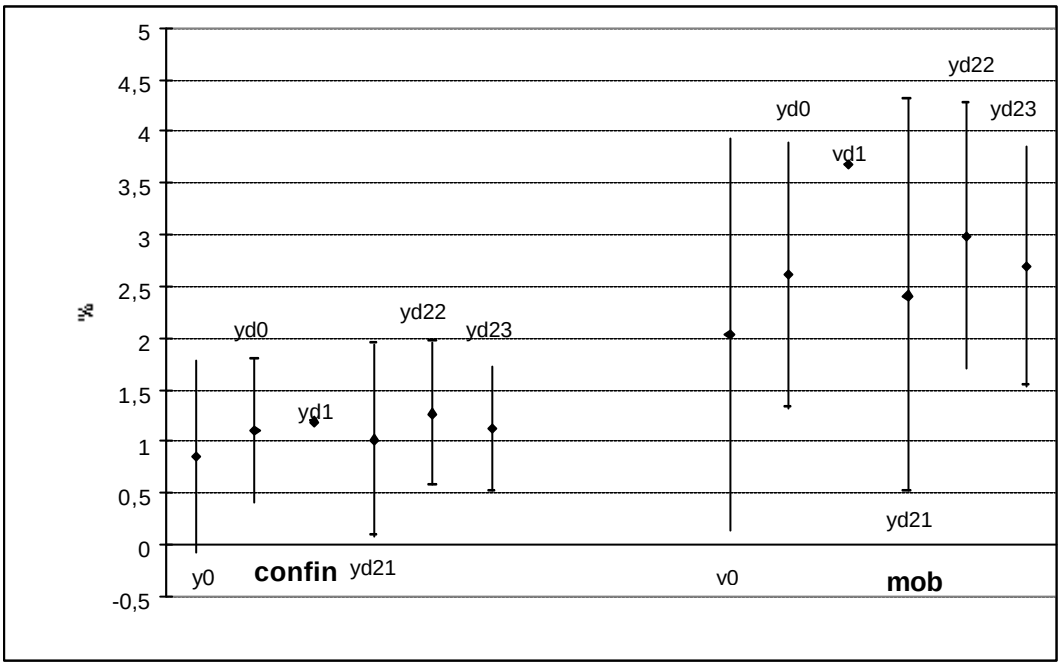
Les moyennes et les intervalles de confiance à 95 % des estimations concernant quelques unes des 10 variables habituelles d'HID sont graphiqués page suivante (les tableaux statistiques sont présentés à la suite ainsi qu'une des courbes de densité de probabilité). On retiendra comme enseignement que :

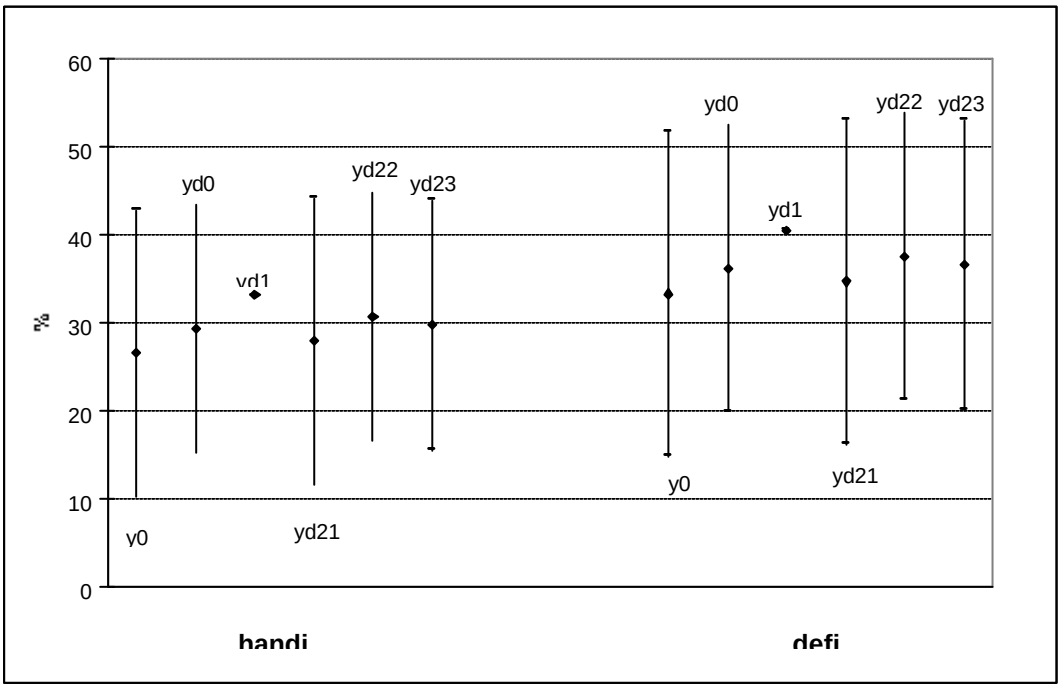
- La variance de l'estimateur **indirect** post-stratifié « classique » **yd1** est très faible.

- Les 3 estimateurs **indirects à deux facteurs** ont des variances beaucoup plus importantes, comparables à celles des **estimateurs directs**. Plus précisément l'intervalle de confiance de **yd21** rappelle celui de l'estimateur direct simple **y0**, l'amplitude de l'intervalle de **yd22** et **yd23** est à peine plus réduite, comparable à celle de l'estimateur direct post-stratifié **yd0.**

**En conclusion**, l'estimateur indirect « combiné », avec effet résiduel mesuré sur un département, n'offre pas plus de précision que l'estimateur direct.

Moyennes et intervalles de confiance à 95 % des estimations de quelques prévalences




On sait par ailleurs que mesurer l'effet résiduel à partir **d'une classe de départements**[21], conduit à réintroduire du biais. On se rapproche alors des résultats de l'estimateur post-

[21] voir annexes suivantes.

stratifié «classique » **yd1,** avec d'autant plus de biais et d'autant moins de variance que la classe de départements s'élargit. En final, **yd1 peut paraître supérieur** à cette deuxième catégorie d'estimateurs combinés dans la mesure où les pondérations associées à cet estimateur ne sont jamais négatives.

**Résultats détaillés du test :**
**Les principales statistiques (moyennes, écart type, …)**
**concernant chaque estimateur pour les 10 variables d'HID**

confin

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.008534 | 0.004741 | 202.702666 | 0 | 0.039785 |
| YD0 | 23751 | 0.011049 | 0.003531 | 262.433819 | 0 | 0.022796 |
| YD1 | 23751 | 0.011925 | 0.000005747 | 283.241317 | 0.011911 | 0.011950 |
| YD21 | 23751 | 0.010203 | 0.004742 | 242.326206 | 0.001668 | 0.041455 |
| YD22 | 23751 | 0.012718 | 0.003532 | 302.057359 | 0.001668 | 0.024465 |
| YD23 | 23751 | 0.011244 | 0.003079 | 267.058999 | 0.001668 | 0.021720 |
| YFR | 23751 | 0.010257 | 0.000004703 | 243.617777 | 0.010245 | 0.010277 |

aidki

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.071495 | 0.028000 | 1698.082562 | 0.005363 | 0.176942 |
| YD0 | 23751 | 0.086168 | 0.022713 | 2046.578531 | 0.024216 | 0.185708 |
| YD1 | 23751 | 0.096731 | 0.000046238 | 2297.468866 | 0.096592 | 0.096930 |
| YD21 | 23751 | 0.080510 | 0.028001 | 1912.197364 | 0.014382 | 0.185953 |
| YD22 | 23751 | 0.095183 | 0.022714 | 2260.693333 | 0.033230 | 0.194719 |
| YD23 | 23751 | 0.088898 | 0.022606 | 2111.415658 | 0.032043 | 0.193140 |
| YFR | 23751 | 0.087716 | 0.000045927 | 2083.354065 | 0.087588 | 0.087920 |

dadapt

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.024134 | 0.014186 | 573.200959 | 0 | 0.097318 |
| YD0 | 23751 | 0.029488 | 0.012197 | 700.375412 | 0 | 0.081081 |
| YD1 | 23751 | 0.026732 | 0.000023918 | 634.908753 | 0.026678 | 0.026827 |
| YD21 | 23751 | 0.027130 | 0.014188 | 644.362210 | 0.002993 | 0.100317 |
| YD22 | 23751 | 0.032484 | 0.012198 | 771.536663 | 0.002993 | 0.084079 |
| YD23 | 23751 | 0.029515 | 0.011470 | 701.014492 | 0.002993 | 0.079594 |
| YFR | 23751 | 0.023736 | 0.000022968 | 563.747501 | 0.023689 | 0.023824 |

ralloc

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.025794 | 0.011275 | 612.644396 | 0.003248 | 0.074381 |
| YD0 | 23751 | 0.031575 | 0.009215 | 749.940044 | 0.015896 | 0.079834 |
| YD1 | 23751 | 0.040956 | 0.000025081 | 972.745666 | 0.040908 | 0.041033 |
| YD21 | 23751 | 0.026941 | 0.011277 | 639.883629 | 0.004395 | 0.075538 |
| YD22 | 23751 | 0.032722 | 0.009218 | 777.179278 | 0.017042 | 0.080991 |
| YD23 | 23751 | 0.032033 | 0.009386 | 760.816327 | 0.016806 | 0.081739 |
| YFR | 23751 | 0.039809 | 0.000021627 | 945.506432 | 0.039763 | 0.039877 |

rinval

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.038469 | 0.015891 | 913.674954 | 0.004853 | 0.080258 |
| YD0 | 23751 | 0.047648 | 0.010999 | 1131.680247 | 0.021318 | 0.105034 |
| YD1 | 23751 | 0.063131 | 0.000030574 | 1499.434186 | 0.063035 | 0.063227 |
| YD21 | 23751 | 0.040633 | 0.015892 | 965.075198 | 0.007015 | 0.082422 |
| YD22 | 23751 | 0.049812 | 0.011001 | 1183.080491 | 0.023482 | 0.107207 |
| YD23 | 23751 | 0.048024 | 0.011104 | 1140.611475 | 0.023371 | 0.106321 |
| YFR | 23751 | 0.060967 | 0.000027701 | 1448.033941 | 0.060875 | 0.061052 |

handi

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.265647 | 0.083158 | 6309.375191 | 0.095580 | 0.942044 |
| YD0 | 23751 | 0.293728 | 0.071898 | 6976.324321 | 0.110732 | 0.941569 |
| YD1 | 23751 | 0.331503 | 0.000135 | 7873.520017 | 0.331073 | 0.331950 |
| YD21 | 23751 | 0.279161 | 0.083156 | 6630.353702 | 0.109093 | 0.955556 |
| YD22 | 23751 | 0.307242 | 0.071897 | 7297.302833 | 0.124244 | 0.955080 |
| YD23 | 23751 | 0.297576 | 0.072367 | 7067.718538 | 0.118300 | 0.954163 |
| YFR | 23751 | 0.317988 | 0.000137 | 7552.541505 | 0.317544 | 0.318446 |

mob

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.020346 | 0.009670 | 483.235639 | 0.003187 | 0.085210 |
| YD0 | 23751 | 0.026104 | 0.006530 | 619.984441 | 0.007317 | 0.056340 |
| YD1 | 23751 | 0.036784 | 0.000013440 | 873.655134 | 0.036736 | 0.036822 |
| YD21 | 23751 | 0.024130 | 0.009670 | 573.102968 | 0.006971 | 0.088995 |
| YD22 | 23751 | 0.029887 | 0.006532 | 709.851770 | 0.011101 | 0.060127 |
| YD23 | 23751 | 0.026963 | 0.005916 | 640.386780 | 0.010439 | 0.053851 |
| YFR | 23751 | 0.033000 | 0.000011528 | 783.787806 | 0.032959 | 0.033031 |

defi

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.333428 | 0.094090 | 7919.255492 | 0.100288 | 0.942044 |
| YD0 | 23751 | 0.362424 | 0.082927 | 8607.928610 | 0.144862 | 0.941569 |
| YD1 | 23751 | 0.405592 | 0.000171 | 9633.226324 | 0.404975 | 0.406183 |
| YD21 | 23751 | 0.347025 | 0.094088 | 8242.192793 | 0.113881 | 0.955635 |
| YD22 | 23751 | 0.376021 | 0.082926 | 8930.865912 | 0.158451 | 0.955159 |
| YD23 | 23751 | 0.365752 | 0.084143 | 8686.976729 | 0.147018 | 0.954243 |
| YFR | 23751 | 0.391996 | 0.000174 | 9310.289023 | 0.391347 | 0.392576 |

cotor

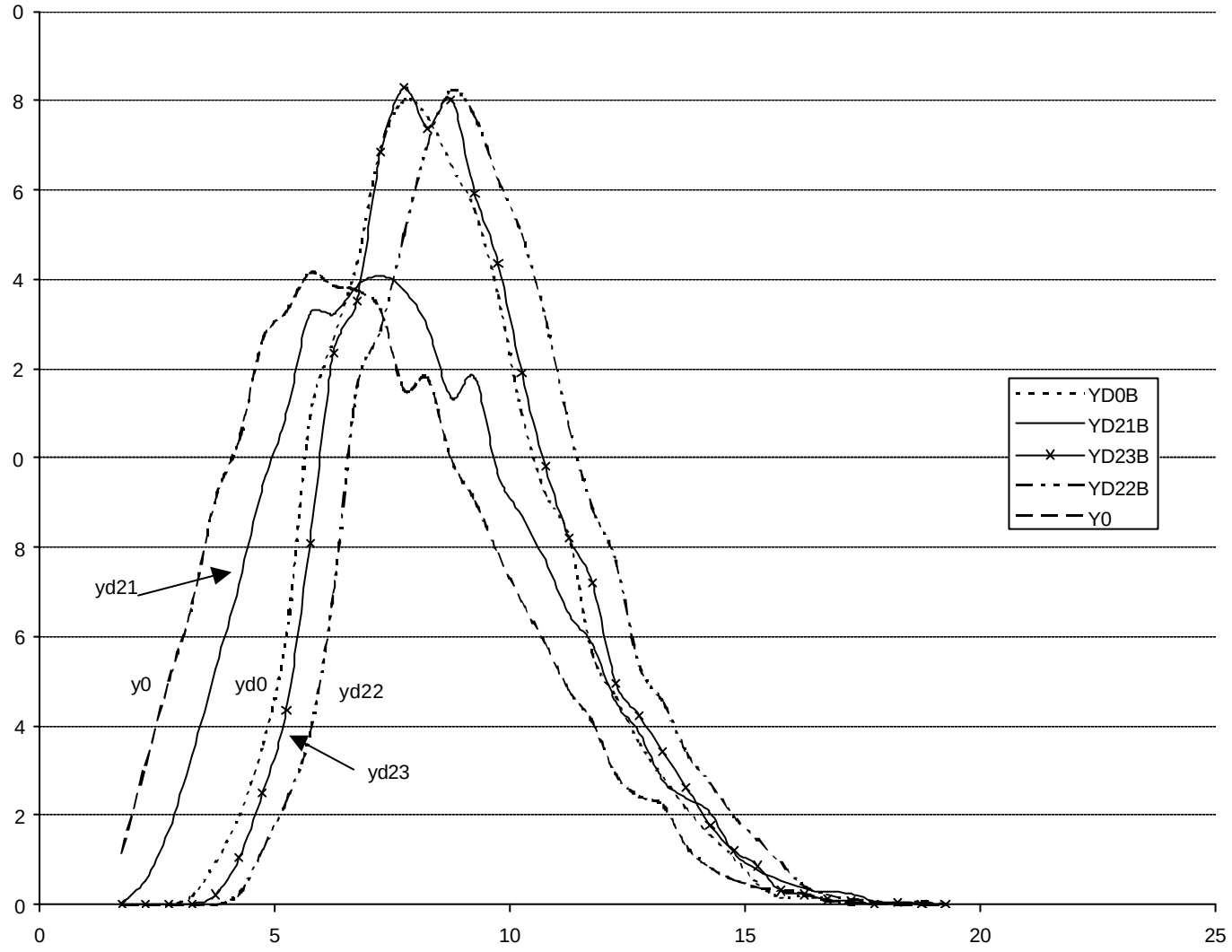
| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.035109 | 0.015057 | 833.870920 | 0 | 0.107926 |
| YD0 | 23751 | 0.040174 | 0.011243 | 954.162416 | 0 | 0.095430 |
| YD1 | 23751 | 0.041393 | 0.000026460 | 983.126820 | 0.041332 | 0.041480 |
| YD21 | 23751 | 0.035833 | 0.015057 | 851.061120 | 0.000724 | 0.108650 |
| YD22 | 23751 | 0.040897 | 0.011242 | 971.352615 | 0.000724 | 0.096152 |
| YD23 | 23751 | 0.039944 | 0.011195 | 948.713937 | 0.000724 | 0.094562 |
| YFR | 23751 | 0.040669 | 0.000027023 | 965.936621 | 0.040609 | 0.040759 |

expr

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Y0 | 23751 | 0.009750 | 0.009852 | 231.582195 | 0 | 0.048244 |
| YD0 | 23751 | 0.011014 | 0.010325 | 261.584331 | 0 | 0.055881 |
| YD1 | 23751 | 0.025926 | 0.000026418 | 615.757501 | 0.025862 | 0.026012 |
| YD21 | 23751 | 0.009254 | 0.009852 | 219.785851 | -0.000499 | 0.047745 |
| YD22 | 23751 | 0.010517 | 0.010325 | 249.787987 | -0.000499 | 0.055383 |
| YD23 | 23751 | 0.010083 | 0.010351 | 239.477950 | -0.000499 | 0.055668 |
| YFR | 23751 | 0.026422 | 0.000026497 | 627.553845 | 0.026357 | 0.026507 |

**Courbe de densité de probabilité associée aux différents estimateurs**

***(exemple sur la variable : « avoir besoin d'une aide »)***



# *ANNEXE XV*

## LES ESTIMATEURS COMBINES ASSOCIES A UNE CLASSIFICATION DEPARTEMENTALE

Dans l'hypothèse où le nombre d'observations HID d'un département quelconque serait insuffisant pour mesurer de façon fiable le deuxième terme de l'estimateur combiné (à savoir les écarts entre comportements locaux et nationaux), il avait été envisagé d'effectuer des regroupements de départements ayant des prévalences de handicap voisines. Les similitudes entre départements devaient être appréciées au niveau de leurs comportements résiduels, c'est à dire après élimination des effets dus à la structure de leur population selon le sexe, l'âge, …etc (effets déjà pris en compte dans le premier terme de l'estimateur). En suivant cette démarche, nous examinerons successivement :

- Une classification des départements sur la base de l'enquête VQS.

- Les résultats d'estimateurs combinés dont les comportements résiduels ont été observés sur un groupe de départements « voisins » (calculs effectués dans l'Hérault et en Ille et Vilaine).

### I – Classification des départements d'après l'enquête VQS.

La classification a été faite à partir du fichier VQS pour lequel on dispose d'observations départementales en nombre suffisant (à l'exceptions des départements 09, 11, 32, 43, 58, 65 et 72 qui n'ont pas d'individus VQS).

La répartition départementale de la population en ménage dans les six groupes de handicap croissant (gp1 à gp6) constitue les six variables actives de l'analyse. Afin d'éliminer les effets de la structure socio-démographique départementale sur cette répartition, celle-ci a été calculée sur des départements standardisés[22] : leur structure par sexe et groupe d'âges a été calquée sur la structure nationale en 10 strates (2 sexes * 5 tranches d'âges).

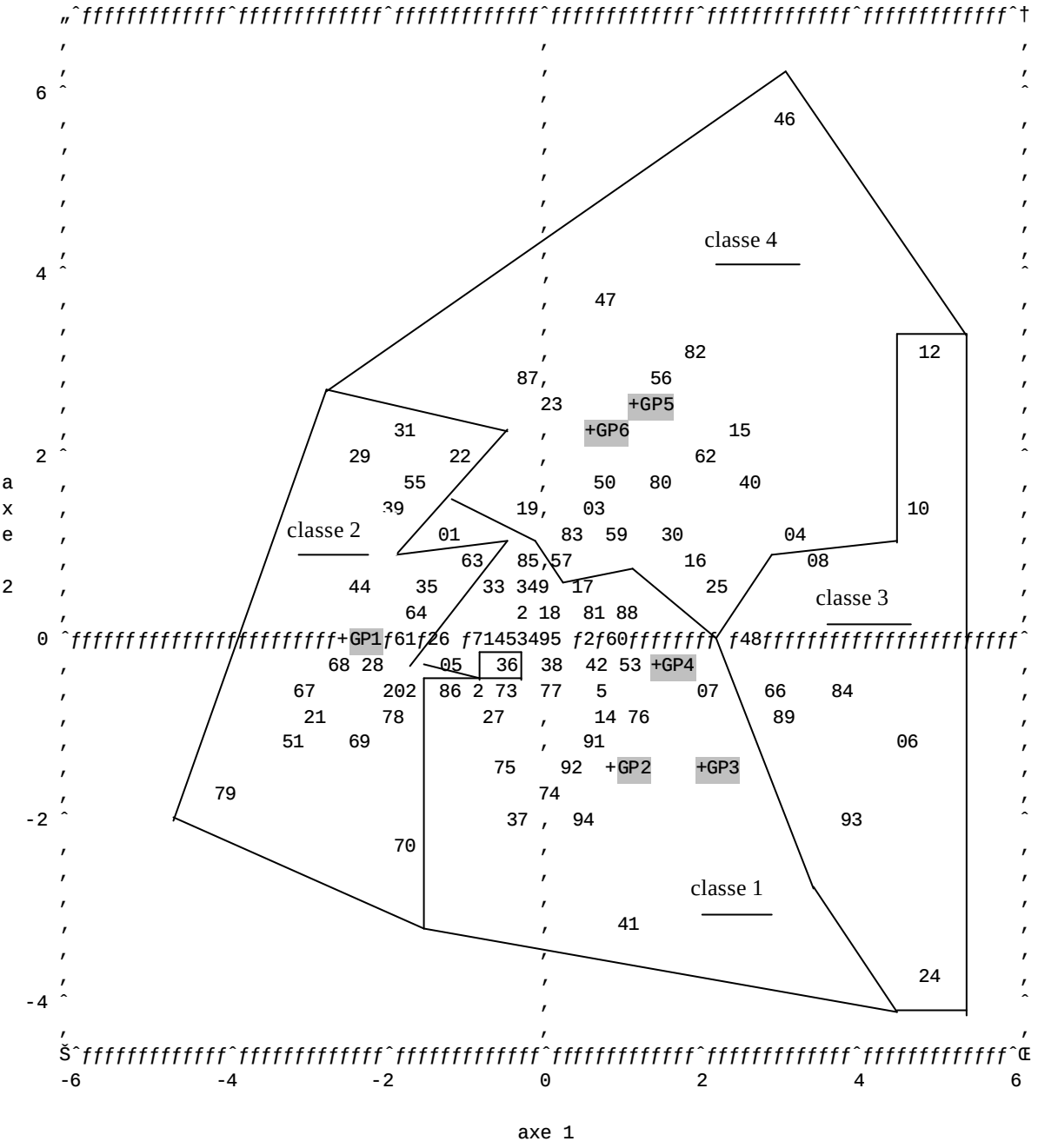
14 variables supplémentaires ont été ajoutées, caractérisant la répartition de la population vivant en ménage selon le sexe (2), le groupe d'âges (5), la taille de la commune (en 3 postes : rurales, urbaines de moins de 100 000 habitants et de plus de 100 000) et le niveau d'études de la personne de référence du ménage (en 4 positions). Toutes ces variables sont tirées de l'exploitation départementale du RP99.

[22] La standardisation a été obtenue en multipliant les pondérations des individus VQS de la strate **s** dans le département **d** par le ratio des poids de cette strate entre les échelons national et départemental. La répartition départementale des individus selon le groupe VQS, calculée avec cette nouvelle pondération, est alors représentative de comportements locaux résiduels.

L’existence au sein du département d’une capitale régionale est introduite sous la forme d’une variable supplémentaire qualitative.

Une APC et une CAH ont été réalisées sur ces données.

Selon l’ACP, la part d’inertie expliquée par le plan des deux premiers axes est de 78%. Le graphique ci-dessous, associée à cette ACP, représente les individus actifs (départements) et les axes unitaires sur le plan 1 – 2.




L’axe 1 oppose des départements ayant une proportion élevée de leur population ne souffrant d’aucune déficience (départements 51, 67, 69) aux départements ayant une part relativement importante de leur population dans les groupes de handicap 3 et 4 (06, 24, 84, 93).

L'axe 2 oppose les départements à déficience moyenne marquée (groupe 3 fort dans le 24 et le 93 par exemple) à ceux connaissant plutôt des déficiences lourdes (groupes 5 et 6 importants dans les départements 46, 56, 62).

Sur cette représentation graphique les départements ont été rassemblés en 4 classes qui correspondent à la partition la plus équilibrée, résultant de la CAH.

## *Partition en 4 classes : composition et particularité des classes*

### *Classe numéro 1 (en noir sur la carte page suivante)*

01 05 07 13 14 17 18 26 27 33 34 37 38 41 42 45 49 52 53 54 60 71 73 74 75 76 77 81 85 86 88 91 92 94 95

Dans cette classe les handicaps lourds sont peu fréquents, tandis que les déficiences légères semblent assez bien représentées. Le comportement résiduel en matière de handicap y est donc plutôt moyen, voire faible. Par ailleurs, ces départements sont constitués d'une population plutôt d'âge actif (les 20-60 ans sont assez présents), vivant dans des villes de taille plutôt moyenne à grande et ayant un niveau d'études assez élevé.

### *Classe numéro 2 (en quadrillé à grandes mailles)*

02 21 22 28 29 31 35 36 39 44 51 55 61 63 64 67 68 69 70 78 79

C'est dans cette classe que l'on trouve la plus forte proportion de personnes ne souffrant d'aucun handicap. Les déficiences légères y sont peu présentes et, dans une moindre mesure, il en est de même des handicaps lourds (groupes 5 et 6). Ces populations, bien portantes, sont plutôt masculines et vivent un peu plus en milieu rural que la moyenne nationale.

### *Classe numéro 3 (symbolisée par des bouliers sur la carte)*

06 08 10 12 24 48 66 84 89 93

Ces départements se caractérisent par une faible proportion d'individus ne présentant aucune déficience. En revanche, la proportion des groupes 2, 3 et 4 y est assez élevée, tandis que les groupes 5 et 6 pèsent à peine un peu plus lourds que la moyenne. Globalement cette classe présente un handicap «résiduel » un peu supérieur à celui de la classe 1. Cette population, légèrement plus féminine que la moyenne nationale, a un niveau d'études inférieur à la moyenne. Elle est bien représentée dans les tranches d'âges élevés et vit plutôt en milieu rural mais, curieusement, elle est aussi présente dans les grandes cités urbaines.

## *Classe numéro 4 (en pointillés fins)*

03 04 15 16 19 23 25 30 40 46 47 50 56 57 59 62 80 82 83 87

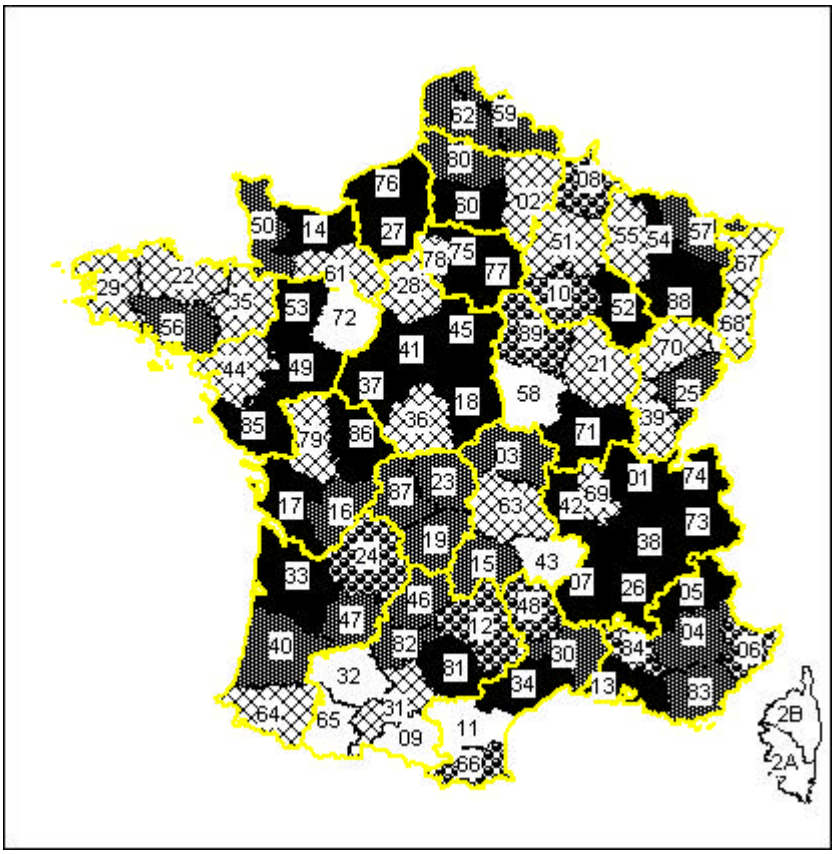
C'est la classe des handicaps lourds, avec une population âgée bien representée, vivant plutôt en milieu rural et dont le niveau d'études s'est arrêté, plus souvent qu'ailleurs, au niveau du primaire ou du collège.

La présence d'une capitale régionale ne semble pas jouer un rôle discriminant au sein des départements.

Au niveau géographique, la partition en 4 classes ne regroupe pas forcément des départements contigus et ne se recoupe pas vraiment avec le découpage régional traditionnel comme le montre la carte suivante.

## *Partition géographique de la France en fonction du comportement résiduel départemental*

relatif au handicap




NB : les départements 09, 11, 32, 43, 58, 65 et 72 n'ont pas d'individus VQS et la Corse a été exclue de l'analyse.

**I I – Application de la typologie en 4 classes au calcul de l'estimateur combiné.**

L'estimateur local prend en compte, en plus d'un effet de structure local, un effet de comportement résiduel propre à la classe à laquelle appartient le département étudié. Cet estimateur s'obtient à partir du fichier national HID, en modifiant les pondérations individuelles. Les nouvelles pondérations sont définies en **annexe XVI** : seul le deuxième terme change par rapport aux expressions présentées précédemment.

II.1 – Le cas particulier de l'Hérault.

Soit,
yd0 = estimateur direct avec calage sur marges VQS + RP99,

yd1 = estimateur post-stratifié indirect à un facteur, sans effet résiduel départemental,

yd21 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel départemental simple,
yd22 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel départemental redressé,
yd23 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel départemental normalisé,

yd31 = estim. post-stratifié indirect à 2 facteurs, avec effet départemental simple mesuré sur une classe de départements,
yd32 = estim. post-stratifié indirect à 2 facteurs, avec effet départemental redressé mesuré sur une classe de départements,
yd33 = estim. post-stratifié indirect à 2 facteurs, avec effet départemental normalisé mesuré sur une classe de départements,

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR | DISTANCE (1) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **yd0** | **1,11** | **8,42** | **2,64** | **4,02** | **5,77** | **27,71** | **2,56** | **36,27** | **5,93** | **1,98** | |
| **yd1** | 1,19 | 9,66 | 2,68 | 4,08 | 6,29 | 33,09 | 3,82 | 40,50 | 4,73 | 2,57 | 0,68 |
| **yd21** | 0,93 | 7,60 | 2,47 | 2,53 | 3,64 | 25,77 | 2,11 | 32,65 | 2,84 | 1,03 | 0,93 |
| **yd22** | 1,30 | 9,72 | 3,31 | 3,30 | 4,93 | 29,75 | 2,95 | 36,67 | 3,80 | 1,23 | 0,69 |
| **yd23** | 1,14 | 9,12 | 3,01 | 3,29 | 4,82 | 28,84 | 2,66 | 35,70 | 3,77 | 1,20 | 0,62 |
| (yd1-yd0) | 0,08 | 1,24 | 0,04 | 0,07 | 0,53 | 5,38 | 1,26 | 4,23 | -1,20 | 0,59 | |
| (yd21-yd0) | -0,19 | -0,83 | -0,17 | -1,49 | -2,13 | -1,93 | -0,45 | -3,62 | -3,09 | -0,95 | |
| (yd22-yd0) | 0,19 | 1,29 | 0,68 | -0,72 | -0,84 | 2,04 | 0,39 | 0,41 | -2,12 | -0,75 | |
| (yd23-yd0) | 0,03 | 0,70 | 0,37 | -0,73 | -0,95 | 1,14 | 0,10 | -0,56 | -2,16 | -0,78 | |
| **yd31** | 1,16 | 8,43 | 2,53 | 3,77 | 5,40 | 33,26 | 3,17 | 39,07 | 3,41 | 2,54 | 0,61 |
| **yd32** | 1,16 | 8,36 | 2,51 | 3,65 | 5,24 | 33,15 | 3,14 | 38,97 | 3,29 | 2,51 | 0,62 |
| **yd33** | 1,24 | 8,83 | 2,69 | 3,92 | 5,65 | 33,82 | 3,38 | 39,63 | 3,58 | 2,55 | 0,64 |
| (yd31-yd0) | 0,05 | 0,01 | -0,11 | -0,25 | -0,37 | 5,55 | 0,61 | 2,81 | -2,51 | 0,56 | |
| (yd32-yd0) | 0,05 | -0,07 | -0,12 | -0,37 | -0,53 | 5,44 | 0,58 | 2,71 | -2,64 | 0,53 | |
| (yd33-yd0) | 0,13 | 0,40 | 0,05 | -0,10 | -0,11 | 6,11 | 0,82 | 3,36 | -2,34 | 0,57 | |

Le tableau de l'estimation locale dans l'Hérault a donc été complété par 3 nouveaux estimateurs : yd31, yd32 et yd33, dont le mode de calcul est défini **en annexe XVI**.

Les nouvelles pondérations associées à ces estimateurs ont assez souvent un signe négatif, comme le souligne **l'annexe XIX**.

Le critère de distance retenu (1) pour un estimateur donné est la racine carrée de la somme sur les dix variables, du carré de l'écart relatif à l'estimateur direct.

Le gain, mesuré en terme de distance, d'un estimateur combiné à deux facteurs ne semble pas déterminant.

Avec l'estimateur yd23, ce gain par rapport à yd1 apparaît moins nettement que sur les graphiques de la note précédente. Une amélioration appréciable de l'estimation est toutefois obtenue pour les variables handi, mob, defi, confin et aidki, tandis que la situation se dégrade plutôt sur les autres variables, plus institutionnelles.
En revanche, la situation est plus figée avec yd33, même si globalement la distance à yd0 est à peine plus importante qu'en yd23. En définitive, c'est comme si avec yd33 on retrouvait une situation voisine de yd1, estimateur post-stratifié indirect à un facteur. Cela revient à dire que le deuxième terme de l'estimateur, la correction d'un comportement résiduel, n'a pas eu beaucoup d'effet, et donc que la classification retenue est contestable. Il est vrai que la classe 1, dont dépend l'Hérault, a les comportements résiduels les plus proches de la moyenne nationale.

II.2 – Le cas de l'Ille et Vilaine.

Voici à titre indicatif, l'éventail des résultats des différents estimateurs dans un département avec extension VQS et sans extension HID. L'Ille et Vilaine appartient à la classe 2.

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR |
|---|---|---|---|---|---|---|---|---|---|---|
| **yd1** | 0,91 | 7,93 | 2,14 | 3,69 | 5,68 | 30,04 | 3,12 | 37,36 | 4,33 | 2,48 |
| **yd21** | 0,06 | 8,29 | 2,74 | 2,83 | 5,89 | 33,01 | 2,61 | 38,19 | 3,41 | 2,42 |
| **yd22** | 0,27 | 10,73 | 3,87 | 3,79 | 7,94 | 36,92 | 3,51 | 42,43 | 4,99 | 2,32 |
| **yd23** | 0,32 | 11,76 | 4,43 | 3,92 | 8,45 | 38,75 | 3,67 | 44,47 | 5,39 | 2,22 |
| **yd31** | 0,65 | 7,37 | 1,77 | 3,90 | 5,76 | 24,65 | 2,55 | 33,64 | 3,14 | 1,87 |
| **yd32** | 0,77 | 8,13 | 2,04 | 4,25 | 6,27 | 26,13 | 2,83 | 35,04 | 3,51 | 1,91 |
| **yd33** | 0,83 | 8,66 | 2,17 | 4,63 | 6,81 | 27,54 | 3,01 | 36,38 | 3,94 | 1,98 |

Les estimateurs indirects à deux facteurs, avec effet résiduel mesuré sur le département, ne sont pas fiables. Lorsque cet effet est mesuré sur l'ensemble de la classe 2, la prévalence du handicap a tendance à baisser par rapport à l'estimation yd1, puisque cette classe est caractérisée par son bon état physique.

# *ANNEXE XVI*

## NOUVELLES PONDERATIONS DU FICHIER HID NATIONAL ASSOCIEES A UN ESTIMATEUR LOCAL A DEUX FACTEURS

### COMPRENANT UN EFFET RESIDUEL MESURE SUR UNE CLASSE DE DEPARTEMENTS

Soit c une classe de dépatement (c=1,..,4). En reprenant les expressions présentées dans la note précédente et après correction du deuxième terme, l'indicateur combiné avec effet résiduel mesuré sur la classe c devient :

Estimateur post-stratifié indirect avec effet départemental simple mesuré sur la classe c:

$$\hat{Y}_{d31} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \quad p_k} + \frac{\hat{N}_d}{\hat{N}_c \quad p_k} \; 1_c - \frac{\hat{N}_d}{\hat{N} \quad p_k} \right) y_k$$

Estimateur post-stratifié indirect avec effet départemental redressé mesuré sur la classe c:

$$\hat{Y}_{d32} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \quad p_k} + \frac{\hat{N}_d}{\hat{N}_c} \; \frac{\hat{N}_{s,c}}{\hat{N}_{s,c} \quad p_k} \; 1_c - \frac{\hat{N}_d}{\hat{N} \quad p_k} \right) y_k$$

Estimateur post-stratifié indirect avec effet départemental normalisé mesuré sur la classe c:

$$\hat{Y}_{d33} = \sum_{k \in HID} \left( \frac{\hat{N}_{h,d}}{\hat{N}_h \quad p_k} + \frac{\hat{N}_d \quad \hat{N}_s}{\hat{N} \quad \hat{N}_{s,c} \quad p_k} \; 1c - \frac{\hat{N}_d}{\hat{N} \quad p_k} \right) y_k$$

avec **1**c une variable indicatrice qui vaut :

- 1 si l'individu k appartient à la classe correspondant au département étudié,
- et 0 sinon.

et avec toujours comme règle d'écriture un $\hat{N}$ incliné quand il s'agit de la source HID et un $\hat{N}$ redressé quand l'information provient de VQS.

# *ANNEXE XVII*

## Bilan des résultats du modèle combiné à deux facteurs

(l'exemple de l'Hérault)

## 1. Bilan des résultats obtenus

Plusieurs méthodes ont été testées dans le département de l'Hérault. La difficulté d'effectuer un choix tient au fait que les résultats des tests sont inégaux selon la variable d'intérêt retenue.

### 1.1 Rappel des différents estimateurs utilisés

8 estimateurs ont été produits.

Le premier est la valeur à cibler, celle que l'on cherche à atteindre, sa connaissance dans l'Hérault étant estimée grâce à l'extension départementale de l'enquête HID.

yd0 = estimateur direct avec calage sur marges VQS + RP99,

Vient ensuite l'estimateur de type « petits domaines » classique :

yd1 = estimateur post-stratifié indirect à un facteur, sans effet résiduel départemental,

et parce que ce dernier donnait des résultats jugés peu satisfaisants, divers estimateurs combinés ont été définis :

un premier ensemble qui n'est applicable qu'au cas de l'Hérault, dans la mesure où pour les autres départements on ne dispose pas d'un échantillon HID suffisant pour évaluer un "résidu départemental" :

yd21 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel départemental simple,
yd22 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel départemental redressé,
yd23 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel départemental normalisé,

et un second ensemble applicable à tous les départements, qui évalue le "résidu départemental" au résidu moyen de la classe de départements dans laquelle on se situe (classe résultant d'une analyse et d'une CAH) :

yd31 = estim. post-stratifié indirect à 2 facteurs, avec effet départemental simple mesuré sur une classe de départements,

yd32 = estim. post-stratifié indirect à 2 facteurs, avec effet départemental redressé mesuré sur une classe de départements,

yd33 = estim. post-stratifié indirect à 2 facteurs, avec effet départemental normalisé mesuré sur une classe de départements,

### 1.2 Bilan des tests

**a) L'estimateur yd1** a été calculé sur une population locale stratifiée en 52 catégories selon le sexe, la tranche d'âges, le groupe VQS et la taille de la commune. La confection du fichier national associé à cet estimateur et son calage sur les résultats départementaux du RP99 ne pose pas de problème. De plus, par définition, ce fichier respecte les marges socio-démographiques du département étudié, ce qui lui confère un aspect plutôt rassurant.

Mais l'hypothèse d'égalité des comportements à l'intérieur de chaque strate, qui nous autorise à appliquer localement les comportements observés au niveau national, conduit à des résultats assez biaisés dans l'Hérault pour l'ensemble des 10 variables d'intérêt étudiées (comme s'il persistait un effet local propre qui tirerait les prévalences de ce département vers le bas).

**b) Quant aux estimateurs combinés,** ils sont mieux ciblés sur les particularités de l'Hérault lorsque l'effet résiduel est mesuré à partir des données départementales (amélioration sensible de l'estimation des variables handi et defi avec yd23 par exemple). En revanche le biais réapparaît partiellement quand cet effet est calculé sur un la classe de départements dont fait partie l'Hérault (avec yd33 l'estimation de ces deux variables se rapproche de celle obtenue avec yd1). Or, pour un département quelconque, il sera probablement nécessaire de passer par la construction de classes de départements, étant donné la faiblesse des échantillons HID locaux (voir point n°2). Mais, si le gain d'un estimateur combiné pour un département comme l'Hérault reste assez faible, il en va peut-être autrement des autres départements. Ainsi, selon François Clanché, faudrait-il poursuivre les tests sur d'autres zones géographiques (les « super régions » de Valérie Albouy par exemple).

**c)** La discussion sur le choix d'un estimateur combiné à conduit Laurent Wilms à s'interroger sur **la raison du calage** de $\bar{y}_d$ sur les marges nationales dans le calcul de l'effet résiduel $(\bar{y}_d - \bar{y})$ (autrement dit pourquoi choisir yd23 plutôt que yd22 ou bien yd33 à yd32 ?). D'après la définition du modèle à deux facteurs, cela ne s'imposerait pas selon lui. En revanche, pour Pierre Mormiche, ce calage permet de ne retenir comme résidu, que l'effet provenant du comportement spécifique au département. La question n'a finalement pas été tranchée mais il a été reconnu qu'après calage (évaluations des types 23 et 33), la somme des estimateurs départementaux n'avait plus de raison d'être égale aux estimations nationales (ce qui peut être gênant).

**d)** D'une façon plus générale, la critique a porté sur une présence importante de **poids négatifs** dans les fichiers associés aux estimateurs combinés. Les poids négatifs se trouvant sur les observations extérieures à la zone sur laquelle porte le calcul de l'effet résiduel, leur présence est donc moindre brsque ce calcul porte sur une classe de plusieurs départements (cas des estimateurs yd32 et yd33).

L'existence de poids négatifs peut avoir pour conséquence la production d'effectifs négatifs lorsque l'on s'intéresse à des sous-populations peu nombreuses (ce qui fait désordre). La présence de poids négatifs rend aussi impossible le calage du fichier sur les marges locales VQS et RP99 (CALMAR rejetant les observations ayant un poids négatif). De ce fait le fichier ne respecte plus les marges socio-démographiques du département ; il peut même, par construction, en être assez éloigné. Cela peut être troublant pour les utilisateurs.

Pour éviter d'avoir à fournir un fichier comportant des poids négatifs, Jean-Claude Deville a envisagé plusieurs solutions. La première consiste à remplacer notre modèle additif par un modèle multiplicatif : effet de la structure locale x effet résiduel de la zone. Ainsi à la place de poids négatifs on aurait des poids proches de zéro. Mais cette solution s'est avérée ne pas être opérationnelle, car elle ne permet pas d'obtenir des pondérations uniformes pour l'ensemble des variables d'intérêt. La deuxième solution consiste à masquer la présence des poids négatifs en associant à chaque individu de la zone d'étude (à poids positifs) les individus qui lui sont identiques ( du point de vue du handicap) situés à l'extérieur de cette zone et à ne retenir que la pondération résultante. Cette solution est malheureusement irréaliste étant donné le nombre considérable de variables de cette enquête.

**e) En conclusion,** il a été décidé de mieux apprécier l'importance des poids négatifs. Si leur présence n'entraîne pas de risque majeur, le choix se ferait plutôt en faveur d'un estimateur de type yd32 ou yd33. Dans le cas contraire on retiendrait l'estimateur yd1.

**f)** Par ailleurs, constatant que certaines **variables dépendent** énormément des **instances administratives locales** (la reconnaissance par les COTOREP par exemple), il a été suggéré d'améliorer la connaissance sur tout le domaine lié à cette variable par l'introduction d'une information externe (telle que le nombre de dossiers COTOREP du département). Ceci permettrait d'améliorer la qualité de l'enquête tout en confirmant des données déjà connues par ailleurs. Des contacts vont être pris avec la DREES (via Claude Gissot) pour obtenir une information départementale sur le nombre de COTOREP. Ce type d'information externe, dont les effets suivraient un découpage géographique propre à la variable, pourrait compléter le modèle d'estimation indirecte classique (yd1).

# *ANNEXE XVIII*

## Estimations locales sur des zones géographiques plus vastes que le département

(Tests complémentaires)

Lors de la dernière réunion du groupe de travail, il a été admis que l'application d'estimateurs combinés améliorait légèrement l'estimation de la prévalence du handicap dans le département de l'Hérault. Cependant, ces résultats présentaient l'inconvénient de découler de fichiers à pondération partiellement négative.

Il existe une interprétation à la présence de poids négatifs : ces poids correspondent à des individus qui appartiennent à des catégories sur-représentées. Toutefois cette présence peut entraîner certains désagréments pour les utilisateurs (cf. compte rendu de la réunion du groupe de travail du 19/03).

En conséquence, pour mieux évaluer le gain apporté par l'utilisation d'estimateurs combinés, on a étendu les tests à des zones géographiques plus vastes que le département, zones suffisamment vastes pour qu'une estimation directe ait un sens. Les résultats de ces tests sont présentés dans le classeur Excel ci-joint.

Définition des 8 super-zones (déjà utilisées par Valérie Albouy) :

ZONE 1 (Ile de France) : les départements 75,77,78,91,92,93,94,95.
ZONE 2 (Nord Ouest) : les départements 27,76,59,62,2,60,80.
ZONE 3 ( Ouest) : les départements 14,50,61,22,29,35,56,44,49,53,72,85,16,17,79,86.
ZONE 4 (Sud Ouest) : les départements 24,33,40,47,64,9,12,31,32,46,65,81,82.
ZONE 5 (Centre) : les départements 3,15,43,63,21,58,71,89,18,28,36,37,41,45,19,23,87.
ZONE 6 (Sud Est) : les départements 20,11,30,34,48,66,4,5,6,13,83,84.
ZONE 7 (Est) : les départements 25,39,70,90,1,7,26,38,42,69,73,74.
ZONE 8 (Nord Est) : les départements 67,68,8,10,51,52,54,55,57,88.

Rappel des 7 principaux estimateurs indirects :

L'estimateur de type « petits domaines » classique :

y1 = estimateur post-stratifié indirect à un facteur, sans effet résiduel.

Divers estimateurs combinés :

- un premier ensemble qui n'est applicable que sur une zone où on dispose d'un échantillon HID suffisant pour évaluer un "résidu", mais qui n'est pas transposable au niveau d'un département :

y21 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel simple mesuré sur la super-zone,
y22 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel mesuré sur la super-zone et redressé,
y23 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel mesuré sur la super-zone et normalisé ;

- un second ensemble applicable à tous les départements, qui évalue le "résidu de la zone" à l'aide du résidu moyen des classes de départements dans lesquelles on se situe :

y31 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel simple mesuré sur plusieurs classes de départements,
y32 = estim. post-stratifié indirect à 2 facteurs, avec effet résiduel redressé mesuré sur plusieurs classes de départements,
y33 = estim. post-stratifié indirect à 2 facteurs, avec effet départemental normalisé mesuré sur plusieurs classes de départements.

Définition des espaces sur lesquels est mesuré l'effet résiduel des estimateurs y31-32-33 :

Dans le cadre d'estimations départementales, on avait défini 4 classes de départements (résultat d'une CAH) qui permettaitent d'attribuer un comportement moyen à chacun des départements de la classe.

**Classe 1 :** 01, 05, 07, 13, 14, 17, 18, 26, 27 , 33, 34, 37, 38, 41, 42, 45, 49, 52, 53, 54, 60, 71, 73, 74, 75, 76, 77, 81, 85, 86, 88, 91, 92, 94, 95.

**Classe 2 :** 02, 21, 22, 28, 29, 31, 35, 36, 39, 44, 51, 55, 61, 63, 64, 67, 68, 69, 70, 78, 79.

**Classe 3** : 06, 08, 10, 12, 24, 48, 66, 84, 89, 93.

**Classe 4 :** 03, 04, 15, 16, 19, 23, 25, 30, 40, 46, 47, 50, 56, 57, 59, 62, 80, 82, 83, 87.

Maintenant, chacune des 8 super-zones qui constituent le cadre géographique de l'estimation chevauchent plusieurs de ces classes (3 ou 4 selon les cas). C'est donc sur les ensembles constitués par 3 ou 4 classes de départements que sont définis les comportements résiduels des 8 zones (l'ensemble représenté par les 4 classes de départements correspond au territoire national).

Exemple : la zone Ile de France touche l'ensemble constitué par la classe 01 (pour les départements 75, 77, 91, 92, 94 et 95), la classe 02 (département 78) et la classe 03 (département 93).

Conclusion :

Pour apprécier globalement les résultats, une notion de distance a été définie. Le critère de distance retenu pour un estimateur donné est la racine carrée de la somme sur les dix variables, du carré de l'écart relatif entre les résultats de cet estimateur et l'estimateur direct (y0).

$$\sqrt{\sum_{var}\left(\frac{\hat{\bar{y}}_{xx}-\hat{\bar{y}}_0}{\hat{\bar{y}}_0}\right)^2}$$

Les résultats des estimateurs combinés, avec effet résiduel mesuré localement sur chaque super-zone (estimateurs y21-22-23) sont sans surprise : ils sont bien meilleurs que ceux provenant d'un estimateur indirect « classique » (y1) (malheureusement, il ne sont pas applicables au niveau départemental).

Mais, comme dans le cas de l'Hérault, le biais réapparaît quand l'effet résiduel est mesuré sur une zone encore plus large (estimateurs y31-32-33). C'est le cas dans les zones 1, 2, 3, 5 et 8, où est localisée la majeure partie des départements qui nous intéressent (départements 77, 95, 27, 76, 62, 35 pour lesquels doivent être réalisées des estimations locales).

En résumé, dans le cas d'estimations départementales, la balance ne me semble pas pencher en faveur des estimateurs combinés mais plutôt du côté de l'estimateur indirect classique.

# Résultats des estimations locales

# sur les huit super-zones géographiques du territoire métropolitain

**ESTIMATION EN ZONE 1 (Ile de France)**

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR | DISTANCE | DISTANCE 2 hors ralloc rinval et cotor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ESTIMATION DIRECTE** | | | | | | | | | | | | |
| estim. directe simple | 0,73 | 5,25 | 1,53 | 2,54 | 3,68 | 29,99 | 2,05 | 34,77 | 2,71 | 4,48 | | |
| estim. directe après calage sur la zone VQS | 0,77 | 5,57 | 1,62 | 2,58 | 3,76 | 30,43 | 2,14 | 34,89 | 2,76 | 4,05 | | |
| estim. directe après calage sur la zone VQS +RP99 **(y0)** | **0,75** | **5,59** | **1,65** | **2,73** | **3,97** | **29,88** | **2,12** | **34,02** | **2,90** | **3,75** | | |
| **ESTIMATION INDIRECTE** | | | | | | | | | | | | |
| estim. indirecte simple | 0,88 | 7,66 | 1,96 | 3,38 | 5,16 | 29,83 | 2,97 | 37,28 | 3,93 | 2,66 | | |
| estim. indirecte après calage sur la zone RP99 **(y1)** | **0,77** | **6,83** | **1,90** | **3,26** | **5,24** | **28,64** | **2,67** | **34,28** | **4,12** | **2,81** | **0,72** | **0,45** |
| y21 | 0,60 | 4,10 | 1,11 | 1,96 | 2,70 | 28,04 | 1,57 | 32,87 | 1,89 | 4,51 | **0,80** | **0,58** |
| y22 | 0,63 | 4,35 | 1,15 | 2,09 | 2,88 | 28,59 | 1,65 | 33,46 | 2,05 | 4,53 | **0,69** | **0,51** |
| y23 | 0,82 | 5,35 | 1,43 | 2,57 | 3,67 | 30,06 | 2,15 | 34,89 | 2,81 | 4,52 | **0,29** | **0,27** |
| y31 | 0,81 | 7,07 | 1,72 | 3,30 | 4,67 | 28,63 | 2,60 | 35,78 | 3,52 | 2,71 | **0,57** | **0,46** |
| y32 | 0,83 | 7,20 | 1,77 | 3,32 | 4,71 | 28,83 | 2,67 | 35,98 | 3,55 | 2,71 | **0,61** | **0,50** |
| y33 | 0,89 | 7,55 | 1,88 | 3,56 | 5,06 | 29,40 | 2,83 | 36,53 | 3,88 | 2,76 | **0,80** | **0,60** |

Cette zone comprend les départements 75,**77**,78,91,92,93,94,**95.**
Elle est incluse dans l'ensemble formé par les classes 01, 02 et 03.

**ESTIMATION EN ZONE 2 (Nord Ouest)**

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR | DISTANCE | DISTANCE 2 hors ralloc rinval et cotor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ESTIMATION DIRECTE** | | | | | | | | | | | | |
| estim. directe simple | 1,10 | 11,23 | 3,14 | 5,85 | 7,62 | 36,45 | 3,76 | 46,51 | 6,72 | 3,17 | | |
| estim. directe après calage sur la zone VQS | 0,98 | 9,84 | 2,61 | 5,08 | 6,49 | 33,87 | 3,29 | 43,83 | 5,89 | 3,14 | | |
| estim. directe après calage sur la zone VQS +RP99 **(y0)** | **1,00** | **10,04** | **2,68** | **5,50** | **6,98** | **34,40** | **3,23** | **43,61** | **6,52** | **3,19** | | |
| **ESTIMATION INDIRECTE** | | | | | | | | | | | | |
| estim. indirecte simple | 0,99 | 8,65 | 2,34 | 4,06 | 6,26 | 31,62 | 3,39 | 38,96 | 4,91 | 2,70 | | |
| estim. indirecte après calage sur la zone RP99 **(y1)** | **1,03** | **9,17** | **2,45** | **4,53** | **6,72** | **32,10** | **3,61** | **40,21** | **5,63** | **3,02** | **0,31** | **0,21** |
| y21 | 1,07 | 11,08 | 3,09 | 5,95 | 7,75 | 36,29 | 3,73 | 46,29 | 6,95 | 3,23 | **0,30** | **0,26** |
| y22 | 0,93 | 9,97 | 2,72 | 5,44 | 6,94 | 34,72 | 3,20 | 44,77 | 6,15 | 3,02 | **0,12** | **0,10** |
| y23 | 0,94 | 9,94 | 2,70 | 5,32 | 6,76 | 34,73 | 3,18 | 44,78 | 5,94 | 2,97 | **0,14** | **0,10** |
| y31 | 0,95 | 8,35 | 2,37 | 4,04 | 6,25 | 31,09 | 3,34 | 38,43 | 4,89 | 2,60 | **0,50** | **0,32** |
| y32 | 0,96 | 8,43 | 2,39 | 4,02 | 6,24 | 31,22 | 3,37 | 38,57 | 4,86 | 2,59 | **0,50** | **0,31** |
| y33 | 0,98 | 8,56 | 2,42 | 4,11 | 6,37 | 31,50 | 3,41 | 38,85 | 4,99 | 2,62 | **0,46** | **0,29** |

Cette zone comprend les départements **27**, **76**,59,**62**,2,60,80.
Elle est incluse dans l'ensemble formé par les classes 01, 02 et 04.

**ESTIMATION EN ZONE 3 ( Ouest)**

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR | DISTANCE | DISTANCE 2 hors ralloc rinval et cotor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ESTIMATION DIRECTE** | | | | | | | | | | | | |
| estim. directe simple | 0,75 | 10,55 | 2,93 | 4,89 | 7,47 | 35,53 | 4,13 | 43,08 | 5,63 | 1,50 | | |
| estim. directe après calage sur la zone VQS | 0,75 | 10,55 | 2,94 | 4,71 | 7,15 | 35,14 | 4,06 | 42,93 | 5,52 | 1,62 | | |
| estim. directe après calage sur la zone VQS +RP99 **(y0)** | **0,75** | **10,32** | **2,92** | **4,96** | **7,18** | **34,56** | **3,86** | **42,53** | **5,52** | **1,73** | | |
| **ESTIMATION INDIRECTE** | | | | | | | | | | | | |
| estim. indirecte simple | 1,08 | 9,11 | 2,52 | 4,04 | 6,20 | 32,23 | 3,59 | 39,59 | 4,70 | 2,54 | | |
| estim. indirecte après calage sur la zone RP99 **(y1)** | **1,11** | **9,50** | **2,75** | **4,59** | **6,79** | **32,74** | **3,79** | **40,66** | **5,37** | **2,40** | **0,63** | **0,62** |
| y21 | 0,81 | 10,85 | 3,07 | 4,97 | 7,54 | 35,98 | 4,29 | 43,49 | 5,63 | 1,40 | **0,26** | **0,25** |
| y22 | 0,82 | 10,74 | 3,03 | 4,88 | 7,42 | 35,70 | 4,28 | 43,19 | 5,51 | 1,39 | **0,25** | **0,25** |
| y23 | 0,77 | 10,56 | 2,91 | 4,96 | 7,52 | 35,71 | 4,16 | 43,22 | 5,62 | 1,39 | **0,22** | **0,22** |
| y31 | 1,04 | 8,81 | 2,56 | 4,02 | 6,19 | 31,70 | 3,54 | 39,06 | 4,68 | 2,44 | **0,67** | **0,61** |
| y32 | 1,06 | 8,88 | 2,58 | 4,00 | 6,18 | 31,83 | 3,57 | 39,20 | 4,65 | 2,43 | **0,67** | **0,61** |
| y33 | 1,07 | 9,01 | 2,61 | 4,10 | 6,31 | 32,11 | 3,61 | 39,48 | 4,78 | 2,46 | **0,67** | **0,62** |

Cette zone comprend les départements
14,50,61,22,29,**35**,56,44,49,53,72,85,16,17,79,86.
Elle est incluse dans l'ensemble formé par les classes 01, 02 et 04.

**ESTIMATION EN ZONE 4 (Sud Ouest)**

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR | DISTANCE | DISTANCE 2 hors ralloc rinval et cotor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ESTIMATION DIRECTE** | | | | | | | | | | | | |
| estim. directe simple | 1,13 | 8,34 | 1,95 | 3,07 | 6,07 | 30,27 | 3,13 | 36,04 | 5,21 | 4,00 | | |
| estim. directe après calage sur la zone VQS | 1,18 | 8,43 | 2,10 | 3,53 | 6,73 | 34,47 | 3,31 | 39,98 | 5,47 | 3,29 | | |
| estim. directe après calage sur la zone VQS +RP99**(y0)** | **1,27** | **9,11** | **2,45** | **4,64** | **7,75** | **36,95** | **3,75** | **42,56** | **6,86** | **3,62** | | |
| **ESTIMATION INDIRECTE** | | | | | | | | | | | | |
| estim. indirecte simple | 1,36 | 10,48 | 3,02 | 4,51 | 6,96 | 34,09 | 4,28 | 41,41 | 5,37 | 2,64 | | |
| estim. indirecte après calage sur la zone RP99 **(y1)** | **1,33** | **10,46** | **3,12** | **4,89** | **7,36** | **34,24** | **4,29** | **41,65** | **5,75** | **2,40** | **0,52** | **0,49** |
| y21 | 1,47 | 10,02 | 2,57 | 3,62 | 6,90 | 32,58 | 4,01 | 38,26 | 5,92 | 4,00 | **0,40** | **0,28** |
| y22 | 1,56 | 10,66 | 2,75 | 3,92 | 7,34 | 33,96 | 4,30 | 39,63 | 6,25 | 4,03 | **0,42** | **0,38** |
| y23 | 1,24 | 9,22 | 2,22 | 3,50 | 6,67 | 31,70 | 3,57 | 37,48 | 5,80 | 4,10 | **0,41** | **0,25** |
| y31 | 1,36 | 10,48 | 3,02 | 4,51 | 6,96 | 34,09 | 4,28 | 41,41 | 5,37 | 2,64 | **0,49** | **0,42** |
| y32 | 1,36 | 10,45 | 3,01 | 4,45 | 6,89 | 34,00 | 4,27 | 41,32 | 5,30 | 2,63 | **0,50** | **0,42** |
| y33 | 1,36 | 10,48 | 3,02 | 4,51 | 6,96 | 34,09 | 4,28 | 41,41 | 5,37 | 2,64 | **0,49** | **0,42** |

Cette zone comprend les départements 24,33,40,47,64,9,12,31,32,46,65,81,82.
Elle est incluse dans l'ensemble formé par les classes 01, 02, 03 et 04 (= l'ensemble de la France).

**ESTIMATION EN ZONE 5 (Centre)**

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR | DISTANCE | DISTANCE 2 hors ralloc rinval et cotor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ESTIMATION DIRECTE** | | | | | | | | | | | | |
| estim. directe simple | 1,25 | 9,22 | 2,54 | 3,95 | 5,66 | 28,83 | 3,38 | 40,95 | 4,15 | 1,18 | | |
| estim. directe après calage sur la zone VQS | 1,50 | 10,46 | 3,10 | 4,49 | 6,50 | 31,69 | 3,96 | 43,46 | 4,66 | 1,14 | | |
| estim. directe après calage sur la zone VQS +RP99 **(y0)** | **1,70** | **11,03** | **3,31** | **4,79** | **6,95** | **33,38** | **4,31** | **44,99** | **5,08** | **1,21** | | |
| **ESTIMATION INDIRECTE** | | | | | | | | | | | | |
| estim. indirecte simple | 1,02 | 8,98 | 2,45 | 4,15 | 6,27 | 32,36 | 3,47 | 39,78 | 4,72 | 2,60 | | |
| estim. indirecte après calage sur la zone RP99 **(y1)** | **1,20** | **10,26** | **2,95** | **5,11** | **7,45** | **34,35** | **4,07** | **42,45** | **5,87** | **2,38** | **1,03** | **1,02** |
| y21 | 1,26 | 9,40 | 2,60 | 4,14 | 5,80 | 29,40 | 3,43 | 41,54 | 4,19 | 1,14 | **0,53** | **0,45** |
| y22 | 1,40 | 10,54 | 3,01 | 4,71 | 6,60 | 31,69 | 3,87 | 43,56 | 4,87 | 1,22 | **0,24** | **0,24** |
| y23 | 1,39 | 10,41 | 2,94 | 4,72 | 6,60 | 31,45 | 3,82 | 43,35 | 4,91 | 1,23 | **0,27** | **0,26** |
| y31 | 1,02 | 8,98 | 2,45 | 4,15 | 6,27 | 32,36 | 3,47 | 39,78 | 4,72 | 2,60 | **1,28** | **1,27** |
| y32 | 1,03 | 8,94 | 2,44 | 4,09 | 6,19 | 32,27 | 3,46 | 39,69 | 4,64 | 2,58 | **1,28** | **1,26** |
| y33 | 1,02 | 8,98 | 2,45 | 4,15 | 6,27 | 32,36 | 3,47 | 39,78 | 4,72 | 2,60 | **1,28** | **1,27** |

Cette zone comprend les départements 3,15,43,63,21,58,71,89,18,28,36,37,41,45,19,23,87.
Elle est incluse dans l'ensemble formé par les classes 01, 02, 03 et 04 (= l'ensemble de la France).

**ESTIMATION EN ZONE 6 (Sud Est)**

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR | DISTANCE | DISTANCE 2 hors ralloc rinval et cotor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ESTIMATION DIRECTE** | | | | | | | | | | | | |
| estim. directe simple | 1,77 | 11,76 | 2,52 | 4,35 | 8,84 | 34,22 | 4,65 | 42,12 | 5,57 | 2,38 | | |
| estim. directe après calage sur la zone VQS | 1,35 | 9,49 | 1,99 | 3,37 | 7,28 | 30,75 | 3,61 | 38,67 | 4,43 | 1,95 | | |
| estim. directe après calage sur la zone VQS +RP99 **(y0)** | **1,35** | **9,73** | **2,08** | **3,93** | **7,87** | **32,52** | **3,80** | **40,59** | **5,12** | **2,44** | | |
| **ESTIMATION INDIRECTE** | | | | | | | | | | | | |
| estim. indirecte simple | 1,21 | 9,85 | 2,71 | 4,36 | 6,71 | 33,59 | 3,87 | 41,02 | 5,21 | 2,69 | | |
| estim. indirecte après calage sur la zone RP99 **(y1)** | **1,24** | **10,19** | **2,86** | **4,99** | **7,43** | **34,50** | **4,02** | **41,83** | **5,94** | **2,73** | **0,52** | **0,42** |
| y21 | 1,97 | 12,80 | 2,84 | 4,76 | 9,42 | 36,03 | 5,08 | 43,95 | 6,05 | 2,43 | **0,83** | **0,76** |
| y22 | 1,60 | 10,52 | 2,33 | 3,74 | 7,99 | 32,47 | 4,10 | 40,03 | 4,80 | 1,95 | **0,33** | **0,32** |
| y23 | 1,42 | 9,55 | 2,08 | 3,38 | 7,47 | 30,89 | 3,67 | 38,35 | 4,37 | 1,76 | **0,36** | **0,30** |
| y31 | 1,30 | 10,03 | 2,87 | 4,28 | 6,67 | 35,81 | 3,96 | 42,56 | 5,29 | 2,87 | **0,47** | **0,44** |
| y32 | 1,25 | 9,63 | 2,75 | 4,03 | 6,32 | 35,13 | 3,81 | 41,85 | 4,97 | 2,80 | **0,43** | **0,37** |
| y33 | 1,23 | 9,49 | 2,71 | 3,98 | 6,24 | 34,90 | 3,75 | 41,61 | 4,90 | 2,77 | **0,41** | **0,36** |

Cette zone comprend les départements 20,11,30,**34**,48,66,4,5,6,**13**,83,84.
Elle est incluse dans l'ensemble formé par les classes 01, 03 et 04.

**ESTIMATION EN ZONE 7 (Est)**

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR | DISTANCE | DISTANCE 2 hors ralloc rinval et cotor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ESTIMATION DIRECTE** | | | | | | | | | | | | |
| estim. directe simple | 0,63 | 6,33 | 1,74 | 3,02 | 4,69 | 29,13 | 2,76 | 33,72 | 4,62 | 2,61 | | |
| estim. directe après calage sur la zone VQS | 0,60 | 6,00 | 1,66 | 2,94 | 4,53 | 29,24 | 2,49 | 33,76 | 4,50 | 2,40 | | |
| estim. directe après calage sur la zone VQS +RP99 **(y0)** | **0,65** | **6,73** | **1,85** | **3,45** | **5,04** | **28,02** | **2,60** | **32,74** | **5,12** | **2,39** | | |
| **ESTIMATION INDIRECTE** | | | | | | | | | | | | |
| estim. indirecte simple | 0,82 | 7,76 | 2,03 | 3,78 | 5,73 | 30,26 | 2,93 | 37,66 | 4,43 | 2,56 | | |
| estim. indirecte après calage sur la zone RP99 **(y1)** | **0,90** | **8,15** | **2,23** | **3,95** | **5,94** | **30,75** | **3,13** | **38,41** | **4,63** | **2,48** | **0,61** | **0,56** |
| y21 | 0,44 | 5,28 | 1,37 | 2,84 | 4,29 | 27,61 | 2,27 | 32,19 | 4,35 | 2,53 | **0,56** | **0,49** |
| y22 | 0,45 | 5,24 | 1,38 | 2,62 | 4,07 | 27,37 | 2,29 | 31,93 | 4,11 | 2,49 | **0,59** | **0,47** |
| y23 | 0,62 | 6,25 | 1,74 | 2,90 | 4,62 | 29,00 | 2,73 | 33,60 | 4,58 | 2,60 | **0,26** | **0,15** |
| y31 | 0,78 | 7,46 | 2,06 | 3,76 | 5,73 | 29,73 | 2,88 | 37,13 | 4,42 | 2,46 | **0,38** | **0,32** |
| y32 | 0,80 | 7,53 | 2,09 | 3,74 | 5,71 | 29,87 | 2,91 | 37,26 | 4,39 | 2,45 | **0,41** | **0,35** |
| y33 | 0,81 | 7,66 | 2,12 | 3,83 | 5,84 | 30,15 | 2,95 | 37,55 | 4,51 | 2,48 | **0,45** | **0,38** |

Cette zone comprend les départements 25,39,70,90,1,7,26,38,**42**,69,73,74.
Elle est incluse dans l'ensemble formé par les classes 01, 02 et 04.

**ESTIMATION EN ZONE 8 (Nord Est)**

| | CONFIN | AIDKI | DADAPT | RALLOC | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR | DISTANCE | DISTANCE 2 hors ralloc rinval et cotor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ESTIMATION DIRECTE** | | | | | | | | | | | | |
| estim. directe simple | 1,15 | 9,55 | 3,03 | 4,38 | 6,36 | 29,39 | 4,09 | 36,84 | 3,73 | 1,49 | | |
| estim. directe après calage sur la zone VQS | 1,29 | 9,73 | 3,09 | 4,26 | 6,18 | 29,15 | 4,32 | 36,59 | 4,03 | 1,76 | | |
| estim. directe après calage sur la zone VQS +RP99 **(y0)** | **1,17** | **9,00** | **2,61** | **4,30** | **6,26** | **26,99** | **3,78** | **35,64** | **4,15** | **1,99** | | |
| **ESTIMATION INDIRECTE** | | | | | | | | | | | | |
| estim. indirecte simple | 0,95 | 8,43 | 2,28 | 3,92 | 5,99 | 31,33 | 3,28 | 38,73 | 4,53 | 2,57 | | |
| estim. indirecte après calage sur la zone RP99 **(y1)** | **0,94** | **8,71** | **2,41** | **4,47** | **6,60** | **31,84** | **3,36** | **40,00** | **5,28** | **2,56** | **0,52** | **0,43** |
| y21 | 1,09 | 9,17 | 2,92 | 4,34 | 6,21 | 28,94 | 3,95 | 36,38 | 3,58 | 1,42 | **0,36** | **0,33** |
| y22 | 1,16 | 9,52 | 3,05 | 4,59 | 6,54 | 29,52 | 4,11 | 36,92 | 3,90 | 1,48 | **0,35** | **0,34** |
| y23 | 1,25 | 9,95 | 3,20 | 4,74 | 6,73 | 30,21 | 4,29 | 37,61 | 4,06 | 1,53 | **0,42** | **0,40** |
| y31 | 0,95 | 8,43 | 2,28 | 3,92 | 5,99 | 31,33 | 3,28 | 38,73 | 4,53 | 2,57 | **0,46** | **0,44** |
| y32 | 0,95 | 8,40 | 2,28 | 3,86 | 5,91 | 31,24 | 3,27 | 38,63 | 4,45 | 2,55 | **0,45** | **0,43** |
| y33 | 0,95 | 8,43 | 2,28 | 3,92 | 5,99 | 31,33 | 3,28 | 38,73 | 4,53 | 2,57 | **0,46** | **0,44** |

Cette zone comprend les départements 67,68,8,10,51,52,54,55,57,88.
Elle est incluse dans l'ensemble formé par les classes 01, 02, 03 et 04 (= l'ensemble de la France).

# *ANNEXE XIX*

## IMPORTANCE DES POIDS NEGATIFS

*(cas du département de l'Hérault)*

**Univariate Procedure**

Variable=W (associée à l'estimateur yd33)

Moments

| | | | |
|---|---|---|---|
| N | 16945 | Sum Wgts | 16945 |
| Mean | 51.70009 | Sum | 876058 |
| Std Dev | 204.2309 | Variance | 41710.26 |
| Skewness | 5.371683 | Kurtosis | 28.55613 |
| USS | 7.5203E8 | CSS | 7.0674E8 |
| CV | 395.0301 | Std Mean | 1.568919 |
| T:Mean=0 | 32.95267 | Pr>\|T\| | 0.0001 |
| Num ^= 0 | 16945 | Num > 0 | 13397 |
| M(Sign) | 4924.5 | Pr>=\|M\| | 0.0001 |
| Sgn Rank | 53681950 | Pr>=\|S\| | 0.0001 |

Quantiles(Def=5)

| | | | |
|---|---|---|---|
| 100% Max | 2010.261 | 99% | 1222.382 |
| 75% Q3 | 23.1281 | 95% | 133.3633 |
| 50% Med | 4.052847 | 90% | 54.10696 |
| 25% Q1 | 0.207057 | 10% | -1.40803 |
| 0% Min | -52.6395 | 5% | -4.18585 |
| | | 1% | -30.1648 |
| Range | 2062.901 | | |
| Q3-Q1 | 22.92104 | | |
| Mode | 3.977911 | | |

Extremes

| Lowest | Obs | Highest | Obs |
|---|---|---|---|
| -52.6395( | 222) | 1687.08( | 626) |
| -47.2888( | 376) | 1809.703( | 1382) |
| -46.059( | 230) | 1870.621( | 1374) |
| -45.4194( | 599) | 1890.453( | 1347) |
| -44.2774( | 623) | 2010.261( | 1352) |

## Contribution des départements enquêtés à l'estimation du handicap dans l'Hérault estimateur yd33

| Provenance | _TYPE_ | _FREQ_ | CONFIN | AIDKI | DADAPT | RALLOC |
|---|---|---|---|---|---|---|
| ensemble | 0 | 16945 | 10897.73 | 77341.05 | 23528.69 | 34302.53 |
| dep de la classe 1 | 1 | 8323 | 10030.26 | 73162.27 | 22226.93 | 33606.42 |
| autres dep | 1 | 8622 | 867.48 | 4178.77 | 1301.75 | 696.11 |

| Provenance | RINVAL | HANDI | MOB | DEFI | COTOR | EXPR |
|---|---|---|---|---|---|---|
| ensemble | 49519.44 | 296285.15 | 29628.89 | 347158.16 | 31394.91 | 22342.08 |
| dep de la classe 1 | 48842.38 | 289892.78 | 27790.75 | 340571.48 | 31078.88 | 22585.61 |
| autres dep | 677.05 | 6392.37 | 1838.14 | 6586.68 | 316.03 | -243.53 |

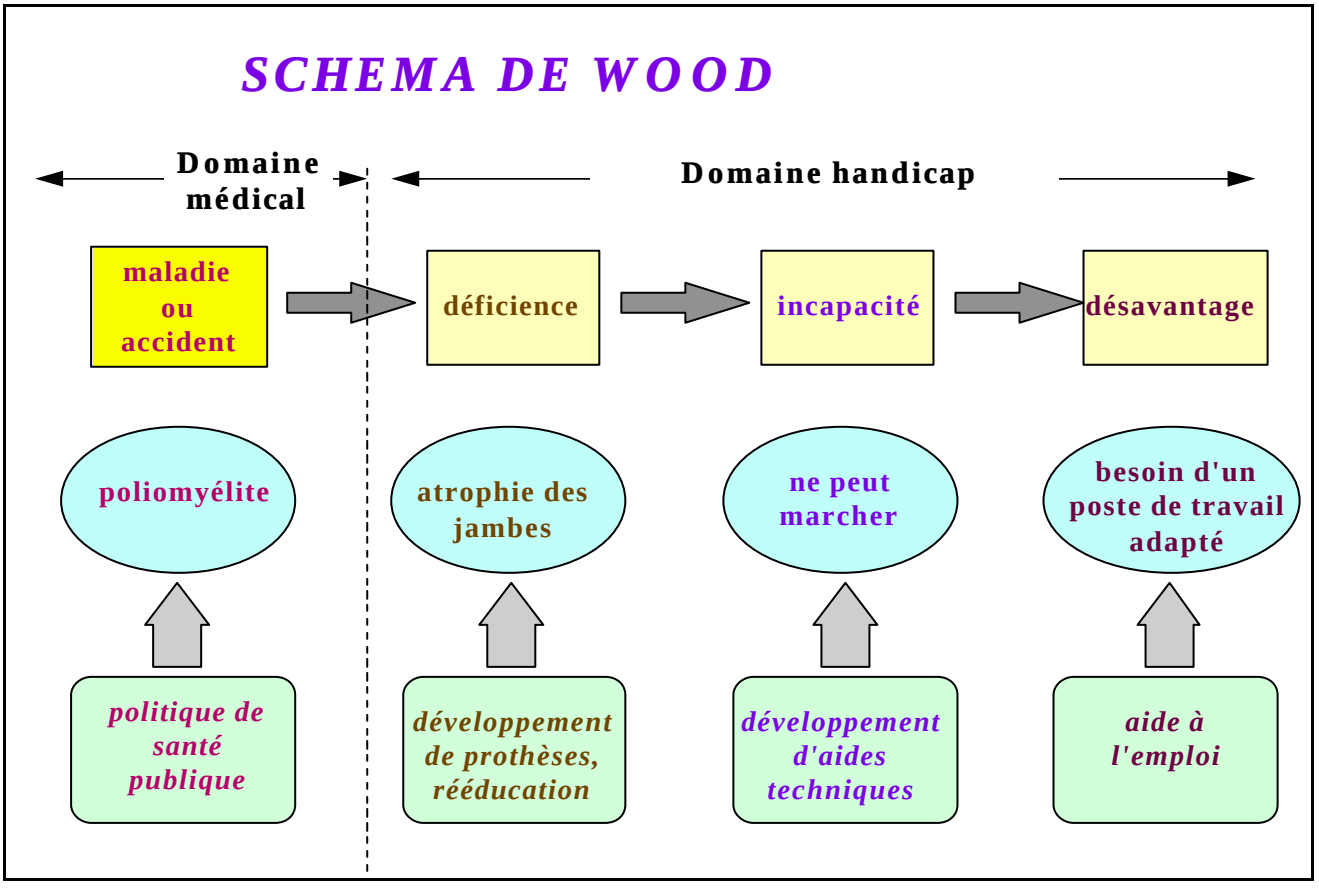
## *ANNEXE XX*

# Thèmes de l'enquête

### I La classification internationale des handicaps, fil conducteur du questionnaire

Le développement des études et statistiques dans le domaine des handicaps est encore récent. Ainsi, alors que la CIM, classification internationale des maladies, a vu le jour il y a un siècle et en est à sa dixième révision, ce n'est qu'en 1980 que le Britannique Philipp WOOD a construit pour l'OMS une nomenclature des « déficiences, incapacités, désavantages » (en abrégé CIH : classification internationale des handicaps), adoptée officiellement en mai 1988 par le ministère français chargé des affaires sociales.

Depuis, les spécialistes de ces domaines ont pris l'habitude de représenter les relations entre maladies et handicaps selon le schéma dit «séquence de WOOD », illustré dans le schéma ci-après.



Les **maladies** (au sens large, c'est à dire y compris les accidents et autres traumatismes moraux ou physiques) sont à l'origine de la chaîne. Elles relèvent du diagnostic et des traitements médicaux.

Les **déficiences** sont les pertes (amputations, scléroses...) ou les dysfonctionnements des diverses parties du corps (membres, muscles, organes) ou du cerveau. Elles résultent en général d'une maladie ou d'un traumatisme. Une notion voisine plus couramment utilisée est celle d'invalidité.

Les **incapacités** sont les difficultés ou impossibilités de réaliser des actes élémentaires (physiques comme se tenir debout, se lever, monter un escalier, psychiques comme mémoriser...), ou plus complexes (s'habiller, se servir d'un téléphone, parler avec plusieurs personnes...). Elles résultent en général d'une ou plusieurs déficiences.

Les **désavantages**, terme préféré à handicaps par les spécialistes francophones, désignent les difficultés ou impossibilités que rencontre une personne à remplir les rôles sociaux auxquels elle peut aspirer ou que la société attend d'elle : suivre les cours scolaires, accomplir un travail, communiquer avec ses semblables, remplir un rôle parental... Ils se situent à la croisée de l'environnement naturel ou social et des caractéristiques propres de l'individu. Pour prendre un exemple, une personne en fauteuil roulant pourra ne pas être désavantagée dans le domaine de l'emploi si les transports pour se rendre à son lieu de travail lui sont accessibles sans difficulté et si son poste de travail, aménagé, ne comporte pas d'obstacle particulier. On comprend ainsi que le désavantage dans un domaine donné, ici l'emploi, est le produit d'une situation personnelle, la paralysie des membres inférieurs, et de conditions environnementales, l'aménagement des transports et du poste de travail.

Cette façon de décrire les problèmes est intéressante pour la politique sociale et de santé, car elle montre qu'on dispose d'une batterie d'actions possibles pour réduire le handicap :

- la recherche et les soins médicaux pour guérir ou prévenir la maladie ;

- la mise au point et à disposition de prothèses pour réduire une déficience ;

- la diffusion d'aides techniques ou l'apport d'aide humaine pour la réalisation des tâches quotidiennes (une baignoire adaptée aux difficultés d'une personne âgée, une aide-soignante pour la toilette matinale...) ;

- une action environnementale (aménager les rues, les transports, les postes de travail...).

## II Traitement des différents thèmes dans l'enquête nationale

Le questionnement HID couvre les trois dimensions du handicap - déficience, incapacité et désavantage - dont les définitions viennent d'être présentées.

**Les déficiences** relèvent d'un concept peu utilisé parmi les non-spécialistes. Ses frontières avec les champs voisins (pathologies d'un côté et incapacités de l'autre) sont difficiles à faire percevoir. Enfin ses limites (à partir de quand peut-on dire qu'il y a une déficience - la myopie

plus ou moins bien corrigée en est-elle une ? un mal de dos chronique mais supportable est-il une déficience ?) sont floues. Aussi leur relevé a-t-il été réalisé en trois étapes :

1. une question introductive ("Rencontrez-vous dans la vie de tous les jours des difficultés, qu'elles soient physiques, sensorielles, intellectuelles ou mentales - dues aux conséquences d'un accident, d'une maladie chronique, d'un problème de naissance, d'une infirmité, du vieillissement...) ouvre un relevé en clair de la nature et de l'origine de ces éventuelles "difficultés", qui constituent le plus souvent des déficiences ;

2. par la suite, dans le chapitre sur les incapacités, chaque incapacité relevée donne lieu à un questionnement sur ses causes. Cette procédure a permis de "récupérer" environ 6 600 oublis s'ajoutant aux 26 500 déclarées initialement.

3. enfin, les fichiers ont été ensuite repris par une équipe médicale constituée à cet effet, qui a mis en pratique le contrôle, le recodage, le complément des déficiences. Sur les 33 167 déficiences relevées par les enquêteurs pour les 16 924 personnes de l'échantillon ménage, un peu moins de 3 000 ont été supprimées, mais 7242 nouvelles ont été créées.

On reviendra sur l'interprétation de ces éléments et les précautions à prendre dans l'exploitation des données sur les déficiences.

**Les incapacités** ont été délibérément placées au centre de la procédure d'enquête. C'est en effet la dimension du handicap la plus commodément accessible à un questionnaire par interview : la plupart des questions classiquement posées dans ce domaine font référence à des actes concrets et précis de la vie quotidienne, parfois à de simples gestes, et les réponses en paraissent mieux assurées. Pour la même raison sans doute, la plupart des instruments d'évaluation de la sévérité du handicap sont fondés sur une synthèse de questions relatives aux incapacités.

**Le désavantage** pose un problème différent. Il relève en effet d'un point de vue comparatif : dans le domaine de l'emploi par exemple, il ne suffit pas d'établir que x % des hommes aveugles de 30 à 35 ans ne trouvent pas d'emploi pour avoir repéré ou mesuré un quelconque désavantage. Il est nécessaire de comparer ce taux à celui affectant la moyenne des hommes du même âge (tel qu'il ressort des enquêtes sur l'emploi).

Pour faciliter la comparaison, on a inclus au questionnaire de chaque domaine abordé - logement, transports, scolarité, emploi, revenus, loisirs et contacts - des questions qui reproduisent aussi fidèlement que possible, celles tirées de l'enquête correspondante de l'Insee.

Il est rempli avec le pensionnaire. Mais si la personne sélectionnée pour l'enquête est inapte à y répondre, elle peut se faire assister ou suppléer par un parent voire un éventuel tuteur s'il y a lieu.

# SOMMAIRE

**PRÉSENTATION DES TRAVAUX DU GROUPE « ESTIMATIONS LOCALES » DANS LE CADRE DE L'ENQUÊTE HID ..... 2**

1 Les estimations locales, un sous-produit de l'enquête nationale ..... 2
2 Les éléments disponibles pour produire des estimations locales ..... 3
3 Méthodologie des « petits domaines » ..... 4
*3.1 Le modèle d'estimation sur « petits domaines »* ..... *5*
*3.2 La formalisation du modèle* ..... *6*
*3.3 Choix des critères de post-stratification* ..... *7*
*3.4 Réduction du nombre de post-strates* ..... *8*
*3.5 Modification des poids individuels du fichier national* ..... *9*
*3.6 Précision des résultats* ..... *10*
*3.7 Tests de validité du modèle* ..... *10*
*3.8 Calages ultimes* ..... *11*
4 Tentatives visant à améliorer le modèle ..... 12
*4.1 Une seule méthode de calcul quelle que soit la variable étudiée* ..... *12*
*4.2 Intégrer ou pas une composante locale résiduelle* ..... *13*
5 Précautions particulières propres aux exploitations locales ..... 14

**ANNEXES ..... 19**

Annexe I : L'architecture d'ensemble de l'enquête HID ..... 19
Annexe II : Tirage et pondération de l'échantillon national ..... 21
Annexe III : Bibliographie ..... 24
Annexe IV : Propositions d'estimateurs pour l'enquête HID ..... 25
Annexe V : Choix d'un modèle de comportement ..... 31
Annexe VI : Définition des post-strates ..... 44
Annexe VII : Estimation de la variance de l'estimateur régional ..... 47
Annexe VIII-a : Premiers tests du modèle de comportement ..... 53
Annexe VIII-b : Test du modèle de comportement dans le département de l'Hérault ..... 62
Annexe IX : Calage sur les marges socio-démographiques du RP 99 dans l'Hérault ..... 66
Annexe X : Estimateurs de type « petits domaines » sous un modèle à un ou deux facteurs ..... 69
Annexe XI : Les nouvelles pondérations du fichier national correspondant à l'estimateur post-stratifié départemental à deux facteurs ..... 72
Annexe XII : Résultats des quatre estimateurs de type « petits domaines » dans le département de l'Hérault ..... 77
Annexe XIII : Réflexions du groupe de travail à propos de la stabilité de la mesure de l'effet départemental « résiduel » ..... 84

Annexe XIV : Contrôle de la stabilité départementale de la mesure de l'effet « résiduel » dans le modèle de comportement à deux facteurs (Application de la méthode de « Boot-Strap ») ..... 87
Annexe XV : Les estimateurs combinés associés à une classification départementale ..... 93
Annexe XVI : Nouvelles pondérations du fichier HID national associées à un estimateur local à deux facteurs comprenant un effet résiduel mesuré sur une classe de départements ..... 99
Annexe XVII : Bilan des résultats du modèle combiné à deux facteurs ..... 100
Annexe XVIII : Estimations locales sur des zones géographiques plus vastes que le département ..... 103
Annexe XIX : Importance des poids négatifs ..... 109
Annexe XX : Thèmes de l'enquête ..... 110