LA RÉALISATION D'ESTIMATIONS LOCALES DANS LE CADRE DE L'ENQUÊTE HID

Christine COUET, Pierre MORMICHE

Insee, Département de la Démographie

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 5).

Toutefois ce travail ne concerne qu'un des volets de l'enquête qui au total en compte quatre[1]. 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[2]

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.

[1] **Cf, Annexe 1** : L'architecture d'ensemble de l'enquête HID, [1] Document de travail n° F0207, INSEE.
[2] – sont décrites en **Annexe 2** : Tirage et pondération de l'échantillon national du document de travail n° F0207, INSEE.

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[3].

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 la construction de modèles [4].

C'est de cette deuxième catégorie d'estimateurs que se sont inspirés les travaux du groupe[5] 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 listés au § 2 ci-dessus) pour lesquelles on disposait d'extension de l'enquête VQS.

[3] 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.
[4] Voir la liste d'articles traitant de méthodes d'estimation sur petits domaines dans l'**annexe** bibliographique **3**, [1] Document de travail n° F0207, INSEE.
[5] 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).

3.1. 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[6] 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 »[7] - 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. Conformément à cette observation, 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. Comme c'est sur l'échantillon le plus vaste, l'échantillon national, que la mesure de ces prévalences est la plus fiable, on utilise les résultats nationaux par post-strate pour estimer les prévalences locales.

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 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 : le « groupe VQS ». On mesure l'importance qu'a revêtu pour le travail d'eestimations 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[8]. 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 :

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

[6] En épidémiologie, la prévalence désigne simplement la proportion de sujets qui dans une population donnée souffre de telle pathologie ou de tel handicap.
[7] Appelées ainsi car elles sont définies après la réalisation du plan de sondage.
[8] Cf, **Annexe 4** : Propositions d'estimateurs pour l'enquête HID, L.Wilms (UMS), [1] Document de travail n° F0207, INSEE.

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

$$\hat{\bar{Y}}_h = \frac{\sum_{k \in s_{HID} \cap postrateh} \frac{y_k}{p_k}}{\sum_{k \in s_{HID} \cap postrateh} \frac{1}{p_k}}$$

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 :

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

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) [9].

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 ;

[9] Un résumé de ces travaux est présenté en **Annexe 5** : Choix d'un modèle de comportement (extrait du rapport de Valérie Albouy), [1] Document de travail n° F0207, INSEE.

- 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[10] (**cf. annexe 6**).

Les regroupements sont surtout réalisés chez les individus à faible risque : essentiellement les populations jeunes n'appartenant pas au groupe 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 d'intérêt exploitables 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.

[10] Cf **Annexe 6** : Définition des post-strates, [1] Document de travail n° F0207, INSEE.

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[11,12]. 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 ».

[11] Les détails de la démarche sont dans le rapport de stage de Valérie Albouy – ENSAE, été 2000.
[12] Cf **Annexe 7** : Choix d'un modèle de comportement (extrait du rapport de Valérie Albouy), [1] Document de travail n° F0207, INSEE.

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 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[13]. Par la suite, le critère « taille de la commune » en 3 postes a été préféré à la notion de type d'habitat[14].

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.

[13] Cf. les résultats des tests en **annexe 8-a :** Premiers tests du modèle de comportement, [1] Document de travail n° F0207, INSEE.

[14] Cf. les résultats dans l'Hérault en **annexe 8-b :** Test du modèle de comportement dans le département de l'Hérault, [1] Document de travail n° F0207, INSEE.

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[15] 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.

[15] **cf. annexe 9 :** Calage sur les marges socio-démographiques du RP 99 dans l'Hérault, [1] Document de travail n° F0207, INSEE.

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 10** et l'expression des pondérations nationales associées à ce modèle en **annexe 11**. Mais son expérimentation dans le département de **l'Hérault** souligne quelques problèmes spécifiques à cette nouvelle approche (cf. les résultats et commentaires en **annexe 12**).

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 13)** 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 14**).

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 15**).
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 16,** 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 17** et les tests complémentaires, sur des zones plus vastes que le département, en **annexe 18**);
- 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 19**). 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 présentée 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 20)**.

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.

Bibliographie

[1] Couet C., « Estimations locales dans le cadre de l'enquête HID - Démarche suivie », *Document de travail n° F0207, INSEE, Direction des Statistiques Démographiques et Sociales,* novembre 2002.

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

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

• Les contributions des participants à la conférence de Riga : « Small area estimation ».
dont :

[4] Attal-Toubert K., «Local estimations for the "Handicap, Disability and Dependence" survey ».

[5] Christine M., «The French local estimates on labour statistics,based on administrative registers and surveys : current situation and future work ».

[6] Albouy V., « Estimation sur petits domaines : le cas de l'enquête Handicaps, Incapacités, Dépendance », ENSAE, été 2000.

[7] Bourgeois A., Bonnery D., Valentino J., Bunales R., sous la direction de Qualité L., « Calcul de la variance des estimateurs sur petits domaines dans l'enquête HID par une méthode empirique », *Projet Statistique ENSAI 2ème année,* 2002.

[8] Couet C., «Estimations locales sur les personnes handicapées vivant en domicile ordinaire - Enquête HID 1999, Résultats détaillés », *INSEE-Résultats série Société* n° 12, décembre 2002.

Liste des annexes du document de travail référencé dans ce texte

Annexe 1 : L'architecture d'ensemble de l'enquête HID

Annexe 2 : Tirage et pondération de l'échantillon national

Annexe 3 : Annexe bibliographique

Annexe 4 : Propositions d'estimateurs pour l'enquête HID L. Wilms (UMS)

Annexe 5 : Choix d'un modèle de comportement (extrait du rapport de Valérie Albouy)

Annexe 6 : Définition des post-strates

Annexe 7 : Estimation de la variance de l'estimateur régional (extrait du rapport de Valérie Albouy

Annexe 8a : Premiers tests du modèle de comportement

Annexe 8b : Test du modèle de comportement dans le département de l'Hérault

Annexe 9 : Calage sur les marges socio-démographiques du RP 99 dans l'Hérault

Annexe 10 : Estimateur de types petits domaines sous un modèle à 1 ou 2 facteurs (Laurent Wilms – UMS)

Annexe 11 : Les nouvelles pondérations du fichier national correspondant à l'estimateur post-stratifié départemental à deux facteurs

Annexe 12 : Résultats des quatre estimateurs de type «petits domaines »dans le département de l'Hérault

Annexe 13 : Réflexions du groupe de travail à propos de la stabilité de la mesure de l'effet départemental « résiduel »

Annexe 14 : 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 »

Annexe 15 : Les estimateurs combinés associés à une classification départementale

Annexe 16 : Nouvelles pondérations du fichier HID national associées à un estimateur local à deux facteurs comprenant un effet résiduel mesuré sur une classe de départements

Annexe 17 : Bilan des résultats du modèle combiné à deux facteurs (l'exemple de l'Hérault)

Annexe 18 : Estimations locales sur des zones géographiques plus vastes que le département (Tests complémentaires)

Annexe 19 : Importance des poids négatifs (cas du département de l'Hérault)

Annexe 20 : Thèmes de l'enquête HID