Estimations communales exploitant les données de l'enquête *Famille et logements* 2011 et du recensement : une opération à hauts risques

Pascal Ardilly *

À l'été 2013, l'Insee a diffusé à plus de 1 400 communes des estimations d'effectifs pour diverses sous-populations vivant sur leurs territoires : personnes ayant souscrit un pacs, personnes en couple non cohabitant, grands-parents, personnes âgées vivant seules et ayant des enfants résidant à proximité, etc.

Ces estimations communales se sont appuyées sur la collecte nationale de l'enquête *Famille et logements*, enquête de très grande taille associée à l'enquête annuelle de recensement 2011. Afin de leur assurer une qualité suffisante, on a eu recours à une démarche par modélisation, de type « petits domaines ». La première étape s'appuie sur l'ensemble de l'échantillon de l'enquête. Elle consiste à modéliser les probabilités individuelles d'appartenir à la sous-population d'intérêt. Les variables explicatives utilisées sont des variables binaires disponibles dans le recensement : sexe, groupe d'âge, statut matrimonial, etc. Elles délimitent donc des catégories de population au sein desquelles ces probabilités d'appartenir à la sous-population d'intérêt seront considérées homogènes et indépendantes de la commune de résidence. Une fois évaluées ces probabilités, on obtient les estimations communales en les multipliant par les effectifs communaux des catégories associées (fournis par le recensement). Enfin, on procède à un calage sur l'effectif national de la sous-population d'intérêt issu de l'enquête.

Parce qu'elle fonde les estimations sur des échantillons de grande taille, cette procédure réduit beaucoup la variance d'échantillonnage par rapport à une estimation qui utiliserait seulement les informations de l'enquête au niveau communal. En contrepartie, le recours à un modèle génère un biais car on assimile un comportement communal à un comportement supra communal. L'erreur qui en résulte peut être appréciée à un niveau agrégé et on constate, sur les variables diffusées, qu'elle reste le plus souvent très acceptable.

Codes : C42, C51, C52, J12, J13.

Mots clés : estimation sur petits domaines, modélisation des comportements, modèle logistique, appréciation du biais, *shrinkage*, *benchmarking*, pacs, modes de garde des enfants, type de famille.

** Direction de la méthodologie et de la coordination statistique et internationale, Département des méthodes statistiques (Insee).*

L'auteur remercie Nathalie Blanpain, Guillemette Buisson et Aude Lapinte (division Enquêtes et études démographiques, Insee) pour leurs relectures, leurs conseils et l'aide efficace qu'elles ont apportée à l'occasion de ce travail, ainsi que les deux rapporteurs anonymes pour leurs remarques pertinentes et leurs suggestions. L'auteur demeure seul responsable d'éventuelles erreurs ou imprécisions.

Produire des statistiques communales à partir de l'enquête *Famille et logements* : le contexte

L'enquête *Famille et logements* (*EFL*) est une enquête associée à l'*Enquête annuelle de recensement* (*EAR*) de 2011, menée auprès d'individus adultes résidant en ménage ordinaire et en métropole (Breuil-Genier *et al.*, dans ce numéro). Par adulte, on entend les individus ayant 18 ans ou plus au 1er janvier 2011. Elle s'est déroulée dans 1 438 communes volontaires de toutes tailles et se place dans la longue tradition des enquêtes *Famille* dont la dernière édition était adossée au recensement de 1999. Elle a pour objectifs essentiels d'actualiser notre connaissance de la fécondité, de mettre en regard des informations portant sur trois générations familiales (grands-parents, parents, enfants), d'étudier le recours au pacte civil de solidarité, de permettre une meilleure connaissance des familles recomposées, du mode de garde des jeunes enfants ainsi que de la multi-résidence des adultes et des enfants. Pour cela, elle s'appuie sur un échantillon de très grande taille, constitué de 239 000 femmes et 121 000 hommes répondants à un questionnaire confié aux agents recenseurs, donc géré sur le terrain par dépôt-retrait. Le questionnaire est individuel, mais il comprend aussi quelques questions portant sur le niveau logement / ménage.

L'échantillonnage des individus dans *EFL* résulte d'un plan de sondage particulièrement complexe et s'appuie sur une intersection de deux échantillons de logements : d'une part l'échantillon des logements de l'*EAR* de 2011 (Godinot, 2005), d'autre part un échantillon de logements spécifiques à l'enquête *EFL* résultant d'un plan à deux degrés. Ce dernier est constitué par un tirage stratifié de communes (recensées en 2011) dans lesquelles on sélectionne au hasard (par tirage aléatoire simple, donc à probabilités égales) des grappes prédéfinies de logements, lesquelles sont en fait des regroupements de districts contigus, voire d'« îlots regroupés pour l'information statistique » (Iris)[1] contigus lorsqu'il s'agit de grandes communes. Le nombre de grappes tirées dépend de la taille de la commune, mais dans beaucoup de communes on échantillonne une ou deux grappes, ce qui fait que l'enquête concerne le plus souvent entre 200 et 400 logements dans la commune. Seuls les logements abritant des ménages ordinaires et ayant le statut de résidence principale sont pris en compte. Pour obtenir l'échantillon d'individus physiques, l'enquête a la particularité d'associer une modalité de la variable sexe à chaque grappe tirée, selon un processus dit d'enquête en deux phases : l'échantillon national de grappes étant tiré, on sélectionne au hasard, sans se préoccuper de la commune concernée, deux tiers des grappes dans lesquelles on distribue un questionnaire « femme » à toutes les femmes adultes de l'échantillon de logements, le tiers de grappes restant donnant lieu à la remise d'un questionnaire « homme » à tous les hommes adultes de l'échantillon de logements. On ne contrôle donc pas la structure de l'échantillon par sexe à l'intérieur des communes concernées par l'enquête.

Afin de minimiser le nombre de non-réponses résultant des communes réfractaires (l'enquête n'étant pas obligatoire et la coopération des municipalités étant de ce fait indispensable), l'Insee s'était engagé à restituer des estimations « personnalisées » à l'intention de chaque commune ayant accepté de participer à l'enquête. Dans la pratique, l'estimation a été effectuée pour toutes les communes du territoire métropolitain. L'idée était de produire de l'information communale intéressante pour le débat local et articulée autour des principaux thèmes de l'enquête. La situation s'avérait néanmoins très délicate parce que l'enquête *EFL*, à l'image de toutes les grandes enquêtes par sondage si l'on exclut les enquêtes de recensement, n'était pas à l'origine conçue pour produire des statistiques à un niveau aussi fin. Le volume et le choix des statistiques restituées ont relevé de la responsabilité de l'institut, qui a opté pour un document de quatre pages associant des informations exclusivement issues du recensement (*RP*) avec des informations provenant de l'enquête *EFL*. Le choix s'est porté sur des informations pour lesquelles les élus locaux ont exprimé leur intérêt au fil du temps à savoir, dans le cas des informations basées sur *EFL* :

- Le nombre d'hommes, de femmes et de couples ayant souscrit un pacs ;

- Le nombre de grands-pères et de grands-mères, respectivement de moins et de plus de 75 ans ;

- Le nombre d'enfants de moins de 4 ans gardés principalement par les parents, les grands-parents, une assistante maternelle, une crèche, ou un autre mode de garde ;

- Le nombre d'hommes et de femmes vivant en couple mais non cohabitant ;

1. *Un Iris est un groupe d'îlots jugé homogène en type d'habitat.*

- Le nombre d'enfants de moins de 18 ans vivant dans une famille recomposée, traditionnelle, monoparentale, en distinguant trois tranches d'âge ;

- Le nombre de familles recomposées et traditionnelles ;

- Le nombre d'hommes et de femmes de plus de 75 ans vivant seuls et ayant au moins un enfant résidant dans la même commune ;

- Le nombre d'hommes et de femmes de plus de 75 ans vivant seuls et ayant au moins un enfant résidant à moins de 30 minutes de leur domicile ;

- Le nombre d'enfants de moins de 18 ans résidant habituellement dans deux logements pour cause de vie avec un autre parent;

- Le nombre d'hommes et de femmes résidant habituellement dans deux logements.

La méthodologie mise en œuvre

Où est la difficulté ?

Toute enquête par sondage génère une erreur d'échantillonnage, c'est-à-dire que les estimations qu'elle permet de produire sont, sauf miracle, différentes des vraies valeurs qui resteront inconnues. La théorie des sondages montre que l'erreur totale attribuée à l'échantillonnage, dite « erreur quadratique moyenne », est constituée de deux composantes, le biais et la variance. Imaginons l'ensemble (gigantesque) des échantillons que l'on peut former dans la population complète et l'ensemble des estimations qui leur correspondent. Le biais mesure l'écart existant entre la moyenne (pondérée par les probabilités de sélection des échantillons) de toutes ces estimations et la vraie valeur que l'on veut mesurer, tandis que la variance traduit la sensibilité de ces estimations à l'échantillon tiré. Autant on parvient le plus souvent à limiter fortement le biais par une pondération adéquate, autant on doit subir le phénomène de variance. Cette variance est d'autant plus grande que la taille de l'échantillon répondant est petite et c'est pourquoi l'estimation communale *EFL* pose un redoutable problème : à part dans quelques rares cas correspondant aux plus grandes communes, l'échantillon communal *EFL* comprend moins de 500 répondants, qui plus est géographiquement proches (existence d'un effet de grappe pénalisant). Cela ne génère pas de biais, mais s'avère largement insuffisant pour obtenir des estimations dont la variance répond à nos standards de qualité. Il est donc nécessaire d'utiliser une méthode d'estimation qui se démarque de la théorie classique des sondages, laquelle pourrait fréquemment aboutir dans un tel cas de figure à des estimations aberrantes[2].

Quelle stratégie adopter ?

Pour éviter une explosion de la variance, c'est-à-dire une instabilité excessive des estimations, il faut baser celles-ci sur un échantillon de grande taille, qui va donc au-delà du simple échantillon communal, et faire appel à un modèle de comportement. Ce dernier assimilera le comportement moyen de la commune qui nous intéresse à celui qui caractérise d'autres communes de même nature. On pourra ainsi construire l'estimation communale sur un échantillon qui associe plusieurs communes, donc sur un échantillon supra communal.

Le modèle à appliquer dépend de la variable considérée ; il dépend aussi, en général, de la commune en ce sens où on peut s'appuyer sur une typologie de communes préalablement définie et appliquer séparément la procédure pour chaque classe de communes. Dans notre contexte, le modèle relie la probabilité d'appartenir à la sous-population d'intérêt (celle dont on cherche l'effectif) et un ensemble de variables explicatives, relation dans laquelle la géographie n'aurait, par hypothèse (presque) plus d'influence[3] (cf. encadré 1).

Considérons par exemple la sous-population des personnes pacsées. On explique la probabilité d'être pacsé par une batterie de variables individuelles en neutralisant tout effet de la commune (en tout cas, dès lors que la commune appartient à une catégorie de communes donnée). Ces variables explicatives peuvent être par exemple le sexe, l'âge et le niveau de diplôme si on suppose que deux individus appartenant à deux communes différentes mais qui ont le même sexe, le même âge et le même niveau de diplôme ont la même probabilité d'être pacsés.

---

2. On trouvera dans Ardilly (2006) les éléments théoriques éclairant la nature, la mesure et les techniques classiques de réduction de l'erreur quadratique moyenne.

3. Le terme de modèle peut donc être compris de deux façons : soit on raisonne en population finie, et on parle d'égalités, entre communes, de proportions définies par sous-population (approche sondage), soit on se place en population infinie et l'hypothèse prend la forme d'un mécanisme stochastique universel qui relie une probabilité à des variables explicatives individuelles (approche économétrique). Mais tout cela est équivalent, et dans les deux approches le cœur de l'hypothèse est que la localisation à la commune n'apporte pas d'information.

La sanction immédiate de ce type d'hypothèse est l'introduction d'un biais, puisqu'il est difficile d'imaginer qu'il n'existe aucune spécificité locale au-delà de la structure qu'offre la commune en matière de variables explicatives : si le modèle ignore une variable explicative importante et si la commune possède une spécificité selon cette variable, on obtient un estimateur biaisé. Or aucun modèle n'est capable de prendre en compte toute la complexité des phénomènes étudiés. On substitue donc nécessairement du biais à la variance : le pari qui justifie toute cette approche, c'est que le gain de variance surpasse la perte en biais, ce qui permet de réduire l'erreur d'échantillonnage à un point compatible avec nos standards de qualité. En la circonstance, le gain en variance est considérable ; la perte en biais sera quant à elle appréciée variable par variable. Le lecteur trouvera dans Rao (2003) un exposé des principales techniques utilisées pour produire des estimations lorsque les échantillons sont de petite taille.

Les statistiques qui nous intéressent *in fine* sont des effectifs communaux. Dans toute la suite, sauf lorsqu'on dénombre des familles ou des enfants, on ne considère que des adultes (18 ans ou plus au 1er janvier 2011), puisqu'il s'agit du champ couvert par l'*EFL*. Reprenons l'exemple du nombre de pacsés. On partitionne au préalable la population de la commune en sous-populations définies d'après les critères explicatifs retenus à l'issue de la procédure initiale d'ajustement du modèle probabiliste : mettons qu'il s'agisse du sexe, de l'âge et du niveau de diplôme. L'approche en population finie ayant été privilégiée, l'effectif total estimé d'individus pacsés dans la commune s'obtient en multipliant les proportions estimées d'individus pacsés au sein d'une sous-population par les effectifs communaux respectifs de ces sous-populations, puis en sommant ces produits. Le modèle intervient de manière déterminante car il permet, sous-population par sous-population, de remplacer la proportion de personnes pacsées propre à la commune (mais que l'on ne peut pas estimer de manière fiable du fait d'un nombre insuffisant d'observations) par la proportion de personnes pacsées évaluée sur l'ensemble de communes auxquelles cette commune s'assimile. Dans l'expression de l'estimateur communal final, les estimations

Encadré 1

PRINCIPE DU MODÈLE D'ESTIMATION COMMUNALE

On note com_0 la commune dans laquelle on veut obtenir l'estimation d'un effectif $N_{com_0}^{ssp}$ caractérisant une sous-population notée *ssp* (par exemple celle des individus pacsés). Le modèle porte sur la proportion des individus appartenant à cette sous-population. La proportion n'est pas nécessairement calculée sur l'ensemble de la population communale, mais bien souvent sur un champ que l'on notera *ch* (par exemple l'ensemble des individus adultes non mariés et vivant en couple cohabitant). Un ensemble de facteurs explicatifs collectés au recensement permet de prédire l'appartenance à *ssp*. Si on n'ose qualifier de satisfaisant dans l'absolu le pouvoir prédictif de ces variables, du moins doit-on au moins dire qu'il est conçu « au mieux » compte tenu de l'information dont on peut disposer. La prise en compte de ces facteurs définit des catégories de population dont l'indice courant est *h* (par exemple un croisement ad-hoc de modalités associant le sexe, la tranche d'âge et le diplôme). On note :

$N_{com_0}^{ch,h,ssp}$ = l'effectif vrai (inconnu) des résidents de com_0 qui sont dans le champ *ch* et la catégorie *h*, et qui appartiennent à la sous-population d'interêt *ssp*.

$N_{com_0}^{ch,h}$ = l'effectif vrai (et observé) des résidents de la commune com_0 qui sont dans le champ *ch* et appartiennent à la catégorie *h*.

On fait l'hypothèse (c'est le modèle) :

$$\forall h,\ \frac{N_{com_0}^{ch,h,ssp}}{N_{com_0}^{ch,h}} \text{ ne dépend que de } h$$

Cette hypothèse s'interprète ainsi : quel que soit l'individu du champ *ch* appartenant à la catégorie *h*, sa probabilité de vérifier le critère *ssp* ne dépend pas de la commune. En pratique, on définit en sus des catégories de communes et on applique ce modèle au sein de chaque catégorie de communes, si bien qu'il faut adapter l'interprétation en disant que ce ratio ne dépend que de l'appartenance à une catégorie de communes donnée. Cela permet d'une certaine façon de redonner un peu d'influence à la composante géographique. De fait, quelle que soit la catégorie *h* donnée, toutes ces proportions sont égales au ratio

$$\frac{\sum_{com} N_{com}^{ch,h,ssp}}{\sum_{com} N_{com}^{ch,h}}$$

où les sommations sont étendues à l'ensemble des communes participant à l'enquête *EFL* (ou selon les circonstances à l'ensemble des communes d'une catégorie de communes donnée).

de chacune de ces proportions s'appuieront sur un système de pondérations adapté, reflétant l'échantillonnage et impliquant un grand nombre d'individus répondants.

D'autres pays ont mis en place des opérations d'estimation communale qui, par certains aspects, sont proches de l'enquête *EFL*. Il s'agit par exemple de l'*Enquête nationale auprès des ménages* (*ENM*) au Canada et de l'*American Community Survey* (ACS) aux États-Unis. Dans les deux cas il s'agit de compléter, au moyen d'une enquête par sondage, une information provenant d'un recensement exhaustif [4]. Les niveaux de diffusion sont très localisés (la commune, voire plus fin) et des problèmes de taille d'échantillon insuffisante se posent. En revanche, contrairement à la situation de l'enquête *EFL*, il s'agit d'opérations annuelles régulières totalement intégrées au processus de recensement et s'appuyant sur un échantillon national beaucoup plus gros. Par ailleurs, l'estimation communale s'effectue en utilisant une technique de calage, sans faire appel à un modèle.

Application du modèle

Le modèle porte sur des vraies probabilités, ou plutôt de vraies proportions (proportions de pacsés par exemple) puisqu'on a à faire à des populations finies. Procéder aux calculs nécessite d'effectuer des estimations de toutes les grandeurs inconnues qui entrent dans la composition de ces proportions. Pour estimer les probabilités d'appartenir à la sous-population dont on cherche l'effectif, il faut entre autres utiliser la pondération de l'*EAR* 2011 puisqu'on rappelle que l'échantillon d'individus *EFL* est un sous-échantillon de l'échantillon des individus recensés par l'*EAR*.

À titre préliminaire, il est important de préciser qu'il n'était pas possible d'utiliser les éléments de pondération (poids d'échantillonnage comme probabilités de non-réponse) provenant du traitement de l'enquête nationale : en effet, la stratégie suivie consistait à effectuer la restitution des estimations communales dans la foulée de la diffusion des premiers résultats nationaux. Or le traitement de l'enquête nationale a naturellement débuté par une longue phase d'apurement : le processus d'estimation locale a donc toujours eu de l'avance sur celui de l'estimation nationale – ce qui n'a pas été sans poser problème. Un argument complémentaire, plus philosophique, vient conforter la technique d'estimation décrite infra : toute la procédure se justifie conditionnellement aux communes tirées, et cela dès la constitution du modèle (cf. encadré 1). Cette situation est très originale parce qu'il est rare que la réunion des domaines d'intérêt ne couvre pas la population complète. Dans ces conditions, puisque les proportions à estimer quand on applique le modèle ne font jamais intervenir le niveau national mais seulement des territoires géographiques sans interprétation, le premier degré de sondage conduisant à l'échantillonnage des communes n'a pas à être pris en compte et la pondération nationale ne trouve pas sa place dans les estimateurs locaux (cf. encadré 2).

Notre contexte est aussi compliqué par le phénomène de non-réponse propre à l'enquête *EFL*. La non-réponse affecte aussi les *EAR* bien entendu, mais l'Insee a mis en place un système d'imputation en amont qui traite la non-réponse totale aussi bien que la non-réponse partielle. De fait, l'utilisateur des fichiers du recensement ne voit jamais apparaître de repondération pour cause de non-réponse. En revanche, il n'y a pas d'imputation dans *EFL* pour cause de non-réponse totale[5] et il est donc indispensable de repondérer les répondants à *EFL* afin qu'ils représentent les non-répondants. Traditionnellement, on recherche des variables au niveau ménage et/ou au niveau individuel qui expliquent la probabilité de réponse. Presque systématiquement, on trouve par exemple que la taille du ménage est une variable significative. En la circonstance, le contexte est assez différent des enquêtes par sondage habituelles parce que le mode de collecte est un mode par dépôt-retrait, ce qui est original. En particulier on imagine que l'effet de la taille du ménage, s'il existe, doit être très atténué lorsqu'on utilise un tel mode de collecte. De plus, les non-répondants à l'*EAR* sont également non-répondants à *EFL* sauf exception, ce qui rend quasiment impossible l'exploitation de variables explicatives « exactes » de la probabilité de réponse qui soient liées au ménage (*a fortiori* aux individus) puisqu'il faut disposer de ces variables pour les non-répondants à *EFL* (l'imputation préalable de la non-réponse au recensement risquant de brouiller les phénomènes explicatifs, s'il y en a).

4. L'information recensée est très simple (questionnaire dit « court »). Ces enquêtes ont remplacé un questionnaire plus complet dit « questionnaire long » qui était déjà distribué auprès d'un échantillon d'individus.

5. On trouve en revanche certaines variables imputées par les responsables de l'enquête pour traiter la non-réponse partielle.

Aussi a-t-on considéré que l'ampleur de la non-réponse découlait essentiellement d'un effet agent recenseur, approche qui permet d'inclure par la même occasion un effet géographique puisque par construction la grappe concentre les logements *EFL*. L'effet géographique traduit de fait un effet de grappe qui rend la probabilité de réponse sensible aux caractéristiques sociodémographiques de la zone de collecte. La probabilité de réponse à *EFL* est donc considérée comme constante par grappe (l'organisation de la collecte attribue une grappe à un agent recenseur), et on l'estime par le taux de réponse empirique dans la grappe. C'est une approche originale que, dans toutes les enquêtes d'ailleurs, les traitements de l'échantillon national ignorent. La pratique montre qu'effectivement, d'un agent recenseur à l'autre, les taux de réponse sont très différents. Signalons que des corrections *ad hoc* sont effectuées en fin de traitement pour corriger des probabilités aberrantes, dues au fait que des erreurs de désignation des districts (ou Iris) échantillonnés ont eu lieu sur le terrain.

Encadré 2

EXPRESSION DE L'ESTIMATEUR COMMUNAL

Le modèle (cf. encadré 1) a distingué des sous-populations indexées par h, en nombre H. On reprend les notations de l'encadré 1. Il convient d'estimer le ratio $\sum_{com} N_{com}^{ch,h,ssp} \Big/ \sum_{com} N_{com}^{ch,h}$ qui s'interprète comme une probabilité (une proportion) inconnue. Le dénominateur est obtenu grâce aux données du recensement de 2009. Le numérateur fait intervenir les effectifs inconnus $N_{com}^{ch,h,ssp}$, que l'on va estimer dans chaque commune grâce aux données *EFL* en utilisant des estimateurs dits « par le ratio ». Ces estimateurs permettent un calage sur les tailles totales communales $N_{com}^{ch,h}$, soit

$$\widehat{N}_{com}^{ch,h,ssp} = N_{com}^{ch,h} \cdot \frac{\widehat{N}_{com}^{ch,h,ssp}}{\widehat{N}_{com}^{ch,h}}$$

La notation $\widehat{N}$ désigne un estimateur de la vraie valeur inconnue N. Lorsque l'effectif à estimer représente un effectif d'adultes (qui sont les unités d'échantillonnage), si on note w_i le poids de l'individu i dans l'*EAR* 2011 et $\widehat{R}_i$ l'estimation de sa probabilité de réponse à l'*EFL*, alors

$$\frac{\widehat{N}_{com}^{ch,h,ssp}}{\widehat{N}_{com}^{ch,h}} = \frac{\sum_{i \in r \cap com \cap h \cap ssp} \frac{w_i}{\widehat{R}_i}}{\sum_{i \in r \cap com \cap h} \frac{w_i}{\widehat{R}_i}}$$

où

$r \cap com \cap h \cap ssp$ est l'échantillon *EFL* d'individus répondants r dans la commune *com*, appartenant au champ *ch*, à la catégorie h et à la sous-population *ssp* qui nous intéresse.

$r \cap com \cap h$ est l'échantillon *EFL* d'individus répondants r dans la commune *com*, appartenant au champ *ch* et à la catégorie h.

On en tire $\widehat{N}_{com_0,L}^{ssp}$ (indice L pour « Local »), l'estimateur de type local de l'effectif inconnu $N_{com_0}^{ssp}$:

$$\widehat{N}_{com_0,L}^{ssp} = \sum_{h=1}^{H} N_{com_0}^{ch,h} \cdot \frac{\sum_{com} N_{com}^{ch,h} \cdot \frac{\sum_{i \in r \cap com \cap h \cap ssp} \frac{w_i}{\widehat{R}_i}}{\sum_{i \in r \cap com \cap h} \frac{w_i}{\widehat{R}_i}}}{\sum_{com} N_{com}^{ch,h}}.$$

On remarquera que les composantes de $\widehat{N}_{com_0,L}^{ssp}$ font référence à des dates différentes, parfois 2009, parfois 2011, mais la probabilité estimée est bien homogène à une grandeur 2011. Comme les effectifs $N_{com_0}^{ch,h}$ faisant office de pondération sont pour leur part représentatifs d'une situation 2009, il restera à en trouver une traduction 2011 afin que l'ensemble de l'estimateur retrouve une allure homogène et puisse prétendre représenter la situation au 1er janvier 2011 (voir infra). Lorsque la statistique porte sur un effectif par sexe (ce qui sera toujours le cas dès que la variable est une variable individuelle portant sur les adultes), les communes *com* impliquées dans les sommations de $\widehat{N}_{com_0,L}^{ssp}$ sont celles où au moins une grappe *EFL* concerne le sexe en question. Si la statistique porte sur le niveau logement / ménage / famille, il s'agit alors de l'ensemble des communes participant à *EFL*, mais dans ce cas l'expression $\widehat{N}_{com_0,L}^{ssp}$ doit être un peu adaptée car il faut faire intervenir des effectifs définis au niveau logement (par exemple un nombre total d'enfants dans le logement i appartenant à une certaine sous-population).

Les effectifs $N_{com_0}^{ch,h}$ et $N_{com}^{ch,h}$, dans les grandes communes du moins, ne sont pas connus de manière « exacte » à cause de l'erreur d'échantillonnage propre au recensement. Dans l'expression $\widehat{N}_{com_0,L}^{ssp}$, il faut donc comprendre que les structures $N_{com_0}^{ch,h}$ et $N_{com}^{ch,h}$ sont en réalité des structures estimées utilisant la pondération du recensement. En terme de variance d'échantillonnage, si on ne prend pas en compte l'incertitude attachée aux estimations des structures $N_{com_0}^{ch,h}$ et $N_{com}^{ch,h}$, la qualité de $\widehat{N}_{com_0,L}^{ssp}$ dépendra de l'inverse de la taille d'échantillon répondant dans l'ensemble des communes *EFL* qui sont impliquées dans son calcul. S'agissant cette fois d'un échantillon supra communal, *a priori* de taille respectable, on peut espérer une variance faible.

Une fois obtenue l'estimation des probabilités de réponse individuelles, chaque individu répondant à *EFL* est pondéré par le produit de trois termes : le poids de l'*EAR* (ramené à 1 dans toutes les petites communes), l'inverse du taux de sondage des grappes dans la commune (prenant en compte la phase spécifique d'affectation des sexes aux grappes) et l'inverse de sa probabilité de réponse estimée. En théorie, l'estimateur du total pondéré de cette manière est sans biais pour estimer un total relatif à un sexe donné[6] (en pratique, il y a un biais dû à l'inexactitude du modèle expliquant la probabilité de réponse). Comme la probabilité qui est au cœur du modèle et que l'on cherche à estimer est un ratio, le taux de sondage des grappes se simplifie et, de fait, n'intervient jamais (cf. encadré 2). Lorsque l'information est définie au niveau du logement / ménage / famille (un nombre d'enfants gardés en crèche par exemple), le sexe de l'enquêté n'a pas à être distingué mais finalement on aboutit à la même formulation de l'estimateur parce que le tirage des logements propre à *EFL* est à probabilités égales. L'expression de l'estimateur communal utilise les effectifs issus du *RP* 2009 pour pondérer les probabilités définies par sous-population (cf. encadré 2) : le nombre d'individus constituant une sous-population donnée dans la commune est estimé par la somme des poids de ces individus. Lorsqu'il s'agit d'une grande commune, on utilise exactement le poids du fichier *RP* 2009. Lorsqu'on a affaire à une petite commune, on prend un poids égal à 1. Il est clair que ces estimations d'effectifs issues du recensement possèdent elles-mêmes une variance d'échantillonnage et font perdre de la précision à l'estimateur final[7]. Noter que pour certaines variables, et seulement dans les petites communes, l'exploitation du recensement a porté sur un sous-échantillon composé du quart de l'échantillon de logements recensés. On parle alors d'exploitation complémentaire. Lorsque les sous-populations sont construites à partir de variables issues de l'exploitation complémentaire (par exemple le type de famille), les logements ont une pondération spécifique et la qualité des estimations des effectifs des sous-populations s'en trouve dégradée.

Malgré ses imperfections, l'estimateur local a considérablement gagné en stabilité : si l'on fait abstraction de la part de variance due au recensement lorsqu'elle est significative, sa variance d'échantillonnage (causée par *EFL*) varie comme l'inverse de la taille de l'échantillon répondant comptabilisée dans l'ensemble des communes impliquées dans le modèle (au sein d'une catégorie de communes donnée). Cette taille est grande, en tout cas très supérieure à celle que l'on trouve dans une commune donnée. L'estimateur a l'avantage de pouvoir se calculer en toutes circonstances, dans n'importe quelle commune. En particulier, on peut estimer un effectif portant sur un sexe qui n'est pas du tout enquêté dans la commune - étant entendu que la commune en question n'a (évidemment) pas participé à l'estimation de la probabilité relative à ce sexe.

Dans certaines petites communes, constituées d'une seule grappe ou partitionnées en deux grappes, le tirage des grappes *EFL* a été exhaustif, si bien que l'ensemble des logements de la commune ont été concernés par l'enquête. Néanmoins au niveau individuel, chaque grappe *EFL* est associée à un sexe : on peut donc distinguer le sous-ensemble des petites communes pour lesquelles l'un des sexes est enquêté exhaustivement. Modulo la non-réponse, l'*EFL* devient alors un recensement auprès des individus de ce sexe. Dans ces communes là et seulement pour le sexe enquêté exhaustivement, on a naturellement retenu comme estimateur le « véritable » effectif, que l'on peut obtenir de manière immédiate puisqu'il suffit de dénombrer les individus (soit homme, soit femme) qui appartiennent à la sous-population d'intérêt. Si la variable se définit au niveau logement / ménage / famille, on applique cette méthode dès lors que le tirage des grappes *EFL* est exhaustif (un peu plus de 600 petites communes sont dans ce cas). La statistique obtenue n'est cependant pas parfaite puisqu'il faut toujours corriger la non-réponse. Si toutes ces communes spécifiques ont été soumises à un traitement particulier au niveau de l'estimation de l'effectif recherché, en revanche elles ont participé aux estimations des probabilités exactement comme toutes les autres communes.

6. La nature des information restituées, distinguant les sexes pour les dénombrements de personnes, incite à une inférence conditionnelle à la labellisation par sexe des grappes. Il est donc préférable de raisonner sexe par sexe pour calculer le taux de sondage des grappes au sein de la commune.

7. De ce point de vue, un recensement exhaustif aurait été plus efficace. En la circonstance, aucune alternative n'est offerte et il faut accepter cette part d'instabilité. Cela étant, le risque est essentiellement concentré dans les « petites grandes communes » parce que c'est là que le recensement est en réalité une enquête par sondage et que les échantillons de logements ont les tailles les plus modestes.

Estimations de probabilités pour les variables exploitées : la phase de modélisation

Nous allons exposer, de manière détaillée pour trois variables choisies parmi l'ensemble des variables exploitées, les principales étapes de la démarche conduisant au modèle retenu en signalant les traitements et les écueils qui sont spécifiques. Le modèle produit toujours un ensemble de proportions estimées. La phase finale d'estimation des effectifs communaux, pour chaque sous-population d'intérêt, sera présentée dans la partie suivante. Pour chacune des variables, la première opération consiste à rapprocher le fichier de collecte *EFL* et le fichier des individus recensés en 2011. En situation idéale, on dispose d'un vaste fichier regroupant, pour chaque individu enquêté, les informations *EFL* et toutes les variables du recensement. En réalité, on perd un certain nombre d'enquêtés *EFL* à cause de défauts d'appariement entre la source *EFL* et l'*EAR*. Cette perte n'a pas de conséquence significative parce que les données *EFL* n'interviennent qu'au travers de ratios dans l'expression des proportions estimées.

La phase de modélisation aboutit à la définition des catégories de population *h* (cf. encadré 1). L'outil statistique utilisé est la régression logistique additive (sans interactions) et pondérée. Les variables explicatives en entrée du modèle sont construites à partir des informations disponibles dans les questionnaires du recensement, qui est la seule source d'information (pseudo) exhaustive que l'on peut mobiliser à un niveau communal et qui présente des variables communes avec l'enquête *EFL* (au moins en terme de concept). Elles dépendent évidemment de choix *a priori* et évoluent selon la variable à expliquer. L'examen des paramètres du modèle ajusté, s'appuyant en particulier sur les tests de significativité des coefficients, permet d'effectuer des sélections puis des regroupements de modalités pour aboutir à une liste de variables significativement explicatives. Pour éviter un degré de complication trop élevé sans impacter significativement les résultats, on a retenu les mêmes catégories pour les hommes et pour les femmes.

Pour toutes les variables à expliquer, deux variables géographiques ont été mobilisées : la ZEAT (Zone d'études et d'aménagement du territoire), qui est un regroupement de régions administratives contiguës (huit modalités au total), et la tranche d'unité urbaine construite à partir du recensement de 2010. En situation standard, le processus de prise en compte de la géographie s'effectue en deux temps. D'abord on ajuste le modèle logistique en distinguant comme variables explicatives l'ensemble des ZEAT et l'ensemble des tranches d'unité urbaines. Sur la base des estimations de coefficient obtenues, on effectue des regroupements de modalités, ce qui donne lieu à des catégories de communes définies d'une manière *ad hoc*, qu'il faut énoncer en extension. La vocation des variables géographiques est de découper le territoire afin de donner lieu à des modèles séparés, tout simplement parce que les communes sont elles-mêmes des entités géographiques réparties entre les ZEAT et les tranches d'unité urbaine. C'est un processus naturel qui conduit généralement à effectuer un ajustement du modèle définitif pour chaque catégorie de communes, sous condition que le nombre de catégories de communes reste limité et donne lieu à des tailles d'échantillon *EFL* globales par catégorie de communes compatibles avec l'approche par modèle. Cela étant, parfois, on peut considérer (et vérifier) qu'une variable géographique joue un rôle essentiel dans l'explication du phénomène alors même que le découpage en catégories de communes conduirait à distinguer trop de catégories. Dans ce cas, pour contourner le caractère peu économe du passage par les catégories de commune, on cherche à croiser cette variable incontournable avec les variables sociodémographiques du recensement mais sans qu'il y ait de croisement deux à deux de toutes les modalités de chacun des deux groupes de variables : ainsi, on limite les termes d'interaction, comme dans le cas du mode de garde, où le caractère plus ou moins urbain de l'environnement (la tranche d'unité urbaine) est déterminant mais sans venir en croisement systématique de toutes les catégories de ménages définies par ailleurs. Dans le prolongement de cette logique, on pourrait enrichir les modèles individuels par des variables explicatives, autres que géographiques, définies au niveau communal, voire supra-communal (ce qui n'a pas été fait). Au-delà de la difficile phase de présélection de telles variables, plutôt *a priori* en surabondance, cela reviendrait à définir des catégories de communes encore plus fines et on se heurterait vite au problème de taille d'échantillon insuffisante par catégorie de communes (sauf à introduire des interactions *ad hoc* avec les variables individuelles mais cela serait extrêmement lourd).

Nombre d'hommes, de femmes, de couples ayant souscrit un pacte civil de responsabilité (pacs)

Le questionnaire *EFL* demande à l'enquêté(e) s'il (ou elle) est ou a été « pacsé(e) » avec la personne avec laquelle il (elle) est en couple. Cette question n'est posée qu'aux personnes déclarant préalablement être en couple. Elle peut appeler une réponse positive même s'il y a eu mariage ultérieur. L'intérêt de l'Insee, dans la perspective de la restitution aux communes, s'est porté sur les effectifs respectivement d'hommes et de femmes pacsé(e)s, puis de couples pacsés, en s'en tenant aux personnes en couple mais non mariées. On a en effet considéré qu'il n'était pas intéressant d'inclure les individus qui ont été d'abord pacsés puis se sont mariés avant la date de l'enquête. On a ajouté la condition de vie dans le même logement (il est en effet possible de se déclarer en couple avec une personne vivant dans un autre logement). On cherche donc à dénombrer, par sexe, les individus pacsés, non mariés, en couple et cohabitant.

À ce stade, le champ (*ch* dans les encadrés 1 et 2) devrait donc tout naturellement être défini comme l'ensemble des personnes non mariées, vivant en couple dans le même logement. Cette définition d'apparence simple crée deux difficultés. La première n'est pas la pire et part du fait que le recensement ne précise pas si la vie en couple s'effectue avec une personne qui réside dans le même logement ou dans un autre logement. Sans espoir d'éclaircir la situation au moyen d'autres variables du recensement et sans engager de traitements complémentaires, on est donc parti du principe qu'en déclaration par mode de dépôt-retrait, dans la plupart des cas l'enquêté allait naturellement interpréter cette question au sens d'une cohabitation (ce que la comparaison entre l'*EFL* et l'*EAR* 2011 valide *a posteriori* ; voir Breuil-Genier *et al.*, dans ce numéro). Cette hypothèse se trouve renforcée par le libellé de la question de l'*EAR* « *vivez-vous en couple ?* » alors que l'*EFL* demande « *êtes-vous en couple ?*». La seconde difficulté est plus sérieuse et provient du fait que non seulement l'information complète définissant le champ doit être disponible dans le *RP* pour permettre le calcul des « vraies » structures communales, mais encore faut-il une parfaite homogénéité de concept entre *EFL* et le *RP*, faute de quoi on introduit un biais (résultant d'une erreur de mesure). Or en la circonstance cette homogénéité n'existe pas (au moins) pour l'état matrimonial. En effet, dans le recensement il est clair que la qualité de l'information individuelle portant sur la situation matrimoniale n'est pas celle de l'enquête *EFL*, plus précise sur ce point et qui requiert davantage d'attention et de cohérence dans les réponses de la part de l'enquêté. En particulier, on trouve un nombre certes faible mais néanmoins non négligeable d'individus qui se déclarent non mariés à l'*EFL* mais (vraisemblablement à tort) mariés au *RP*, et réciproquement. Il existe donc des individus mariés d'après leur déclaration *RP* mais pacsés au sens qui a été défini précédemment. Le terme de célibataire n'a donc pas été perçu par certaines personnes comme reflétant suffisamment bien leur vie conjugale contractualisée. Pour cette raison, le champ *ch* a été limité à l'ensemble des personnes vivant en couple dans le même logement, qu'elles soient mariées ou non au sens du recensement. Cette originalité va permettre de récupérer des individus effectivement pacsés dans les conditions qui ont été définies, c'est-à-dire des personnes vivant en couple, non-mariées selon l'*EFL* (sous-entendu en réalité non-mariées), mais qui se déclarent par ailleurs mariés dans le *RP*. C'est le constat d'une incohérence dans les déclarations (pourtant particulièrement associées…) de l'*EAR* et de l'*EFL* ; on peut le déplorer, mais on ne peut pas raisonnablement exclure en amont les personnes mariées selon le *RP* sans risque d'une sous-estimation significative de l'effectif pacsé. Cela étant, comme cette sous-population concernée par l'incohérence des déclarations d'état matrimonial est probablement très particulière, on a pris le parti, pour les besoins de la modélisation, de créer une catégorie de population spécifique regroupant tous les individus mariés selon le recensement.

Conformément à l'esprit général qui a été développé supra, le modèle de comportement est le suivant : pour un homme (respectivement une femme) déclarant vivre en couple cohabitant au moment de l'enquête, la probabilité d'être pacsé(e) ne dépend que de son appartenance à une sous-population définie par des critères d'âge et de diplôme, auxquels il faut ajouter la prise en compte de l'état marié / non-marié selon la déclaration au *RP* (Davie, 2011). En particulier, cette probabilité ne dépend pas de sa commune de résidence, dès lors que cette commune appartient à une catégorie de communes définie par certains critères qui vont être précisés ci-après et qui sont de toute façon tels que le nombre de communes participant à l'*EFL* et constituant cette catégorie est grand. Pour chaque sexe, les critères d'âge et de diplôme ont été retenus à l'issue d'une régression logistique (pondérée)

qui a permis de sélectionner les modalités les plus explicatives de la probabilité individuelle d'être pacsé. Ont été ainsi retenues sept catégories de population (cf. tableaux 1 et 2) .

Tableau 1
Par catégorie de population et catégorie de communes, probabilité de pacs des hommes

Hommes			Catégorie de communes n° 1		Catégorie de communes n° 2	
Catégorie de population		Effectif répondant	Effectif répondant	Probabilités de pacs estimées (en %)	Effectif répondant	Probabilités de pacs estimées (en %)
Individus mariés selon le recensement		63 709	21 623	0,9	42 086	1,2
Individus non mariés	Diplôme de niveau inferieur au baccalauréat	7 693	2 393	8,3	5 300	8,5
	Diplôme de niveau baccalauréat, naissance en 1969 ou avant	1 841	529	11,8	1 312	12,3
	Diplôme de niveau baccalauréat (naissance entre 1970 et 1979) ou diplôme de niveau supérieur au baccalauréat (naissance en 1969 ou avant)	2 643	1 012	11,7	1 631	9,3
	Diplôme de niveau baccalauréat, naissance en 1980 ou plus tard	3 962	1 220	28,2	2 742	18,4
	Diplôme de niveau supérieur au baccalauréat, naissance entre 1970 et 1979	2 446	675	34,0	1 771	31,0
	Diplôme de niveau supérieur au baccalauréat, naissance en 1980 ou plus tard	2 334	694	23,2	1 640	16,4
Ensemble		84 628	28 146	5,4	56 482	5,0

*Lecture : un homme qui se déclare au recensement non marié mais vivant en couple, ayant un diplôme de niveau inférieur au baccalauréat (catégorie de population = 2) et résidant dans une commune quelconque située en région Alsace (catégorie de communes = 1) aura une probabilité d'être pacsé (donc répondant à l'*EFL *qu'il est pacsé, non marié, et vivant en couple cohabitant) égale à 8,3 %.*
Champ : ensemble des hommes adultes, en métropole, vivant en ménage ordinaire, déclarant au recensement vivre en couple.
Source : EFL *2011,* EAR *2011,* RP *2009, calculs de l'auteur.*

Tableau 2
Par catégorie de population et catégorie de communes, probabilité de pacs des femmes

Femmes			Catégorie de communes n° 1		Catégorie de communes n° 2	
Catégorie de population		Effectif répondant	Effectif répondant	Probabilités de pacs estimées (en %)	Effectif répondant	Probabilités de pacs estimées (en %)
Individus mariés selon le recensement		112 757	38 782	0,9	73 975	0,9
Individus non mariés	Diplôme de niveau inferieur au baccalauréat	10 304	3 434	6,3	6 870	6,4
	Diplôme de niveau baccalauréat, naissance en 1969 ou avant	3 183	910	10,9	2 273	10,4
	Diplôme de niveau baccalauréat (naissance entre 1970 et 1979) ou diplôme de niveau supérieur au baccalauréat (naissance en 1969 ou avant)	5 931	2 171	10,9	3 760	11,0
	Diplôme de niveau baccalauréat, naissance en 1980 ou plus tard	6 382	1 874	14,4	4 508	12,5
	Diplôme de niveau supérieur au baccalauréat, naissance entre 1970 et 1979	5 020	1 446	31,0	3 574	29,0
	Diplôme de niveau supérieur au baccalauréat, naissance en 1980 ou plus tard	6 851	2 231	21,1	4 620	17,2
Ensemble		150 428	50 848	4,5	99 580	4,4

Lecture : se reporter au tableau 1.
Champ : ensemble des femmes adultes, en métropole, vivant en ménage ordinaire, déclarant au recensement vivre en couple.
Source : EFL *2011,* EAR *2011,* RP *2009, calculs de l'auteur.*

Par convention, les catégories sont les mêmes pour les hommes et pour les femmes, quitte à ce que pour un sexe donné on retienne quand même une catégorie non significative. Les communes ont pour leur part été ventilées en deux catégories, distinguées au travers de la variable ZEAT. La significativité de la variable ZEAT provient de la procédure logistique évoquée ci-dessus – laquelle ne sert donc pas à isoler seulement les catégories d'individus. La catégorie de communes 1 regroupe les communes des régions Alsace, Lorraine, Franche-Comté, Nord-Pas-de-Calais, Pays de Loire, Bretagne, Poitou-Charentes, la catégorie 2 regroupe toutes les autres communes. Quel que soit le phénomène social en jeu (ici le pacs), il n'est pas toujours évident de percevoir la logique de ces différents regroupements, qu'il s'agisse des catégories d'individus ou des catégories de communes ; néanmoins, on a fait confiance à la fois aux données et aux outils de la statistique, sans imposer *a priori* que les regroupements issus des procédures logistiques apparaissent conformes à l'intuition.

Faisons le point sur l'utilisation des sources. Pour l'estimation de la probabilité d'être pacsé, on tire de l'*EFL* la variable dichotomique qui repère la situation de pacs, qui elle-même fait appel à trois variables du questionnaire : la signature d'un pacs, l'état matrimonial (non-marié) et la variable « vie en couple dans le même logement », dont l'exploitation est naturelle parce que cette dernière sert de filtre dans le questionnaire pour répondre à la question sur le pacs. S'ajoute la variable sexe (exceptionnellement en incohérence avec celle du questionnaire *EAR*). Parmi les variables provenant de l'*EAR* 2011 on trouve les variables d'état matrimonial, d'âge et de diplôme définissant l'appartenance des individus *EFL* aux sept catégories de population, ainsi que la variable « vie en couple » qui a servi pour sélectionner dans l'*EFL* les individus du champ. Le recensement 2009 fournit pour sa part les variables (sexe, état matrimonial, indicateur de vie en couple, âge, diplôme) permettant d'obtenir les « vrais » effectifs communaux par catégorie pondérant les probabilités précédemment estimées.

On estime, d'une part le nombre d'hommes pacsés et d'autre part le nombre de femmes pacsées. *In fine*, on additionne ces deux effectifs et on divise par deux pour estimer le nombre total de couples pacsés dans la commune.

La modélisation logistique permettant de définir les six catégories de population distinguées parmi les individus non mariés au recensement et vivant en couple a porté sur un échantillon de 20 919 hommes (adultes) répondants. Il faut rajouter 63 709 hommes se déclarant mariés au recensement et vivant en couple, formant la catégorie spécifique n° 1. Cet effectif global de 84 628 hommes est donc ventilé en sept catégories de population, mais aussi deux catégories de communes : 28 146 hommes résident dans des communes de catégorie 1 (434 communes concernées) et 56 482 dans des communes de catégorie 2 (1 004 communes concernées). Les probabilités de pacs des hommes (exclusivement pour les individus non mariés et vivant en couple cohabitant) sont données, pour chacune des catégories de population et de communes, dans le tableau 1.

Dans exactement les mêmes conditions que pour les hommes, l'effectif global répondant des femmes est égal à 150 428 personnes, ce qui apparaît comme un ordre de grandeur correct puisque dans l'*EFL* on a échantillonné deux fois plus de grappes consacrées à l'enquête des femmes qu'à l'enquête des hommes. À l'image exacte du tableau 1, le tableau 2 concerne les femmes.

Qu'il s'agisse des hommes ou des femmes, on constate que la catégorie de population est extrêmement discriminante. La catégorie de communes l'est sensiblement moins, surtout pour les femmes. L'existence d'une probabilité de pacs très faible mais non nulle dans la catégorie des personnes se déclarant mariées au recensement (catégorie n° 1) est caractéristique du problème des erreurs de mesure (*a priori* dans le recensement) qui a été souligné supra. Cette probabilité joue un rôle important car bien qu'étant numériquement très faible (voisine de 1 %), elle vient multiplier les effectifs communaux de loin les plus considérables (ceux des personnes mariées). De fait, la catégorie 1 contribue à hauteur d'environ 15 % à l'estimation du nombre total d'individus pacsés.

Nombre d'enfants de moins de 4 ans, selon leur mode de garde principal

Quelques questions portent, non pas sur l'individu enquêté, mais sur son ménage (ou son logement). C'est le cas du nombre d'enfants de moins de 4 ans qui vivent dans le logement et qui sont gardés selon les modes suivants : garde des parents, garde des grands-parents ou de la famille, assistante maternelle, crèche, ou autre mode de garde. Cette question est censée ne

recueillir que le mode de garde principal. En réalité, on trouve des questionnaires (en faible nombre) où plusieurs modalités sont remplies[8], auquel cas l'enfant concerné a été dupliqué.

Pour cette variable, l'unité modélisée (un enfant) n'est plus l'unité d'échantillonnage (un adulte). En effet, le modèle porte sur la probabilité qu'un enfant donné soit gardé (principalement) par tel ou tel mode de garde. Comme on distingue cinq modes de garde, il y a cinq probabilités associées, dont la somme vaut systématiquement 1. Le modèle fonctionne ainsi : pour un enfant donné dans une catégorie de communes donnée, sa probabilité d'être gardé (principalement) selon un mode de garde donné ne dépend pas de la commune dans laquelle il réside, mais seulement de la catégorie à laquelle appartient son ménage. Par conséquent, une typologie de ménages a été formée, en utilisant l'outil de régression logistique (pondérée), qui s'appuie sur des informations liées à l'effectif et à l'activité des adultes composant le ménage. On procède comme suit. Tout d'abord, on retient pour chaque adulte du ménage trois tranches d'activité possibles selon la situation principale déclarée à l'*EAR* 2011 (question 10 du bulletin individuel (BI)) :

- Emploi, apprentissage, études ;
- Chômage ;
- Retraite, personne au foyer, autre.

Pour les personnes qui travaillent, on prend également en compte le temps de travail déclaré (question 22 du BI) :

- Emploi occupé à temps complet ;
- Emploi occupé à temps partiel.

In fine, huit catégories de ménage ont été définies (cf. tableau 3).

Placer une activité au cœur de la définition des catégories de population discriminantes pour le calcul des probabilités n'est pas sans risque. En effet, l'activité traduit par nature une situation conjoncturelle. Lorsqu'il s'agit seulement de récupérer l'activité au moment de l'enquête (*via* l'*EAR* 2011) pour déterminer la catégorie du ménage, le caractère instable dans le temps de l'activité n'affecte probablement que de façon modérée sa relation avec le mode de garde utilisé. Mais lorsqu'on en arrive au calcul des vrais effectifs communaux d'enfants relatifs à chaque catégorie de ménages distinguée, en exploitant cette fois l'intégralité du recensement 2009, il y a un brouillage inévitable de l'information dû à la superposition de cinq années, ce qui pourrait avoir des conséquences perturbatrices se traduisant en biais d'estimation (sans parler de l'effet du décalage temporel de deux années entre la source *RP* 2009 et la source *EFL* 2011). On peut donc hésiter à franchir le pas ; certes on peut toujours contester le choix des modalités retenues pour définir les catégories de ménage, néanmoins il a paru contraire au bon sens de ne pas exploiter cette information, tant elle apparaît en pratique déterminante dans le choix du mode de garde principal : d'autres variables, certainement plus stables dans le temps, n'auraient eu qu'un (trop) faible pouvoir explicatif. Cela étant, on peut penser qu'au-delà de l'effectif des adultes dans le ménage et de leurs activités, l'environnement plus ou moins urbain a une influence sensible sur les modes de garde. C'est pourquoi on a croisé certaines des catégories définies ci-dessus avec une information sur la tranche d'unité urbaine. Pour cela, on distingue quatre tranches d'unités urbaines : communes rurales (tranche 1), communes d'unités urbaines ayant moins de 20 000 habitants (tranche 2), communes d'unités urbaines ayant 20 000 habitants ou plus mais moins de 200 000 habitants (tranche 3), communes d'unités urbaines ayant 200 000 habitants ou plus (tranche 4). Il s'est avéré que la distinction des communes selon leur tranche d'unité urbaine n'était envisageable que dans les catégories de population 4, 5, 6 et 7, là où la taille d'échantillon globale est suffisante. Par conséquent, pas moins de 20 catégories de ménage ont été finalement distinguées.

Les communes ont pour leur part été séparées en deux catégories, de nouveau sur la base de la variable ZEAT : les communes des régions Champagne-Ardenne, Picardie, Haute-Normandie, Basse-Normandie, Centre, Bourgogne, Nord-Pas-de-Calais, Alsace, Lorraine, Franche-Comté, Languedoc-Roussillon, PACA et Corse forment la catégorie 1 (cf. tableau 3), les autres communes la catégorie 2 (cf. annexe 1).

Le dénombrement des enfants s'appuie fondamentalement sur l'unité ménage dans lequel ils vivent. Le champ associé à la variable mode de garde est constitué par la sous-population des ménages dans lesquels on trouve au moins un enfant de moins de 4 ans vivant dans le logement. C'est une information qui a été construite au niveau ménage en exploitant les seuls BI

8. Un apurement ultérieur du fichier national a fait disparaître ce cas de figure.

du recensement, les déclarations d'*EFL* relatives au nombre et à la liste des enfants vivant dans le ménage n'intervenant jamais : comme pour la variable « pacs », il existe parfois une incohérence entre la déclaration des enfants au recensement et celle faite à l'*EFL* (un tableau du questionnaire *EFL* liste les enfants vivant dans le logement, avec leurs âges). Mais, par raison d'homogénéité, afin de limiter le biais (conséquence de l'erreur de mesure), une source unique (le recensement) a été utilisée pour sélectionner les ménages du champ retenu.

Tableau 3
Par sous-population, en catégorie de communes n° 1, probabilité d'un enfant de moins de 4 ans d'être gardé à titre principal par un des modes de garde

Sous-population		Effectif répondant	Probabilité par mode de garde principal (en %)				
Catégorie de ménage	Tranche d'unité urbaine		Parents	Grands-parents, famille	Assistante maternelle	Crèche	Autre mode
Ménage avec 1 adulte qui travaille à temps plein	Toutes	383	26,0	17,0	25,9	27,5	3,6
Ménage avec 1 adulte qui travaille à temps partiel ou qui est en stage / étude	Toutes	160	41,0	6,8	21,2	28,3	2,7
Ménage avec 1 adulte au chômage	Toutes	210	69,8	2,6	5,7	18,8	3,1
Ménage avec 1 adulte retraité ou au foyer	1	62	93,2	2,1	1,6	2,1	1,0
Ménage avec 1 adulte retraité ou au foyer	2	65	87,9	1,8	8,0	2,3	0,0
Ménage avec 1 adulte retraité ou au foyer	3	168	93,5	0,7	3,3	1,5	1,0
Ménage avec 1 adulte retraité ou au foyer	4	157	87,6	1,4	2,3	4,6	4,1
Ménage avec au moins 2 adultes, et au moins un retraité ou une personne au foyer	1	1 358	89,2	2,2	5,6	2,4	0,6
Ménage avec au moins 2 adultes, et au moins un retraité ou une personne au foyer	2	945	88,7	2,3	5,1	3,7	0,2
Ménage avec au moins 2 adultes, et au moins un retraité ou une personne au foyer	3	1 533	90,9	0,7	3,0	4,8	0,4
Ménage avec au moins 2 adultes, et au moins un retraité ou une personne au foyer	4	1 464	91,0	1,4	2,9	4,0	0,8
Ménage avec au moins 2 adultes, tous les adultes travaillent à temps plein	1	2 553	27,5	9,7	52, 5	7,9	2,4
Ménage avec au moins 2 adultes, tous les adultes travaillent à temps plein	2	1 213	27,3	9,4	44,4	17,1	1,8
Ménage avec au moins 2 adultes, tous les adultes travaillent à temps plein	3	1 376	28,6	10,3	35,9	21,5	3,7
Ménage avec au moins 2 adultes, tous les adultes travaillent à temps plein	4	1 537	28,7	9,8	33,8	24,8	2,9
Ménage avec au moins 2 adultes, tous les adultes travaillent ou sont en stage / étude mais au moins un adulte ne travaille pas à temps plein	1	1 466	33,2	7,1	50,1	8,0	1,6
Ménage avec au moins 2 adultes, tous les adultes travaillent ou sont en stage / étude mais au moins un adulte ne travaille pas à temps plein	2	717	38,5	8,7	36,7	14,2	1,9
Ménage avec au moins 2 adultes, tous les adultes travaillent ou sont en stage / étude mais au moins un adulte ne travaille pas à temps plein	3	822	40,4	5,4	29,3	21,9	3,0
Ménage avec au moins 2 adultes, tous les adultes travaillent ou sont en stage / étude mais au moins un adulte ne travaille pas à temps plein	4	979	37,3	8,1	23,7	27,3	3,6
Ménage avec au moins 2 adultes, aucun n'est retraité ni personne au foyer mais au moins l'un d'eux est au chômage	Toutes	2 174	67,5	3,3	11,7	16,2	1,3
Ensemble		19 342	62,1	5,0	16,8	14,2	1,9

Lecture : considérant un enfant de moins de 4 ans vivant dans un ménage comprenant un seul adulte, retraité ou vivant au foyer, dans une commune de tranche d'unité urbaine 2 et dans une des régions formant la catégorie de communes 1, cet enfant a une probabilité estimée à 8 % d'être gardé (principalement) par une assistante maternelle.
Champ : ensemble des enfants de moins de 4 ans vivant dans un ménage ordinaire de métropole.
Source : EFL *2011,* EAR *2011,* RP *2009, BPE 2010, fichier « Particuliers-employeurs » 2010, distancier Odomatrix Inra Umr 1041 Cesaer, calculs de l'auteur.*

Le dénombrement au recensement des enfants de moins de 4 ans à partir des différentes *EAR* s'effectue en mobilisant l'année de naissance et le mois de naissance. Pour une année de collecte t, on dénombre ainsi les enfants nés en t, $t-1$, $t-2$, $t-3$ ou $t-4$ et dans ce dernier cas on exclut ceux nés en janvier.

Les catégories de ménages (et donc d'enfants) définies sont, globalement, bien discriminantes lorsque l'on considère l'ensemble des modes de garde. Les tailles d'échantillon obtenues dans les deux catégories de communes dénombrent les ménages répondant et non les enfants (qui sont évidemment plus nombreux puisqu'il est possible de préciser le mode de garde principal de deux enfants dans *EFL*). Il ne faut pas s'inquiéter de la présence de quelques catégories de ménage associées à de (très) petites tailles d'échantillon : celles-ci délimitent des sous-populations (très) peu nombreuses au recensement et ont été maintenues pour respecter la logique du modèle afin de limiter (en théorie) le biais. De toute façon, la variance de l'estimateur local dépend de la taille d'échantillon totale dans la catégorie de communes[9].

La vraie répartition communale des effectifs d'enfants selon le mode de garde principal est certainement sensible à l'offre locale en matière de crèches et d'assistantes maternelles. En effet, si la commune ne dispose pas de crèche et que les communes alentours, facilement accessibles, n'en disposent pas non plus, alors on peut penser qu'aucun enfant (ou très peu d'entre eux) n'est gardé en crèche. C'est pourquoi il nous est apparu déraisonnable d'ignorer l'offre locale concernant l'équipement en crèches et les prestations de service des assistantes maternelles. Il existe un répertoire d'équipements communaux appelé « Base permanente des équipements » (BPE) qui donne, pour chaque commune, le nombre de crèches situées sur son territoire (voir www.insee.fr / Définitions et méthodes / Sources et méthodes / BPE). Lorsque la commune ne dispose d'aucune crèche sur son territoire, il est possible d'obtenir, grâce à un distancier (distancier Odomatrix Inra Umr 1041 Cesaer), le temps en minutes nécessaire pour atteindre la plus proche commune qui dispose de l'équipement : cette « distance » (en fait une durée) servant de base au repérage de la commune la plus proche est comptée de mairie à mairie et le temps de trajet est mesuré aux heures creuses, donc en l'absence de toute congestion dans le trafic automobile. Le répertoire utilisé est celui de l'année 2010. Comme tout répertoire, il peut être affecté par des erreurs de mesure, surtout pour un équipement tel qu'une crèche dont il est bien signalé que le périmètre n'est pas défini de manière parfaitement claire. Le principe retenu à ce jour pour l'estimation localisée du nombre d'enfants gardés en crèche est le suivant :

- lorsqu'une crèche au moins est recensée sur le territoire communal, on applique exactement le modèle précédent ;

- lorsque la commune ne dispose d'aucun équipement, on part du principe que les parents recherchent une crèche dans la commune la plus proche. On a considéré que le temps de déplacement ne devait pas dépasser 20 minutes pour atteindre une crèche dans la commune voisine la plus proche et équipée : si le temps relevé est inférieur à 20 minutes, on applique le modèle initial. Sinon, on impose l'absence de garde en crèche pour tout ménage de la commune. Ce dernier cas a concerné finalement 155 communes participant à l'*EFL* (situation qui touche 15,6 % des communes de l'ensemble de la métropole).

Cette méthode est cependant fragile, au moins à trois égards. D'une part il est clair que la bonne variable n'est pas le nombre de crèches mais la capacité d'accueil totale offerte par l'ensemble des équipements de chaque commune. D'autre part, il aurait fallu avoir une vision des conditions d'accès à l'ensemble des communes voisines et non pas à la seule commune la plus proche. Enfin, les distances sont comptées de mairie à mairie, et dans les communes très étendues ou très peuplées, un tel décompte peut s'écarter notoirement de la réalité pour certains ménages. La capacité d'accueil n'est pas disponible dans la BPE, cette dernière a donc été exploitée au mieux. La prise en compte de l'ensemble des communes voisines est en théorie possible grâce au distancier mais cela aurait généré des travaux qui ont été jugés excessifs. Quant à la capacité du distancier à représenter la réalité du terrain en toutes circonstances, il paraît impossible d'en dépasser les limites. Si le mode de garde « crèche » est *in fine* déclaré inexistant, on reporte l'effectif théorique de garde en crèche entre les quatre autres modes de garde, de manière uniforme.

Cette technique a également été appliquée pour moduler la probabilité de garde assurée par des assistantes maternelles. La source est cette fois le fichier administratif Particuliers-employeurs,

9. On retrouve le même contexte lorsqu'on stratifie une population en utilisant une allocation grosso modo *proportionnelle.*

géré par l'Insee (Colin, 2012). On y dénombre, toujours pour l'année 2010, les assistantes maternelles déclarées, ayant travaillé au moins une journée durant l'année. La commune est la commune de résidence de l'assistante maternelle, qui dans la très grande majorité des cas assure les gardes à son domicile. Cette fois, les corrections apportées par la méthode sont négligeables parce que les communes dans lesquelles aucune assistante maternelle n'est accessible en moins de 20 minutes sont – selon le fichier exploité – très rares : 0,4 % de l'ensemble des communes de métropole se trouvent dans cette situation. Finalement, seules 10 communes participant à l'*EFL* ont été corrigées afin de forcer à zéro le nombre total d'enfants gardés par une assistante maternelle (dont 6 communes en Corse).

Nombre d'enfants vivant en familles traditionnelle, recomposée, et monoparentale

Pour certaines exploitations du recensement, notamment pour étudier la composition des familles et en particulier celles avec enfant(s), on mobilise l'information recueillie *via* le bulletin individuel et *via* la feuille de logement. Une famille est alors définie comme un ensemble de personnes vivant dans un même logement (personnes cohabitantes) et constituant l'une des trois configurations (ou « types de familles ») suivantes :

- un seul adulte ayant un ou des enfant(s) : famille monoparentale ;

- deux adultes de sexes opposés vivant en couple, ayant un ou des enfant(s) pour chacun desquels ces deux adultes se déclarent être respectivement le père et la mère : famille traditionnelle ;

- deux adultes de sexes opposés vivant en couple, ayant un ou des enfant(s) et pour au moins l'un d'entre eux, soit l'adulte de sexe masculin n'est pas le père, soit l'adulte de sexe féminin n'est pas la mère : famille recomposée.

Par convention, aucun critère d'âge n'a été retenu pour qualifier les enfants dans la définition de la famille. Toute autre configuration rencontrée au sein du logement n'est pas une famille. L'information individuelle qui permettra de construire le type de famille à partir du questionnaire *EFL* provient d'un tableau spécifique qui liste l'ensemble des enfants vivant dans le logement recensé, même une partie du temps seulement. L'une des questions posées est (si l'adulte est une femme, par exemple) « *Êtes-vous sa mère ?* », suivie de « *Son père est-il votre conjoint / ami actuel ?* ». Les réponses à ces questions permettent de distinguer une famille recomposée d'une famille traditionnelle.

Il a été décidé de produire, pour chaque commune participant à l'*EFL*, l'effectif total d'enfants de moins de 18 ans vivant dans chacun des trois types de familles. On insiste sur le fait que l'âge intervient pour délimiter la sous-population à dénombrer mais qu'il n'intervient pas dans la définition du type de famille. Le modèle fonctionne dans l'esprit de ce qui a été conçu pour la variable « mode de garde », mais en plaçant cette fois le concept de famille au centre du modèle. On commence par partitionner l'ensemble des familles en sous-populations (catégories de familles, dont on verra qu'elles sont au nombre de huit) sur la base d'un ensemble de variables recensées caractérisant (certes plus ou moins bien) le type de famille (comprenant les trois modalités décrites ci-dessus). Puis, étant donné un enfant de moins de 18 ans, on considère que sa probabilité d'appartenir à un type de famille donné dépend exclusivement de la catégorie de la famille dans laquelle il vit. En particulier, cette probabilité ne dépend pas de la commune dans laquelle il réside, du moins lorsqu'on s'en tient aux communes appartenant à une certaine catégorie de communes. Il restera à pondérer ces probabilités propres à la catégorie de famille par les effectifs communaux issus du recensement dénombrant les enfants de moins de 18 ans qui appartiennent à chacune des huit catégories de familles.

Le dénombrement des enfants se heurte à un risque de doubles comptes. En effet, l'enquête demande de lister tous les enfants vivant dans le logement, même une partie du temps. Les motifs de vie dans deux logements sont multiples : pour les enfants des familles recomposées c'est assez clair parce que les deux parents biologiques résident dans deux logements différents, pour les enfants des familles traditionnelles, l'enfant peut résider en partie ailleurs pour ses études (par exemple). Cela étant, dans tous les cas chaque individu a une résidence principale déclarée (où par définition il est censé passer la plus grande partie de son temps). C'est évidemment à l'appréciation des enquêtés et il y a probablement des erreurs de mesure mais il est nécessaire[10] que l'on s'en

10. Nécessaire dans l'optique retenue de ne pas agir sur les poids – car il existe des techniques de repondération (dites de partage des poids) qui autorisent la prise en compte systématique des unités échantillonnées à multiples reprises.

tienne au dénombrement des enfants déclarés au lieu de leur résidence principale. Si ce n'est pas le cas, on comptera certains enfants deux fois, ce qui n'est pas l'optique retenue. Pour activer ce filtre, on exploite une variable identifiante spécifique renvoyant au BI de l'enfant. Si cet identifiant est présent dans le fichier, l'enfant a été recensé dans le logement et on considère alors qu'il s'agit de sa résidence principale. Le cas contraire peut être dû à un défaut du processus sans rapport avec la réalité du terrain ; aussi, à titre de recours et de sécurité, on exploite la variable *EFL* « *Combien de temps vit cet enfant dans ce logement ?* » et on considère qu'il s'agit de sa résidence principale seulement si l'enquêté déclare qu'il y vit tout le temps (même s'il déclare que l'enfant vit dans le logement la moitié du temps ou plus, c'est qu' *a priori* le BI a été rempli dans l'autre logement, où l'enfant a effectivement sa résidence principale et que cet enfant occupe en réalité le logement enquêté la moitié du temps). Le contexte est d'ailleurs compliqué par la présence de non-réponses partielles à cette question qui ont rendu nécessaire une phase spécifique (et complexe) d'imputation préalable de cette variable : *in fine*, sur 272 000 enfants cités comme résidents dans les logements enquêtés par l'*EFL*, 16 000 environ (soit 6 %) ont été imputés pour ce qui concerne le temps passé dans le logement. La sélection des enfants en résidence principale conduit à un échantillon national de 250 500 enfants de tous âges. Parmi eux, 191 500 ont moins de 18 ans, et 1 019 n'ont pas d'année de naissance précisée. Pour ces derniers, on a imputé une année de naissance sur la base de la structure par âge des enfants pour lesquels cette information est disponible.

On considère que les incohérences entre informations déclarées dans le recensement et informations collectées à l'*EFL* sont traitées en amont : on n'a donc pas cherché à corriger, d'une façon ou d'une autre, les incohérences de déclaration concernant une même variable, comme ce peut être le cas par exemple s'agissant du nombre d'enfants de moins de 18 ans vivant dans le logement enquêté ou du nombre d'adultes qui déclarent vivre en couple (1 % des individus se déclarent à l'*EFL* être en couple cohabitant mais ne vivent pas en couple selon le recensement). De plus, les concepts de famille et de ménage coïncident très souvent ; aussi, compte tenu des erreurs introduites par ailleurs dans l'utilisation des modèles, on traite la non-réponse au niveau ménage, en ignorant toute référence à une unité de type famille. Si dans un ménage donné deux adultes du même sexe déclarent chacun au moins un enfant de moins de 18 ans, alors on ne conserve que l'adulte le plus jeune. Ce parti pris permet d'éliminer certaines configurations complexes susceptibles d'être traduites au recensement par une seule famille alors qu'il faudrait en enregistrer deux dans le ménage. Cette opération de suppression d'observations ne concerne que 138 ménages.

Les catégories de familles sont construites pour discriminer au mieux les types de familles. Comme d'habitude, il faut s'en tenir aux variables disponibles dans le recensement. Sauf dans quelques rares cas, il a été possible de récupérer les variables individuelles du recensement relatives au conjoint de l'enquêté(e) lorsqu'il (elle) vit en couple, en particulier son âge, son état matrimonial, son lien familial[11] avec les autres membres du ménage, son mode de cohabitation[12]. Comme pour l'information associée aux enfants, la liaison enquêté-conjoint a été assurée par les gestionnaires de l'enquête et la qualité de l'appariement est apparue tout à fait satisfaisante.

On traite à part l'estimation du nombre d'enfants vivant en famille monoparentale, car le fichier « familles » du recensement permet de repérer les familles monoparentales. Ce fichier étant enrichi par des variables individuelles, on est en mesure d'estimer dans chaque commune le nombre d'enfants qui vivent en famille monoparentale. Concernant les deux autres types de familles (familles biparentales), soumises à la modélisation, le champ associé au modèle est constitué par l'ensemble des enfants de moins de 18 ans déclarés au lieu de leur résidence principale et vivant en famille biparentale. Il y a deux variables essentielles pour définir les catégories de familles parmi les familles biparentales : la taille de la fratrie et le statut matrimonial des parents. À titre accessoire, on ajoute l'âge de la mère mais il n'intervient que dans deux catégories (d'après les régressions logistiques). La première variable dénombre les enfants de moins de 18 ans composant la famille. La seconde variable combine le statut matrimonial du père et celui de la mère.

---

11. Adulte de sexe masculin d'une famille / adulte de sexe féminin d'une famille / enfant d'une famille / hors famille.

12. Personne vivant seule / personne vivant hors famille dans un ménage de plusieurs personnes / adulte d'une famille monoparentale / enfant d'une famille monoparentale / adulte d'un couple sans enfant / adulte d'un couple avec enfant(s) / enfant d'un couple.

Finalement, cela conduit à distinguer les huit catégories suivantes :

1 : père et mère mariés, fratrie de taille 1 ou 2 ;

2 : (père et mère mariés, fratrie de taille 3 ou 4) ou (l'un des parents est célibataire, l'autre marié ou célibataire, fratrie de taille 1 ou 2) ;

3 : (père et mère mariés, fratrie de taille 5 ou plus) ou (l'un des parents est célibataire, l'autre marié ou célibataire, fratrie de taille 3 ou 4, mère âgée de 35 ans ou plus) ;

4 : (l'un des parents est célibataire, l'autre marié ou célibataire, fratrie de taille 3 ou 4, mère âgée de moins de 35 ans) ou (l'un des parents est célibataire, l'autre marié ou célibataire, fratrie de taille 5 ou plus) ;

5 : (l'un des parents est marié, l'autre veuf ou divorcé) ou (l'un des parents est célibataire, l'autre veuf ou divorcé, fratrie de taille 1 ou 2) ;

6 : l'un des parents est célibataire, l'autre veuf ou divorcé, fratrie de taille 3 ou plus ;

7 : père veuf ou divorcé, mère veuve ou divorcée, fratrie de taille 1 ou 2 ;

8 : père veuf ou divorcé, mère veuve ou divorcée, fratrie de taille 3 ou plus.

Trois catégories de communes ont été distinguées, croisant le critère de région et le critère de tranche d'unité urbaine :

1 : (communes des régions Champagne-Ardenne, Picardie, Basse-Normandie, Haute-Normandie, Centre, Bourgogne, Nord-Pas-de-Calais, Aquitaine, Midi-Pyrénées, Limousin) ou (communes des régions Alsace, Lorraine, Franche-Comté, Pays de Loire, Bretagne, Poitou-Charentes, Languedoc-Roussillon, PACA et Corse qui ne se trouvent pas dans une unité urbaine de 200 000 habitants ou plus) ;

2 : (communes des unités urbaines de 200 000 habitants et plus situées dans les régions Alsace, Lorraine, Franche-Comté, Pays de Loire, Bretagne, Poitou-Charentes, Languedoc-Roussillon, PACA et Corse) ou (communes d'Île-de-France, de Rhône-Alpes et d'Auvergne qui ne se trouvent pas dans une unité urbaine de 200 000 habitants ou plus) ;

3 : communes des unités urbaines de 200 000 habitants et plus situées dans les régions Île-de-France, Rhône-Alpes et Auvergne.

Le tableau 4 résume les probabilités d'appartenance aux familles biparentales selon leur type. Pour estimer les probabilités et produire les vrais effectifs communaux d'enfants par catégorie de famille, on doit bien entendu passer par une exploitation des unités statistiques « famille » du recensement. Or les données sur la famille ont la particularité d'être obtenues dans les petites communes auprès d'un sous-échantillon de l'échantillon des logements recensés, appelé « échantillon complémentaire ». Ce dernier résulte d'un échantillonnage de logements à probabilités égales dans l'ensemble des logements recensés de la commune, avec un taux de sondage égal à 1/4 (à l'exception de la Corse où le taux vaut 1). Dans toute grande commune, l'échantillon complémentaire est identique à l'échantillon recensé. De ce fait, dans les petites communes, on abandonne l'exhaustivité du traitement, ce qui dégrade sensiblement la qualité de l'estimation des probabilités issues du modèle et celle des effectifs communaux pondérant ces probabilités. En exploitant les fichiers des individus et des familles au recensement, on peut dénombrer, dans chaque famille, le nombre total d'enfants recensés de moins de 18 ans, ce qui permet de construire les pondérations des probabilités précédemment estimées. La construction de la catégorie de famille nécessite, dans le fichier du recensement (échantillon complémentaire) et pour les familles biparentales, d'identifier les adultes – père et mère – du couple avec lequel vit l'enfant. On utilise pour cela conjointement les variables de lien familial et de mode de cohabitation.

Les tailles d'échantillon répondant proviennent d'un comptage des familles, et non des enfants. Comme on dispose, grâce au *RP* 2009, du nombre total d'enfants (estimé) de moins de 18 ans dans la commune (résidant en ménage ordinaire) on effectue pour chaque commune un calage, par un simple ratio, afin que la somme des effectifs communaux estimés d'enfants appartenant aux trois types de familles redonne bien l'effectif communal issu du *RP*. Cette opération simple et efficace permet au passage de traiter le cas fort gênant d'environ 300 petites communes où le hasard a fait que l'échantillon complémentaire « famille » du recensement n'a inclus aucun enfant de moins de 18 ans alors que le *RP* 2009 en dénombre au moins un. Cette situation conduit évidemment à des dénombrements locaux égaux à zéro enfant par type de famille, ce qui est manifestement faux si on considère les résultats du *RP*. On a donc convenu, pour ces communes, de ventiler forfaitairement l'effectif communal total d'enfants du recensement entre les trois types de familles grâce à une clé de ventilation nationale (largement arrondie) issue des traitements nationaux *EFL* (70 % en famille

traditionnelle, 20 % en famille monoparentale, 10 % en famille recomposée).

Pour chacun des trois types de familles distingués, l'effectif d'enfants par commune ayant été obtenu, une ultime opération de ventilation de cet effectif a été effectuée afin de répartir les enfants entre trois classes d'âge : 0 à 5 ans, 6 à 10 ans, 11 à 17 ans. Cette répartition s'est appuyée sur une clé de répartition nationale issue de *EFL*, calculée de manière précise pour chaque type de famille et distinguant en sus les communes des agglomérations de 200 000 habitants et plus des autres communes. Une alternative consisterait à construire un modèle par tranche d'âge mais cela aurait été lourd informatiquement et la taille de l'échantillon exploitable se serait avérée trop faible dans certaines catégories de familles.

Les autres variables

La production communale utilisant une technique « petits domaines » a concerné six autres variables, dont les principes de la modélisation sont résumés succinctement ci-dessous.

Le nombre de personnes âgées ayant au moins un enfant résidant à proximité de leur domicile

Tableau 4
Par catégorie de communes et catégorie de famille, probabilité pour un enfant de moins de 18 ans en famille biparentale de vivre en famille recomposée ou traditionnelle

Catégorie de communes	Catégorie de famille	Nombre de familles répondantes	Probabilité d'appartenance (en %)	
			Famille recomposée	Famille traditionnelle
1	1	53 750	7,5	92,5
	2	46 848	12,7	87,3
	3	4 379	26,7	73,3
	4	2 127	35,4	64,6
	5	2 788	48,2	51,8
	6	724	76,4	23,6
	7	683	74,1	25,9
	8	331	86,5	13,5
	Ensemble	111 630	13,3	86,7
2	1	12 953	6,5	93,5
	2	11 448	10,4	89,6
	3	1 042	16,5	83,5
	4	420	35,9	64,1
	5	791	44,5	55,5
	6	187	82,3	17,7
	7	185	69,5	30,5
	8	86	60,9	39,1
	Ensemble	27 112	11,0	89,0
3	1	11774	5,4	94,6
	2	10130	10,4	89,6
	3	1546	23,6	76,4
	4	449	25,3	74,7
	5	639	38,0	62,0
	6	204	59,6	40,4
	7	178	59,4	40,6
	8	69	71,8	28,2
	Ensemble	24 989	10,1	89,9
Ensemble		163 731	12,4	87,6

Lecture : considérons un enfant dans une famille biparentale dont le père et la mère sont mariés (catégorie de famille 1). S'il réside dans une commune de catégorie 2, il y a 6,5 chances sur 100 pour qu'il vive dans une famille recomposée.
Champ : ensemble des individus de moins de 18 ans, vivant dans un ménage ordinaire de métropole, en famille biparentale.
Source : EFL *2011,* EAR *2011,* RP *2009, calculs de l'auteur.*

constitue un premier paramètre d'intérêt. Par résider à proximité, on entend soit dans la même commune, soit à moins de 30 minutes (temps compté en heures creuses). Pour caractériser la distance, on utilise un distancier qui quantifie les temps de déplacement entre communes. L'opération est probablement entachée d'erreurs de mesure significatives, compte tenu de l'âge des enquêtés concernés. En particulier, un nombre significatif de personnes âgées déclarent au recensement vivre seules, mais déclarent également dans le questionnaire *EFL* vivre avec au mois un enfant. Il a donc fallu gérer ces incohérences. Par ailleurs, la non réponse partielle était importante et nous avons procédé, pour chaque enfant, à une imputation spécifique de sa commune de résidence lorsqu'elle n'était pas précisée par l'enquêté. Au-delà de ces problèmes, le contexte n'est pas favorable à une modélisation efficace pour deux raisons. D'une part l'information auxiliaire disponible caractérise la personne âgée, alors que c'est celle de l'enfant qui est pertinente. D'autre part, on ne peut pas isoler dans le recensement les personnes ayant eu des enfants, si bien que le champ de la modélisation (voir encadré 1) doit être étendu ici à l'ensemble des personnes de 75 ans ou plus et vivant seules, sans pouvoir se limiter à celles qui ont eu des enfants. Finalement, le modèle a distingué quatre sous-populations en croisant la tranche d'âge et l'état matrimonial, ainsi que trois catégories de communes. Le modèle retenu pour dénombrer les personnes âgées résidant dans la même commune que l'un au moins de leurs enfants est par construction le même que celui qui dénombre les personnes âgées résidant à moins de 30 minutes de l'un au moins de leurs enfants. Dans les deux cas, on constate que les probabilités associées aux hommes sont très souvent nettement inférieures à celles des femmes. C'est étonnant et on peut soupçonner des erreurs de mesure qui traduiraient une différence de comportement entre les deux sexes face au questionnaire. Néanmoins, d'autres explications peuvent être apportées.

Les seconds paramètres d'intérêt sont le nombre de grands-pères et le nombre de grands-mères. L'enquête *EFL* demande en effet à l'enquêté(e) s'il (elle) a des petits-enfants. Le champ de la modélisation retenu est l'ensemble des personnes d'âge supérieur ou égal à 35 ans (car susceptibles d'avoir des petits-enfants). Le modèle a distingué onze catégories de populations, à partir de l'état matrimonial et de l'âge. Pour ces paramètres, l'impact de la géographie et du degré d'urbanisation est apparu assez fort, puisque quatre catégories de communes ont été distinguées.

La multi-résidence des enfants pour cause de vie avec un autre parent a constitué un domaine d'intérêt. On a donc dénombré les enfants de moins de 18 ans qui partagent leur temps entre deux logements pour vivre avec leur autre parent. Les facteurs explicatifs de la multi résidence ont été recherchés du coté de la structure de la famille (vivre en famille recomposée augmente considérablement la probabilité de se partager entre deux logements) et c'est pourquoi la démarche utilisée pour dénombrer les enfants selon leur type de famille a été largement reprise. Pour éviter les doubles comptes, les enfants ont été rattachés à leur résidence principale, information qu'il a donc fallu sécuriser en recoupant les réponses à différentes questions, dont certaines ont donné lieu au préalable à des imputations délicates. La modélisation a conduit à distinguer neuf catégories de familles et deux groupes de communes.

Il a été décidé d'estimer le nombre d'hommes et le nombre de femmes qui déclarent loger de façon « habituelle » dans un autre logement que leur résidence principale. Dans *EFL*, une question spécifique collecte cette information. Cette variable subit une non-réponse partielle assez forte, que l'on a considéré comme distribuée totalement au hasard parmi les répondants totaux (le traitement de la non-réponse totale ayant été effectué au préalable par repondération). Les individus sont interrogés au lieu de leur résidence principale, il n'y a donc pas de double-comptes. Les variables explicatives retenues *a priori* sont l'âge ainsi que l'indicateur de vie en couple. Le caractère explicatif de ces variables paraît naturel, les jeunes adultes pouvant résider ailleurs pour leurs études et les personnes plus âgées partager leur vie entre deux résidences au moment de la cessation d'activité, au moins jusqu'au moment où les déplacements deviennent difficiles. Par ailleurs, la capacité financière augmente souvent avec l'âge et permet donc d'acquérir plus facilement un second logement. Une tentative d'introduction de l'indicateur du lieu d'étude (codé au recensement) s'est avéré être un échec : de manière surprenante la variable n'était pas explicative. Par ailleurs, quatre catégories de communes ont été distinguées, comme d'habitude en croisant de manière spécifique la région et la tranche d'unité urbaine.

Dans chaque commune, le nombre de familles recomposées et le nombre de familles traditionnelles avec au moins un enfant mineur ont également été estimés. Le nombre de familles monoparentales peut être obtenu par une exploitation directe du recensement. La démarche est

très semblable à celle qui a permis de dénombrer les enfants qui vivent dans ces familles, mais cette fois l'unité statistique est la famille, d'un bout à l'autre de la procédure : on caractérise une famille par les statuts matrimoniaux des parents, la taille de la fratrie et, fait original, les classes d'âge du père et de la mère. On aboutit à sept catégories de familles dont la définition complète est complexe (à l'image de ce que l'on a pour les enfants par type de famille, ainsi que pour les enfants résidant dans deux logements), réparties selon trois catégories de communes. Le modèle nous dit qu'une famille donnée biparentale avec au moins un enfant mineur a une probabilité d'être recomposée qui ne dépend pas de sa commune de résidence (hormis au travers de la catégorie de communes), mais seulement de sa catégorie de famille. Le nombre total de familles par catégorie dans chaque commune est bien entendu estimé à partir du recensement.

Enfin, la dernière variable traitée représente les nombres d'hommes et de femmes dans la commune se déclarant en couple mais sans vivre (ou sans être recensé) avec leur conjoint. On repère très simplement dans *EFL* la vie en couple non cohabitant dès lors qu'à la question « *Êtes-vous actuellement en couple ?* » l'enquêté coche la modalité « *Oui, avec une personne qui vit dans un autre logement* ». Le traitement de cette variable ressemble beaucoup à celui des personnes pacsées, avec une originalité de champ de même nature. En effet, pour gérer les biais dus aux erreurs de déclaration concernant la vie en couple, ce qui crée des incohérences entre la déclaration *EFL* et la déclaration *EAR*, le champ de calcul des probabilités est constitué par l'ensemble des adultes (par sexe) et non par l'ensemble des adultes déclarant vivre en couple au recensement (alors que cette variable est disponible dans l'*EAR*). C'est exactement le procédé qui a été utilisé pour dénombrer les personnes pacsées tout en gérant les incohérences affectant l'état matrimonial. Au-delà donc de la prise en compte de l'indicateur de vie en couple au recensement, les catégories de population explicatives de la vie en couple non cohabitant ont été construites à partir de l'âge (en 6 tranches) et de l'état matrimonial de l'individu enquêté. Cela conduit à définir finalement vingt catégories de population, réparties en trois catégories de communes. Comme l'âge est assez détaillé, les catégories de population sont nombreuses, mais leur significativité est confirmée par le modèle. Ce n'est pas un problème parce que la taille de l'échantillon répondant par catégorie de communes reste très importante (ce qui explique aussi la significativité des variables retenues…). La suite relève de la philosophie habituelle.

Concernant les trois premiers paramètres présentés au début de cette partie, le lecteur intéressé par des compléments et par les résultats chiffrés de la modélisation pourra consulter (Ardilly, 2015).

Dernière étape : la production des effectifs communaux

Une fois les probabilités estimées par catégorie de population (cf. encadré 1), la pondération de chacune de ces probabilités par les effectifs adéquats du recensement 2009 produit des estimations communales (cf. encadré 2). Parmi les traitements qui permettront d'achever le processus, il faut évoquer deux opérations spécifiques très importantes. Auparavant, il convient de préciser que pour chaque variable restituée, les effectifs communaux ont été estimés pour l'ensemble des communes de métropole, qu'elles aient ou non participé à l'opération *EFL*. En effet, si la diffusion a été par principe restreinte aux seules communes échantillonnées ayant accepté de participer à l'enquête, l'estimation des effectifs au niveau de chaque commune de métropole (en particulier toutes celles qui n'ont pas été échantillonnées) permet d'apprécier à un niveau global l'ampleur de l'erreur due à l'utilisation d'un modèle, ce qui est un atout précieux.

La première opération est celle du vieillissement des données du recensement de 2009 à 2011. Nous avons vu que les vrais effectifs communaux par catégorie de population sont assis sur les données du recensement de 2009. Or l'*EFL* prétend représenter une situation au 1er janvier 2011. Les deux années de retard du recensement doivent donc être comblées pour actualiser les structures de pondération. Pour chacune des grandes communes (plus de 10 000 habitants), on disposait en mars 2013 d'une série chronologique de populations municipales annuelles et cela jusqu'en 2010 compris. Une estimation de la population municipale 2011 a alors été effectuée par l'équipe en charge du recensement, en tenant compte des éléments récents dont elle disposait provenant de l'*EAR* 2013. Dans les grandes communes, on a calculé le taux d'évolution de la population municipale entre 2009 et 2011 et on a appliqué ce taux aux estimations communales associées à toutes les variables

restituées, considérant ainsi que l'on avait fait « vieillir » au mieux les estimations communales. Pour les petites communes le contexte est très différent. Lors des calculs d'estimation d'effectifs communaux, tous les poids utilisés à partir du recensement ont systématiquement été forcés à la valeur 1. De fait, la situation de chaque petite commune est exactement celle qui reflète son année de recensement, laquelle varie entre 2007 et 2011. Ne produire des effectifs communaux que pour les communes participant à l'*EFL* nous aurait épargné toute opération d'actualisation des effectifs par catégorie parce que, par définition, il s'agit de petites communes qui sont toutes recensées en 2011, donc en coïncidence parfaite avec la date de référence des estimations. La nécessité (nous le verrons un peu plus loin) de produire des estimations sur chaque commune de métropole impose d'enclencher également un processus de vieillissement des estimations des petites communes. Pour les petites communes, il n'y a pas eu de prévision de la population municipale 2011. On a donc procédé par chaînage : considérant l'année de recensement de la petite commune, on a actualisé jusqu'en 2010 ses estimations communales en appliquant les taux d'évolution annuels de la population municipale. Puis, pour passer de 2010 à 2011, on a reconduit l'évolution enregistrée entre 2009 et 2010. Évidemment, cette technique peut paraître un peu grossière et peut-être même inadaptée dans certains cas où l'évolution récente de la démographie communale est perturbée mais il faut considérer, d'une part que les méthodes alternatives ne trouvent pas davantage de justification, d'autre part qu'il existe une multitude d'incertitudes, ne serait-ce que les procédures d'arrondi mises en œuvre pour les besoins de la diffusion (voir infra), qui ont des conséquences numériques allant sans doute bien au-delà de l'erreur due au processus de vieillissement.

La seconde opération est essentielle, car non seulement elle constitue un outil simple permettant, avec d'autres, d'apprécier la qualité de l'ensemble du processus, mais encore elle contribue à améliorer sensiblement cette qualité. Il s'agit ici de procéder à une ultime opération de calage, que l'on a coutume d'appeler « *benchmarking* ». L'idée est la suivante. On dispose d'une enquête nationale (*EFL*) de (très) grosse taille qui fournit, grâce à sa pondération optimisée par une étape de redressement, des estimations nationales que l'on peut qualifier de (très) précises (Insee, 2013). Par ailleurs, on a procédé, pour chaque commune du territoire métropolitain (36 603 communes exactement[13]), à une estimation locale, utilisant la méthodologie *ad hoc* précédemment décrite. Cette méthodologie est basée sur un modèle et génère donc des erreurs spécifiques supplémentaires, de manière inévitable puisqu'un modèle n'est autre qu'une hypothèse simplificatrice de la réalité. Si on somme les estimations communales, on obtient une estimation nationale qui peut être comparée à l'estimation nationale provenant directement de l'*EFL* (donc sans l'intervention des données du recensement, hormis au stade ultime du redressement) : si l'écart entre les deux effectifs est important, parce que les erreurs de modèle ont souvent des conséquences essentiellement systématiques, on peut craindre un biais fort (au sens propre du terme « biais », c'est-à-dire un écart entre une espérance mathématique et une vraie valeur), sinon on pourra penser que le biais reste modéré et que le modèle est somme toute acceptable. Au-delà de la simple constatation des écarts numériques, il est souhaitable de les utiliser à notre profit. C'est pourquoi, puisqu'on fait confiance à l'estimation nationale *EFL*, il est naturel de faire disparaître cet écart, c'est-à-dire d'effectuer une modification de chacune des estimations locales pour que leur somme redonne *in fine* l'estimation nationale. L'opération aura en outre l'excellente vertu de réconcilier les estimations locales diffusées et l'estimation nationale diffusée, c'est-à-dire que les utilisateurs qui disposeront de toute l'information ne constateront pas d'incohérence avec le chiffrage national lorsqu'ils totaliseront les estimations locales.

En la circonstance, on peut même faire mieux. Puisqu'on a raisonné pour chaque variable sur un champ de population spécifique (qui varie d'une variable à l'autre), et que l'effectif de ce champ est connu au niveau communal grâce au recensement (après vieillissement des structures…), on peut effectuer le calage avec l'objectif de respecter à la fois l'effectif national *EFL* et les effectifs communaux du champ considéré, donnés par le *RP*. Prenons l'exemple concret de l'effectif des hommes pacsés : la sommation des estimations communales donne 659 000 et *EFL* conduit à une estimation de 701 000. Par ailleurs, dans chaque commune on dispose d'une estimation 2011 du nombre total d'hommes vivant en couple. Il convient de trouver des effectifs estimés définitifs par commune pour que l'on estime à 701 000 le

13. Les arrondissements de Paris, Lyon et Marseille sont distingués ; en outre, il y a eu quelques opérations très acrobatiques pour tenir compte de modifications géographiques (fusions de communes nécessitant une harmonisation préalable des différentes bases de données communales exploitées, certaines donnant la situation avant fusion et d'autres après fusion).

nombre total d'hommes pacsés tout en respectant dans chaque commune les effectifs des hommes vivant en couple estimés par le recensement. Un algorithme de type *raking-ratio* a été programmé en ce sens (Ardilly, 2006), qui procède par un enchainement de 'règles de trois' réalisant alternativement le calage national et le calage communal, jusqu'à convergence du processus.

Ce processus de *benchmarking* a été appliqué au printemps 2013 sur la base d'estimations nationales de l'*EFL* provisoires, certains traitements d'imputation n'ayant pas encore été effectués au niveau national à cette époque. Son bon fonctionnement nécessite que le nombre des individus du champ au niveau national soit estimé exactement à la même valeur par *EFL* et par le recensement que l'on a fait vieillir auparavant. Pour réconcilier les sources, et avec l'objectif essentiel de ne pas créer d'incohérence avec l'estimation nationale *EFL*, on prend le parti de retenir comme cible l'effectif du champ estimé par *EFL* et, par un simple ratio appelé 'coefficient correcteur', d'adapter les effectifs communaux du champ issus du recensement dans chaque commune. Pour chaque variable diffusée, ce coefficient correcteur est fourni dans le tableau 5.

Tableau 5
Appréciation des erreurs de modèle, avant *benchmarking*, par variable diffusée

Variable d'intérêt	Estimation nationale EFL (1)	Sommation des estimations communales	Erreur relative (1) (en %)	Coefficient correcteur (2)
Nombre d'hommes cohabitants pacsés (et non-mariés par la suite)	701 000	659 000	- 6,0	0,989
Nombre de femmes cohabitantes pacsées (et non-mariées par la suite)	659 000	583 000	- 11,5	0,990
Nombre d'hommes en couple, non cohabitants	602 000	689 000	14,4	0,999
Nombre de femmes en couple, non cohabitantes	714 000	788 000	10,4	0,999
Nombre de grands-pères, moins de 75 ans	4 552 000	4 268 000	- 6,2	1,004
Nombre de grands-mères, moins de 75 ans	6 035 000	5 810 000	- 3,7	1,002
Nombre de grands-pères, 75 ans ou plus	1 584 000	1 495 000	- 5,6	1,033
Nombre de grands-mères, 75 ans ou plus	2 547 000	2 427 000	- 4,7	1,025
Nombre d'hommes âgés de 75 ans ou plus, vivant seuls, ayant au moins un enfant résidant dans la même commune	111 000	100 000	- 9,9	1,062
Nombre de femmes âgées de 75 ans ou plus, vivant seules, ayant au moins un enfant résidant dans la même commune	625 000	671 000	7,4	1,027
Nombre d'hommes âgés de 75 ans ou plus, vivant seuls, ayant au moins un enfant résidant à moins de 30 minutes	216 000	174 000	- 19,4	1,062
Nombre de femmes âgées de 75 ans ou plus, vivant seules, ayant au moins un enfant résidant à moins de 30 minutes	1 103 000	1 025 000	- 7,1	1,027
Nombre d'enfants de moins de 4 ans gardés par les parents	1 652 000	1 608 000	- 2,7	1,027
Nombre d'enfants de moins de 4 ans gardés par les grands-parents	177 000	158 000	- 10,7	1,027
Nombre d'enfants de moins de 4 ans gardés en crèche	397 000	430 000	8,3	1,027
Nombre d'enfants de moins de 4 ans gardés par une assistante maternelle	816 000	749 000	- 8,2	1,027
Nombre d'enfants de moins de 4 ans gardés par un autre mode	97 000	106 000	9,3	1,027
Nombre d'enfants de moins de 18 ans vivant en famille monoparentale	2 450 000	2 478 000	1,1	0,986
Nombre d'enfants de moins de 18 ans vivant en famille recomposée	1 476 000	1 460 000	- 1,1	0,986
Nombre d'enfants de moins de 18 ans vivant en famille traditionnelle	9 774 000	9 941 000	1,7	0,986
Nombre de familles traditionnelles avec au moins un enfant mineur	5 473 000	5 476 000	0,1	0,992
Nombre de familles recomposées avec au moins un enfant mineur	723 000	725 000	0,3	0,992
Nombre d'hommes résidant dans deux logements	2 701 000	2 821 000	4,4	0,999
Nombre de femmes résidant dans deux logements	2 767 000	2 809 000	1,5	0,999
Nombre d'enfants de moins de18 ans résidant dans deux logements pour vivre avec leur autre parent	905 000	835 000	- 7,7	0,998

1. Situation définitive.
2. Situation provisoire (printemps 2013).

*Lecture : l'*EFL *estime à 701 000 le nombre d'hommes pacsés en métropole, contre 659 000 lorsqu'on somme les estimations communales avant* benchmarking*, ce qui conduit à une erreur relative de - 6 %. Par ailleurs, chaque effectif communal du champ (nombre d'hommes vivant en couple) a été multiplié par 0,989 avant d'engager le processus de calages successifs.*
Source : EFL *2011,* EAR *2011,* RP *2009, BPE 2010, fichier « Particuliers-employeurs » 2010, distancier Odomatrix Inra Umr 1041 Cesaer, calculs de l'auteur.*

Un premier outil d'appréciation du biais …

Une première façon d'apprécier le biais généré par la méthode « petits domaines » qui vient d'être décrite consiste à calculer une erreur relative exprimée en pourcentage (on calcule d'abord la différence entre la somme des estimations communales et l'estimation nationale *EFL*, puis on la divise par l'estimation nationale *EFL*). Le tableau 5 fournit pour chaque variable d'intérêt l'erreur relative obtenue avant *benchmarking* par rapport à l'estimation nationale définitive. Le coefficient correcteur (qui n'influe pas sur la somme des estimations communales puisqu'on se place ici avant *benchmarking*) est celui qui a été utilisé pour produire les estimations communales diffusées. Il a donc un caractère provisoire puisqu'il assure la cohérence avec les estimations nationales *EFL* du printemps 2013 (on notera que les champs qui ont été définis concernent des populations suffisamment faciles à identifier pour que les révisions qui ont eu lieu pour certaines estimations *EFL* n'aient qu'un effet minime sur les coefficients correcteurs correspondants). Les estimations nationales issues de l'*EFL* fournies dans ce tableau sont les estimations définitives : elles ont évolué par rapport aux estimations nationales utilisées pour le *benchmarking*. Pour la majorité des variables, ces évolutions sont numériquement (très) faibles et sans conséquence. Néanmoins, dans le cas des trois variables de multi-résidence, ainsi que pour les quatre variables de dénombrement des personnes âgées vivant seules et dont au moins un enfant réside à proximité, la correction s'est avérée non négligeable dans le sens d'une amélioration substantielle de l'adéquation du modèle. Il est rappelé que les quatre dernières variables citées sont fortement affectées par la non-réponse partielle concernant la localisation des enfants, ce qui rend l'estimation nationale elle-même assez fragile. Le lecteur pourra trouver dans Acs (2013), Bailly et Rault (2013), Blanpain et Lincot (2013), Buisson et Lapinte (2013), Domingo (2013) et Lapinte (2013) des études de référence permettant de retrouver les ordres de grandeur des estimations nationales présentées dans ce tableau et de compléter sa connaissance générale des sujets abordés.

Ce tableau constitue un outil d'appréciation de la qualité de l'estimation communale et il convient de s'en faire une opinion en tenant compte des deux commentaires suivants.

Le premier commentaire concerne les coefficients correcteurs : les valeurs éloignées de 1 traduisent un écart significatif entre l'effectif national du champ estimé par l'*EFL* sur la base des données provisoires et l'effectif national du champ donné par le recensement. C'est une situation qui survient à plusieurs reprises et qui est évidemment défavorable à la qualité des estimations. Les causes peuvent être multiples, mais on pense essentiellement au processus de vieillissement du recensement, qui est un peu brutal, et à un décalage des concepts définissant le champ entre d'une part l'*EFL* et d'autre part le *RP*. Par exemple, si l'on considère les personnes âgées vivant seules, encore faut-il que ce que l'on entend par « vivre seul » soit traduit de la même façon dans l'*EFL* et dans le recensement. L'erreur d'échantillonnage de l'*EFL* peut aussi y contribuer, d'autant plus que l'on dénombre une population plus petite : par exemple, il ne devrait y avoir que peu d'ambiguïté sur ce qu'on appelle « enfant de moins de 4 ans » mais il y a néanmoins sur ce champ plutôt simple une différence d'estimation nationale de 2,7 % entre les deux sources (cf. tableau 5). Ces décalages peuvent expliquer une part non négligeable de l'erreur relative : par exemple lorsqu'on dénombre les hommes de plus de 75 ans vivant seuls et ayant au moins un enfant situé à moins de 30 minutes de leur domicile, comme il y a (cf. tableau 5) 6,2 points de pourcentage attribuables à un décalage dans le dénombrement du champ (les hommes de 75 ans et plus vivant seuls), on peut dire que l'erreur spécifiquement due au modèle n'est pas égale à - 19,4 % mais qu'elle se situe plutôt aux alentours de - 15 %.

Le second commentaire concerne les erreurs relatives, qui sont à la fois l'ultime étape de toute la procédure et le seul indicateur quantitatif (simple) que l'on obtienne pour juger de la qualité d'ensemble des estimations locales. On constate la diversité des situations, les erreurs relatives – dont on rappelle le caractère fragile du calcul pour certaines d'entre elles – variant en valeur absolue de (presque) zéro à (presque) 20 %. L'utilisateur habitué aux coefficients de variation nationaux des grandes enquêtes par sondage (souvent situés entre 1 et 5 points de pourcentage) pourrait percevoir comme excessives une bonne moitié de ces erreurs. Nous pensons au contraire qu'il faut s'en satisfaire, pour les raisons suivantes :

- il est essentiel d'apprécier les erreurs en abandonnant les standards de qualité des estimations nationales, car bien entendu il faut comparer ce qui est comparable. En la circonstance, c'est le résultat d'un exercice extrêmement acrobatique qui a fait prendre tous les risques. Il a

fallu en effet faire face à un cumul d'obstacles : *primo*, sauf pour quelques rares cas les zones d'estimation (les communes) sont extrêmement petites ; *secundo*, l'échantillonnage souffre par construction d'un fort effet de grappe qui en dégrade l'efficacité ; *tertio*, on s'est intéressé à des sous-populations qui, pour certaines d'entre elles, sont rares et/ou assez compliquées à isoler ; quarto, l'information disponible pour expliquer des phénomènes qui, pour la plupart, relèvent de déterminants complexes est celle du recensement, c'est-à-dire une information sociodémographique assez restreinte en gamme et dont le pouvoir explicatif reste limité ; enfin, on ne dispose pour la pondération des probabilités d'aucun effectif estimé sans erreur dans les grandes communes (puisque le recensement n'y est pas exhaustif), ainsi que dans les petites communes à chaque fois que l'unité « famille » entre en jeu.

- d'autres expériences menées sur d'autres sources et avec d'autres variables conduisent à des erreurs d'ordre de grandeur comparables. On trouvera des indicateurs chiffrés dans Ardilly (2014), Ardilly (2012), Platek *et al.* (1987), ESSnet on Small Area Estimation (2012) ou encore IASS Satellite Conference (1999). Il semblerait donc que l'on soit coutumier de tels décalages dès qu'on applique des modèles portant sur des sous-populations de petite taille.

- s'il faut se donner des normes, on peut considérer qu'une erreur inférieure à 5 % est très satisfaisante, qu'elle est satisfaisante entre 5 % et 10 %, acceptable entre 10 % et 15 % et dégradée au-dessus de 15 %. Les variables exploitées connaissent sous cet angle des fortunes diverses : les moins avantagées concernent le nombre d'hommes âgés vivant seuls à proximité d'un au moins de leurs enfants, et, dans une moindre mesure , le nombre de personnes en couple non cohabitantes. Les autres se situent dans un périmètre d'erreur très acceptable. Le cas de la première variable est tout à fait conforme à l'intuition : d'une part il s'agit d'effectifs dont l'estimation nationale est elle-même rendue instable par la forte non-réponse partielle, d'autre part les effectifs nationaux de la sous-population concernée et de l'échantillon sont très faibles, ce qui va mécaniquement dans le sens d'une plus grande erreur relative.

Ces écarts relatifs sont peu préjudiciables : s'ils traduisent un risque de biais sensible de certains estimateurs locaux, ils se rapportent à la situation avant *benchmarking*, donc à une situation intermédiaire qui n'inclut aucune correction de biais. En revanche, les estimations finales qui ont été diffusées intègrent le *benchmarking*, qui par construction rend toutes ces erreurs relatives nulles ! Bien sûr, cette correction du biais au niveau national n'annule pas les biais des estimateurs locaux, mais elle en neutralise au moins la composante de nature systématique due à une modélisation imparfaite.

L'analyse des erreurs est très compliquée parce qu'elles résultent d'une superposition de causes. Face à une erreur relative plutôt forte, on est tenté de revenir à la modélisation et à cette occasion on perçoit l'importance des choix empiriques qui ont participé à la modélisation. Si la sélection des variables explicatives ne pose en général pas trop de problème, celui du regroupement initial de leurs modalités détaillées est autrement plus déroutant (construire une classe d'âge par exemple). Le parti pris d'utiliser un même modèle pour les hommes et pour les femmes peut aussi être critiqué, car de ce fait il faut faire des compromis en considérant les résultats de deux régressions différentes. Quant à la construction des catégories de population définitives, elle s'appuie certes sur des tests mais ils sont loin de donner lieu à un processus de construction mécanique. En particulier, les périmètres des sous-populations constituant ces catégories sont construits par regroupement des modalités dont les coefficients de régression associés sont « semblables » (par exemple deux classes d'âge associées à deux coefficients semblables, qui sont regroupées en une seule), décision qui relève en soi d'une approche empirique. Tout cela est effectué en tenant compte en parallèle des tailles des échantillons par sous-catégorie (pour éviter qu'elles ne soient trop petites, ou du moins pour limiter le nombre des petites catégories), et aussi un peu de l'interprétation des regroupements. Sur ce dernier point, il faut bien reconnaitre que nous n'avons pas été trop exigeant, comme en témoigne l'existence de regroupements parfois assez hétéroclites, mais nous nous sommes laissé porter, jusqu'à un certain point, par les résultats de l'outil statistique de régression logistique. Cette stratégie peut être contestée puisqu'on aurait pu s'imposer de ne conserver que des catégories ayant une interprétation naturelle. Enfin, il faut rester prudent dans l'appréciation de la qualité à la lumière de la seule erreur relative, qui n'est jamais qu'une grandeur impliquant une référence nationale (et non locale) et qui peut donc cacher des compensations subtiles d'erreurs au niveau communal. Cette remarque est importante et traduit bien la limite des outils d'appréciation de la qualité de la modélisation

sur la statistique locale diffusée : la faiblesse de l'information locale collectée (donc la très petite taille d'échantillon) représente un risque incompressible qui ne peut pas être contourné au niveau d'une commune. On est bien dans un système d'estimation dépendante d'un modèle : utiliser ces données à un niveau finement localisé reste un acte de foi… Si l'on dispose des indicateurs descriptifs d'alerte détaillés ci-dessus, ils ne constituent en aucun cas des preuves. Surtout, ils s'apprécient de manière globale, c'est-à-dire sur l'ensemble des communes, et non commune par commune. Autrement dit, lorsqu'on parle d'erreur due au modèle, il faut toujours comprendre qu'il s'agit d'une erreur au niveau national et non pas d'une erreur pour la commune considérée.

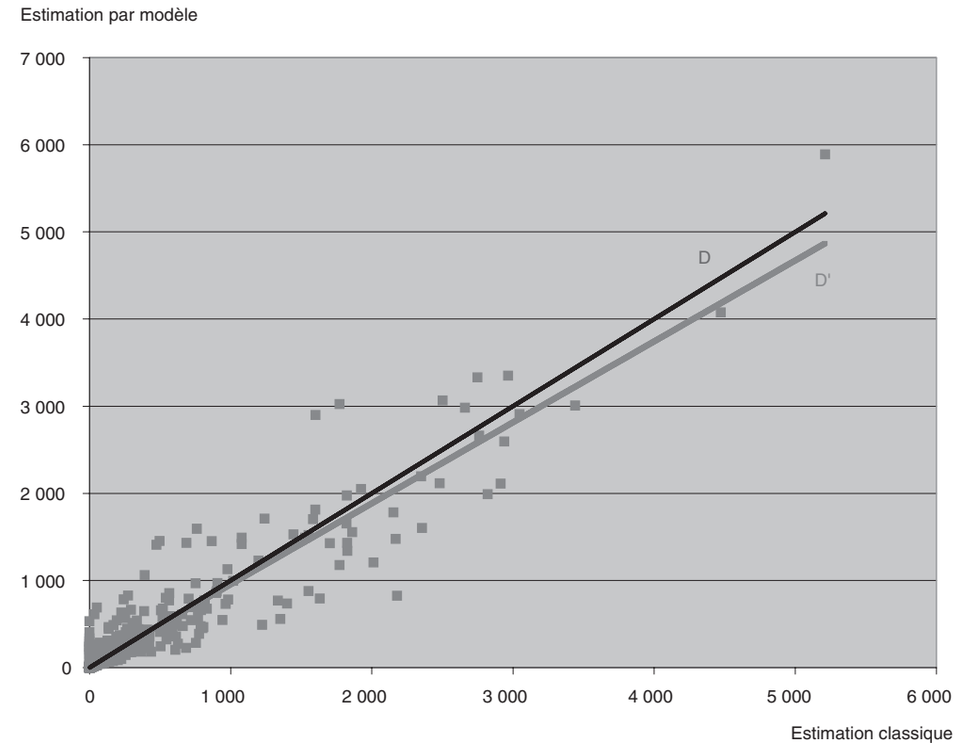
… complété par deux autres méthodes

À titre de second outil de contrôle, on a produit pour chaque variable un nuage de points dont chaque point représente une commune participant à l'*EFL*. On porte sur l'axe des abscisses l'estimation obtenue au moyen de l'estimateur classique limité à la commune (en la circonstance un estimateur par le ratio utilisant seulement les données échantillonnées dans la commune et calé sur les effectifs recensés pour chaque sous-population définie par le modèle[14]). L'ordonnée est l'estimation localisée obtenue avec notre méthode avant *benchmarking*. On rappelle que la première estimation est sans biais (ou le biais est faible) mais de grande variance d'échantillonnage. C'est le contraire pour la seconde. Un nuage de points distribué à peu près symétriquement autour de la première bissectrice (droite de pente 1, passant par l'origine) laisse fortement présager une absence de biais. Inversement, bien que ce ne soit pas systématique, l'absence de symétrie coïncide souvent avec un biais d'estimation. Examinant par exemple les cas respectifs du nombre d'hommes pacsés et du nombre de femmes résidant habituellement dans deux logements, on constate lorsqu'on s'intéresse à l'ensemble des communes participant à l'*EFL* que la symétrie s'avère plus apparente pour la variable « pacs » que pour la multi résidence (cf. figures I et II), alors même que les erreurs relatives du tableau 5 laissaient augurer une situation inverse (ces résultats ne représentent que deux approches empiriques et simplificatrices d'une réalité complexe). La droite D représente la première bissectrice et la droite D' la droite de régression standard du nuage de points. Par ailleurs, un zoom sur la partie du nuage limitée aux communes pour lesquelles l'estimation classique est plutôt petite fait apparaître une dissymétrie. Celle-ci est marquée dans le cas de la multi-résidence des femmes (pour obtenir un tel grossissement, on a retenu une borne supérieure de 2 000 femmes en multi résidence ; cf. annexe 2, figure) : le modèle a manifestement une tendance assez nette à augmenter l'estimation des effectifs lorsque l'estimation classique est petite. On trouve en particulier des communes où l'estimation classique est nulle mais l'estimation par modèle relativement forte. Ce phénomène est assez fréquent et semble traduire le mécanisme mécaniquement uniformisateur de la modélisation : la modélisation simplifie et la simplification rapproche un peu plus les comportements individuels du comportement moyen. La situation extrême serait atteinte si on considérait que seule la variable constante est explicative : par hypothèse, toutes les proportions communales seraient alors considérées comme constantes, toutes seraient égales à la proportion nationale et si en sus toutes les communes avaient la même taille de population, alors le nuage de points serait étalé sur une droite parfaitement horizontale. *A contrario*, un phénomène symétrique se produit à l'autre extrémité du nuage, du coté des plus grandes communes : là aussi, les estimations par modélisation reflètent un certain tassement. Ce tassement a reçu le nom de *shrinkage* (« rétrécissement » des distributions) lorsque l'on passe, sur un grand nombre de petits domaines, d'une distribution d'estimations classiques (sans modèle) à une distribution d'estimations par modélisation. Ce type de graphique a été produit systématiquement pour chaque variable diffusée : à aucun moment il n'est apparu de configuration inquiétante laissant présager un biais fort. On rappelle que le *benchmarking* est effectué ultérieurement.

Le troisième outil consiste à cartographier les résultats pour apprécier la corrélation spatiale des phénomènes mesurés, cette fois sur les données définitives après *benchmarking*. Normalement, on doit obtenir des cartes qui font ressortir une forme de similitude entre des communes voisines et de même type. C'est par exemple le cas en région Rhône-Alpes pour la proportion d'enfants de moins de 4 ans gardés en crèche parmi l'ensemble des enfants

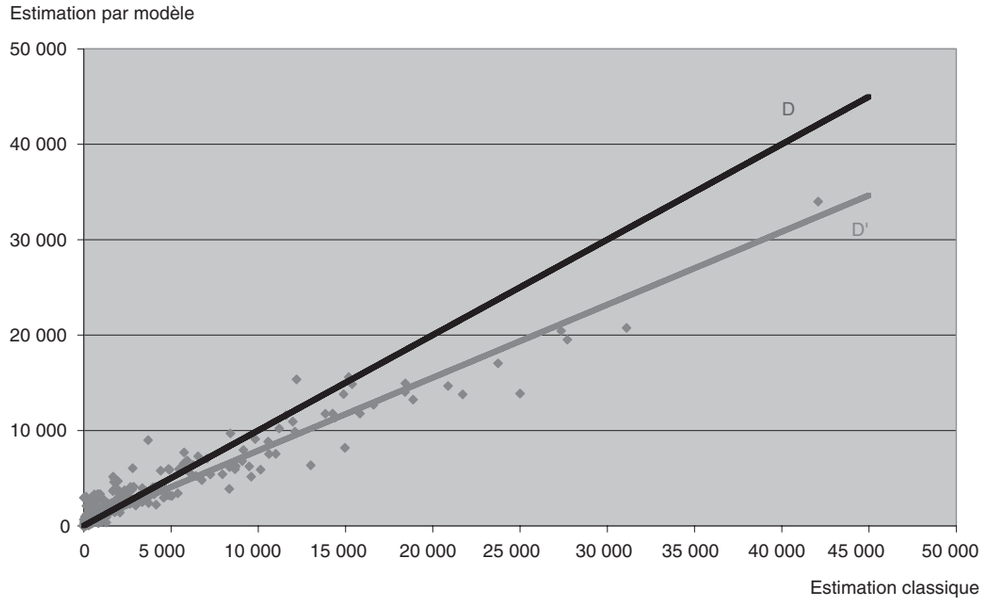
14. *En reprenant les notations de l'encadré 2, il s'agit de* $\sum_{h=1}^{H} N_{com}^{ch,h} \cdot \frac{\widehat{N}_{com}^{ch,h,ssp}}{\widehat{N}_{com}^{ch,h}}$.

Figure I
Estimations classiques et par modèle du nombre d'hommes pacsés : détection du biais



Lecture : une commune est représentée par un point dont l'abscisse est l'estimation du nombre d'hommes pacsés selon l'approche par pondération classique et dont l'ordonnée est l'estimation du nombre d'hommes pacsés selon la méthode proposée utilisant un modèle.
Champ : ensemble des communes participant à l'EFL où les hommes sont enquêtés.
Source : EFL *2011,* EAR *2011,* RP *2009, calculs de l'auteur.*

Figure II
Estimations classiques et par modèle du nombre de femmes en multi résidence : détection du biais



Lecture : une commune est représentée par un point dont l'abscisse est l'estimation du nombre de femmes occupant habituellement deux logements selon l'approche par pondération classique et dont l'ordonnée est l'estimation du nombre de femmes occupant habituellement deux logements selon la méthode utilisant un modèle proposée dans le texte.
Champ : ensemble des communes participant à l'EFL où les femmes sont enquêtées.
Sources : EFL *2011,* EAR *2011,* RP *2009, calculs de l'auteur.*

de moins de 4 ans (cf. annexe 2, carte I) ; on parvient de nouveau à détecter à vue ces corrélations dans le cas d'une variable davantage suspectée de biais, à savoir la proportion d'hommes en couple mais non cohabitant parmi l'ensemble des hommes adultes (cf. annexe 2, carte II). En particulier, dans les deux cas, les unités urbaines de la région ont tendance à se distinguer assez nettement.

Ces deux outils restent des outils graphiques : ils ne constituent pas des preuves à proprement parler, mais ils sont très utiles pour conforter la méthode, avec l'aide que constitue l'erreur relative du tableau 5. Quant à la restitution finale, il a été décidé d'appliquer une politique de diffusion d'estimations arrondies de tous les effectifs. Ainsi, on lisse quelque peu les estimations et on évite de donner aux utilisateurs une impression de précision illusoire. Cela permet également de renforcer le caractère anonyme des données, dans la mesure où certaines communes peuvent être de petite taille.

Il reste à préciser qu'il n'y a pas eu de calcul de variance d'échantillonnage relatif aux estimateurs locaux. D'une part ce n'est vraisemblablement pas l'essentiel de l'erreur, le problème se situant en premier lieu du coté du biais. D'autre part, les calculs seraient très compliqués à mener, peut-être même inextricables dans les conditions extrêmes où l'on se trouve (souvent une seule grappe *EFL* tirée dans la commune, estimateur compliqué avec un échantillonnage qui l'est également, gestion du cas où l'un des sexes n'est pas enquêté dans la commune, erreurs d'échantillonnage propres à l'utilisation des effectifs estimés avec le *RP*, etc.). Néanmoins, par sécurité, la programmation a fourni pour chaque variable la distribution des ratios communaux des effectifs estimés sur les populations municipales pour l'ensemble des communes de métropole, afin de détecter d'éventuelles concentrations de valeurs atypiques. Il n'en a rien été, tous les ratios ont un ordre de grandeur raisonnable et les coefficients de variation s'avèrent en général inférieurs à 1, souvent compris entre 0,3 et 0,6, les rapports entre le quantile d'ordre 95 % et le quantile d'ordre 5 % prenant majoritairement des valeurs de l'ordre de 3 ou 4, s'élevant au pire jusqu'à 10.

* *
*

Une estimation d'effectifs communaux relatifs aux individus physiques n'a à notre connaissance jamais été tentée par l'Insee à partir d'une enquête nationale par sondage – à l'exception évidemment du recensement lui-même (encore s'agissait-il alors d'échantillons exhaustifs ou de très grande taille). En cela, il s'agit d'une opération inédite et susceptible d'être jugée audacieuse. Malgré les critiques que l'on peut formuler sur la gamme relativement modeste des données recensées, le recensement s'avère ainsi capable de produire un système d'information indispensable à la constitution des estimations communales. À ce jour, dans l'attente du développement de répertoires principalement issus des fichiers fiscaux, quelle source pourrait s'y substituer ? Dans ce type d'approche, les hypothèses qui se cumulent et les aléas qui s'additionnent génèrent des erreurs de diverses natures qui finissent par rendre l'estimation délicate : on ne peut que se contenter d'approximations. Du moins ces approximations ont-elles subi l'épreuve du terrain : aucune des 1 438 restitutions personnalisées diffusées aux élus locaux au cours de l'été 2013 n'a à ce jour reçu de critique, alors même qu'il s'agit d'acteurs qui connaissent très bien leurs populations et dont on redoutait l'aptitude à déceler des invraisemblances. Quant à l'esprit de la méthode, expliquée dans le document envoyé aux élus, il n'a pas soulevé non plus de désapprobation, ce qui est remarquable car le fait d'assimiler le comportement moyen d'une commune, même si on se restreint à des sous-populations, à celui d'un ensemble d'autres communes présentées comme similaires aurait pu apparaître incongru aux élus : en effet, il est bien connu qu'on ne ressemble jamais vraiment à son voisin… Aux détracteurs de la méthode, il faut répondre par la question : quelle est l'alternative ? Ne rien produire bien sûr, mais si on doit le faire peut-on raisonnablement accepter, même dans des villes de taille respectable, de s'en tenir aux seules données recueillies dans le périmètre communal ? Alors même que l'enquête *Famille et logements* n'a jamais prétendu à la moindre représentativité communale, ni même régionale d'ailleurs, il faudrait expliquer pourquoi l'estimation que produirait l'application de la théorie classique des sondages ne comptabilise aucune femme pacsée à Saint-Étienne, aucune grand-mère de moins de 75 ans à Aurillac, aucun enfant de moins de 4 ans gardé par une assistante maternelle à Cergy, aucune enfant gardé en crèche à Lorient ou encore aucune femme résidant à Béziers mais vivant habituellement dans deux logements distincts … □

BIBLIOGRAPHIE

Acs M. (2013), « Les spécificités régionales des modes de garde déclarés des enfants de moins de 3 ans », *Études et Résultats*, n° 839, Drees, Ministère des affaires sociales, de la santé et du droit des femmes.

American Community Survey, Design and Methodology (2014), United States Census Bureau.

Ardilly P. (2015), « Estimations communales exploitant les données de l'enquête *Famille et Logements* et du recensement : une opération périlleuse », *Actes des Journées de Méthodologie Statistique de l'Insee*.

Ardilly P. (2014), « Estimation régionale de taux de pauvreté utilisant une technique de calage », *Actes du 8e colloque francophone sur les sondages*, Dijon.

Ardilly P. (2012), « Estimation localisée du chômage : une application des techniques d'estimation sur petits domaines », *Actes des Journées de Méthodologie Statistique de l'Insee*, conférence invitée.

Ardilly P. (2006), *Les techniques de sondage*, 2e édition, Ed. Technip.

Bailly E. et Rault W. (2013), « Les pacsés en couple hétérosexuel sont-ils différents des mariés ? », *Population et Sociétés*, n° 497.

Blanpain N. et Lincot L. (2013), « 15 millions de grands-parents », *Insee Première*, n° 1469.

Breuil-Genier P., Buisson G., Robert-Bobee I. et Trabut L. (2016), « Enquête *Famille et logements* adossée au recensement de 2011 : comment s'adapter à la nouvelle méthodologie du recensement et quels apports au recensement ? », *Économie et Statistique*, dans ce numéro.

Buisson G. et Lapinte A (2013), « Le couple dans tous ses états, non-cohabitation, conjoints de même sexe, Pacs... », *Insee Première*, n° 1435.

Colin C. (2012), *Rapport du conseil national de l'information statistique : services à la personne*, n° 129.

Davie E. (2011), « Un million de pacsés début 2010 », *Insee Première*, n° 1336.

Domingo P. (2013), « Les modalités de résidence des enfants de parents séparés », *L'e-ssentiel*, publication électronique de la Cnaf, n° 139.

ESSnet on Small Area Estimation (2012), Rapport du groupe de travail n° 5, Eurostat.

Godinot A. (2005), « Pour comprendre le recensement de la population », *Insee Méthodes*, n° hors-série.

Guide de l'utilisateur de l'Enquête Nationale auprès des Ménages 2011 (2013), Statistique Canada.

IASS Satellite Conference (1999), « Small area estimation », Conference Proceedings, Riga, Latvia.

Insee (2013), « Calcul des pondérations de l'enquête *Famille et logements* », *note interne*, n° 137/F170.

Lapinte A. (2013), « Un enfant sur dix vit dans une famille recomposée », *Insee Première*, n° 1470.

Platek R., Rao J.N.K., Sarndal C.E. et Singh M.P. (1987), *Small Area Statistics: An international Symposium*,Wiley Series in Applied Probability and Statistics, New York.

Rao J.N.K. (2003), *Small Area Estimation*, Wiley, Hoboken, New Jersey.

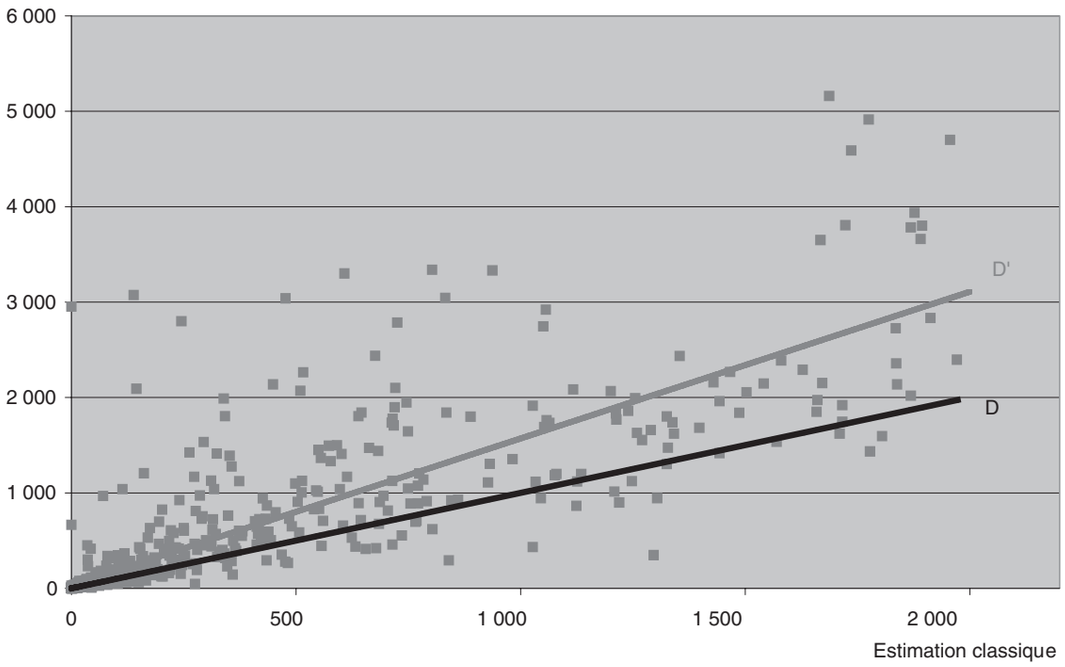
ANNEXE 1

Tableau
Par sous-population, en catégorie de communes n° 2, probabilité d'un enfant de moins de 4 ans d'être gardé à titre principal par l'un des modes de garde

Sous-population		Effectif répondant	Probabilité par mode de garde principal (en %)				
Catégorie de ménage	Tranche d'unité urbaine		Parents	Grands-parents, famille	Assistante maternelle	Crèche	Autre mode
Ménage avec 1 adulte qui travaille à temps plein	Toutes	426	29,8	5,5	28,8	25,3	10,6
Ménage avec 1 adulte qui travaille à temps partiel ou qui est en stage / étude	Toutes	205	34,3	6,3	23,3	24,3	11,8
Ménage avec 1 adulte au chômage	Toutes	185	56,1	10,1	9,6	19,3	4,9
Ménage avec 1 adulte retraité ou au foyer	1	46	88,5	0,0	3,8	5,1	2,6
Ménage avec 1 adulte retraité ou au foyer	2	47	96,2	0,0	0,0	3,1	0,7
Ménage avec 1 adulte retraité ou au foyer	3	54	83,9	0,0	1,1	12,0	3,0
Ménage avec 1 adulte retraité ou au foyer	4	152	84,1	1,9	2,4	8,9	2,7
Ménage avec au moins 2 adultes, et au moins un retraité ou une personne au foyer	1	1 090	85,6	2,5	7,7	3,2	1,0
Ménage avec au moins 2 adultes, et au moins un retraité ou une personne au foyer	2	630	86,1	1,4	7,2	4,7	0,6
Ménage avec au moins 2 adultes, et au moins un retraité ou une personne au foyer	3	656	89,1	0,6	3,1	6,9	0,3
Ménage avec au moins 2 adultes, et au moins un retraité ou une personne au foyer	4	2 100	83,6	2,6	3,7	7,6	2,5
Ménage avec au moins 2 adultes, tous les adultes travaillent à temps plein	1	2 344	25,8	7,1	57,1	8,0	2,0
Ménage avec au moins 2 adultes, tous les adultes travaillent à temps plein	2	1 121	29,3	8,7	49,5	10,1	2,4
Ménage avec au moins 2 adultes, tous les adultes travaillent à temps plein	3	694	27,0	5,4	40,4	24,1	3,1
Ménage avec au moins 2 adultes, tous les adultes travaillent à temps plein	4	3 366	25,8	5,6	31,4	25,9	11,3
Ménage avec au moins 2 adultes, tous les adultes travaillent ou sont en stage / étude mais au moins un adulte ne travaille pas à temps plein	1	1 443	30,2	6,5	53,3	7,8	2,2
Ménage avec au moins 2 adultes, tous les adultes travaillent ou sont en stage / étude mais au moins un adulte ne travaille pas à temps plein	2	663	35,2	5,1	45,9	11,0	2,8
Ménage avec au moins 2 adultes, tous les adultes travaillent ou sont en stage / étude mais au moins un adulte ne travaille pas à temps plein	3	423	36,4	2,9	33,2	22,9	4,6
Ménage avec au moins 2 adultes, tous les adultes travaillent ou sont en stage / étude mais au moins un adulte ne travaille pas à temps plein	4	1 680	32,0	4,6	26,1	29,4	7,9
Ménage avec au moins 2 adultes, aucun n'est retraité ni personne au foyer mais au moins l'un d'eux est au chômage	Toutes	2 170	59,1	4,8	12,5	20,0	3,6
Ensemble		19 495	50,8	4,3	20,2	18,8	5,9

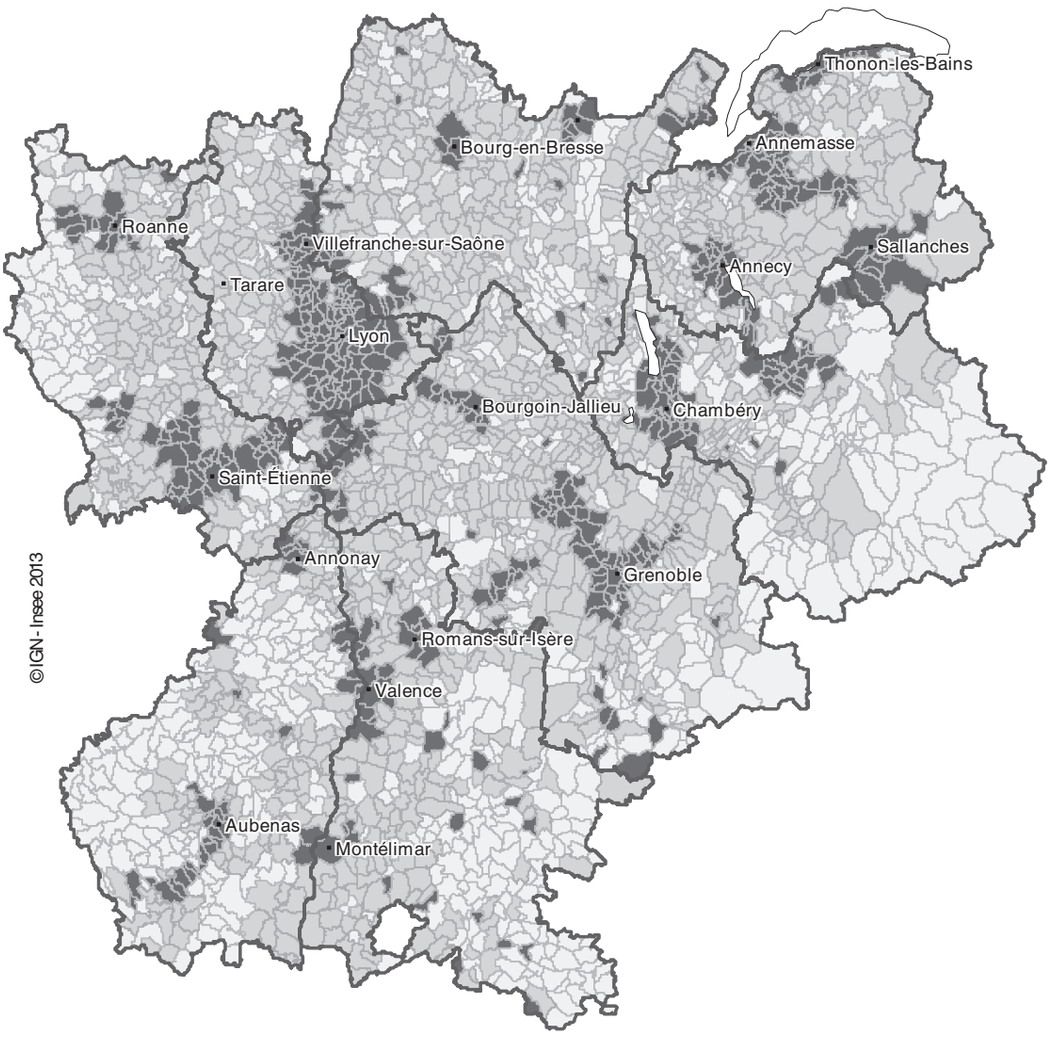
Lecture : considérant un enfant de moins de 4 ans vivant dans un ménage comprenant un seul adulte, travaillant à temps plein et dans une des régions formant la catégorie de communes 2, cet enfant a une probabilité estimée à 25,3 % d'être gardé (principalement) en crèche.
Champ : ensemble des enfants de moins de 4 ans vivant dans un ménage ordinaire de métropole.
Source : EFL *2011,* EAR *2011,* RP *2009, BPE 2010, fichier « Particuliers-employeurs » 2010, distancier Odomatrix Inra Umr 1041 Cesaer, calculs de l'auteur.*

Figure
Estimations classiques et par modèle du nombre de femmes en multi résidence, cas d'une sélection de communes : détection du biais



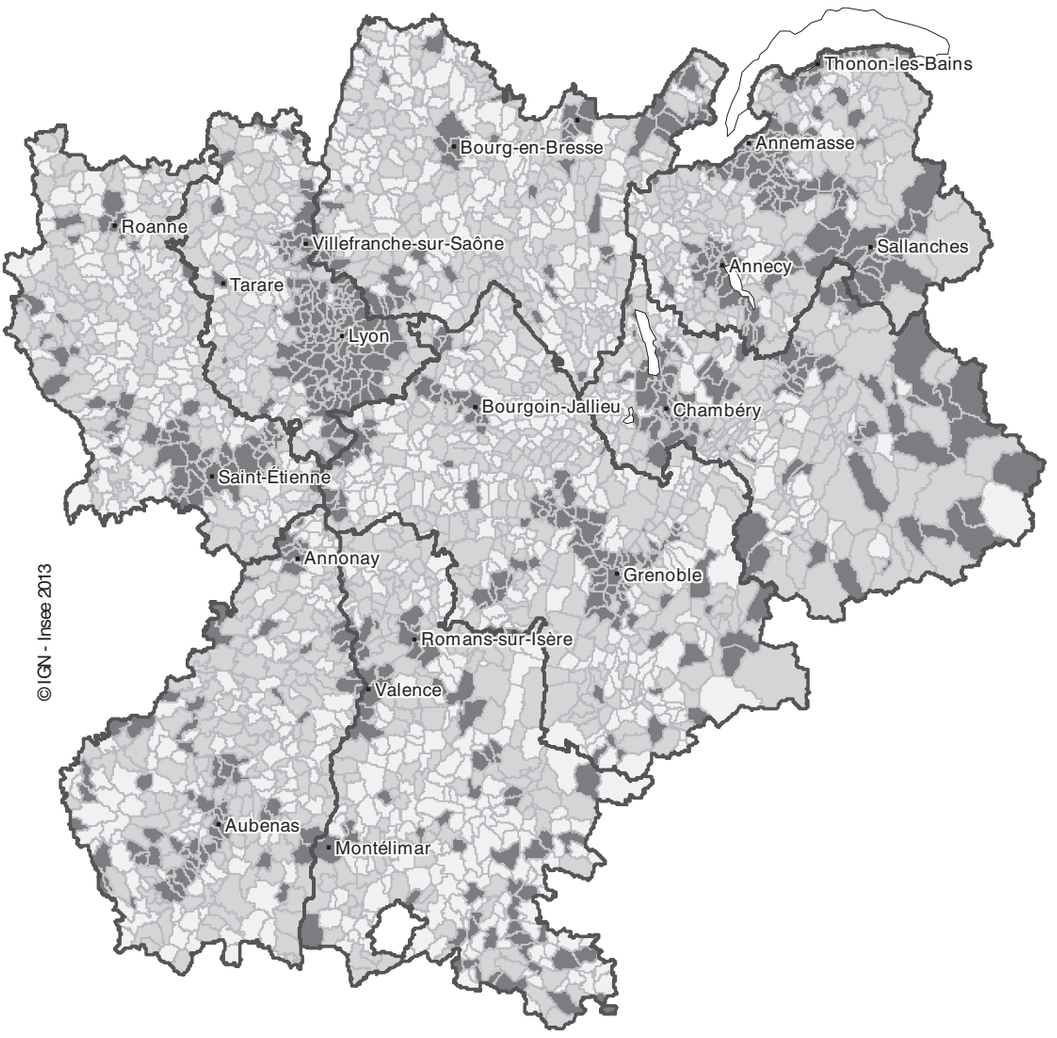
Lecture : une commune est représentée par un point dont l'abscisse est l'estimation du nombre de femmes occupant habituellement deux logements selon l'approche par pondération classique et dont l'ordonnée est l'estimation du nombre de femmes occupant habituellement deux logements selon la méthode proposée utilisant un modèle.
*Champ : ensemble des communes participant à l'*EFL*, où les femmes sont enquêtées et où l'estimation classique est inférieure à 2 000 femmes.*
Source : EFL *2011,* EAR *2011,* RP *2009, calculs de l'auteur.*

Carte I
Proportion d'enfants de moins de 4 ans gardés en crèche



Lecture : toute commune en gris foncé a une proportion d'enfants de moins de 4 ans gardés en crèche (champ : ensemble des enfants de moins de 4 ans, en ménage ordinaire) supérieure à 13,5 %. La proportion est comprise entre 7 % et 13,5 % pour les communes en gris moyen, et inférieure à 7 % pour les communes en gris clair.
Champ : ensemble des communes de la région Rhône-Alpes.
Source : EFL *2011,* EAR *2011,* RP *2009, BPE 2010, distancier Odomatrix Inra Umr 1041 Cesaer, calculs de l'auteur.*

Carte II
Proportion d'hommes vivant en couple non cohabitants



Lecture : toute commune en gris foncé a une proportion d'hommes vivant en couple non cohabitant (champ : ensemble des hommes adultes, en ménage ordinaire) supérieure à 2,3 %. La proportion est comprise entre 1,7 % et 2,3 % pour les communes en gris moyen, et inférieure à 1,7 % pour les communes en gris clair.
Champ : ensemble des communes de la région Rhône-Alpes.
Source : EFL *2011,* EAR *2011,* RP *2009, calculs de l'auteur.*