Recensement 2011 et règlement européen : la procédure d'imputation spécifique à trois variables

Pascal Ardilly *

La modification du questionnaire des *Enquêtes annuelles de recensement* (*EAR*) à partir de 2011, a suscité de nombreux travaux méthodologiques complexes, tels que ceux concernant l'imputation massive de trois variables dans le *Recensement* de 2011 : l'indicateur de résidence antérieure, la période d'achèvement de l'immeuble et l'activité au sens du BIT.

Pour les deux premières, le contexte est celui d'un changement de l'information collectée à partir de 2011. À chaque fois, deux méthodes sont proposées. La première est une approche assez simple au moyen d'un modèle, basée sur l'hypothèse d'une distribution uniforme. La seconde, plus compliquée, utilise un calage par commune ou groupe de communes afin de retrouver en moyenne, à partir des données collectées par les enquêtes annuelles en ancienne nomenclature, les structures estimées en utilisant les données collectées par les enquêtes annuelles en nouvelle nomenclature.

L'imputation de l'activité BIT (trois modalités : actif occupé / chômeur / inactif), en métropole et dans les DOM, relève de la construction complète d'une nouvelle variable à partir de variables auxiliaires individuelles corrélées à l'activité et disponibles à la fois dans le recensement et dans l'enquête *Emploi*. Pour cela, malgré une différence dans les modes de collecte qui crée une hétérogénéité que l'on parvient à corriger en partie, on utilise une variable de déclaration d'activité « spontanée » disponible dans les deux sources. Elle permet d'estimer une liaison assez satisfaisante avec l'activité BIT. On calcule dans un premier temps des probabilités de passage d'une modalité d'activité spontanée vers une modalité d'activité BIT. Puis ces probabilités sont ajustées afin de retrouver en moyenne les effectifs des différentes modalités d'activité BIT donnés par l'enquête *Emploi* dans chaque croisement région / sexe. En particulier, il faut que les taux de chômage BIT par région/ sexe après imputation subissent les perturbations minimales.

Codes JEL : C42, C61, C81, J21, R23.

Mots clés : recensement, imputation, modélisation, changement de nomenclature, réconciliation des sources, mobilité géographique, activité BIT, période d'achèvement d'un immeuble.

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

L'auteur remercie les deux rapporteurs anonymes pour leurs remarques pertinentes et leurs suggestions. L'auteur demeure seul responsable d'éventuelles erreurs ou imprécisions.

Des données sur la population et sur les logements issues du *Recensement de la population* (Godinot, 2005) sont envoyées tous les 10 ans à Eurostat. La dernière transmission a eu lieu fin mars 2014 sur le recensement dont l'année de référence est 2011. Pour la première fois, cet envoi est encadré par un règlement européen (Journal officiel de l'Union européenne n°763 du 9 juillet 2008). Le choix du recensement de 2011 dans le règlement européen génère pour la France une très forte contrainte de calendrier : du fait des délais de collecte, de vérification, de saisie et de traitement des données, les résultats statistiques du recensement de 2011, formé des enquêtes annuelles de 2009 à 2013, ne devaient être disponibles au mieux que début mars 2014, soit très peu de temps avant l'échéance. Aussi, les données françaises envoyées pour répondre au règlement ont été issues du recensement de 2010, moyennant des ajustements de pondération afin d'assurer une cohérence en termes de populations légales avec celles du recensement de 2011 dont le décret a été publié fin décembre 2013.

Les données demandées se présentent sous la forme de 60 tableaux multidimensionnels croisant toutes les modalités des variables d'intérêt (une dizaine de variables par tableau). La plupart de ces tableaux décrivent la population au niveau régional. La confection des tableaux impose de disposer en amont d'une base de données individuelles relatives à chaque variable exploitée : on ne peut donc pas se contenter de produire des estimations agrégées.

La confection des données selon les nomenclatures et concepts requis dans le règlement européen pose problème : certaines données du recensement français doivent être converties afin de s'approcher au mieux de la demande européenne. C'est le cas en particulier du concept de chômage, requis au sens du Bureau international du Travail (BIT), mais appréhendé différemment avec le questionnaire du recensement : la différence de concept entraine plus de 2 points d'écart (voire davantage au niveau régional) entre le taux de chômage national au sens du recensement et celui calculé par l'enquête *Emploi*. Des travaux d'imputations individuelles s'imposent pour produire une variable d'activité au sens du BIT et répondre ainsi au règlement. Par ailleurs, les questions portant sur l'indicateur de résidence antérieure et sur la date d'achèvement de l'immeuble ont changé en 2011, la correspondance s'avérant complexe entre les nouvelles et les anciennes variables. De fait, 60 % des données[1] qui composent le recensement de 2010 ne relèvent pas de ces nouvelles questions : les collectes annuelles antérieures à 2011 ont donc donné lieu à des imputations.

Les méthodes qui ont été appliquées en production et qui vont être décrites par la suite ont une caractéristique commune : elles s'appuient sur des hypothèses de comportement, c'est-à-dire des modèles. De façon générale, la production des statistiques du recensement relève d'une application des techniques de sondage, lesquelles ne dépendent d'aucune modélisation à l'exception notable de la correction de la non-réponse. De fait, c'est ce dernier traitement qui explique qu'une partie des erreurs soit due à des modèles. Dans le cas d'une imputation exhaustive, l'intégralité des données individuelles est produite par des modèles et donc exposée à un risque d'erreur dû à la modélisation. Le processus qui conduit au choix du modèle est guidé par trois considérations. D'une part et en premier lieu, les données auxiliaires disponibles permettant d'expliquer le phénomène à quantifier – lorsqu'il y en a. D'autre part, les exigences en matière de qualité – on entend par là essentiellement les contraintes de cohérence avec des statistiques de même nature déjà produites et le respect d'éventuels seuils d'erreur. Enfin, il faut évidemment tenir compte des moyens que l'on peut y consacrer. La simplicité du modèle, donc son aptitude à être bien compris par l'utilisateur, est également susceptible de jouer, mais cet argument est plus discutable.

Il est légitime que l'on s'interroge, variable par variable, sur la pertinence des procédures d'imputation. Préciser les atouts qu'offrent les procédures d'imputation par modèle dans leur ensemble présente peu d'intérêt : imputer la valeur d'une variable à un individu exige par nature de s'appuyer sur des hypothèses de comportement. En revanche, ce type d'approche a ses limites. Tout d'abord, un modèle n'est qu'une simplification de la réalité, ce qui constitue une source d'erreur incontournable. Par ailleurs, la modélisation est toujours suivie d'une procédure aléatoire qui affecte une valeur à certains individus, cette étape ajoutant sa propre composante d'erreur à celle de la modélisation. De fait, en toute rigueur les statistiques produites ne peuvent pas offrir la même sécurité que celles qui reposent sur des données collectées (même si ces dernières sont sensibles

1. *Trois années d'*EAR *sur les cinq composant un recensement, soit 60 %.*

à bien d'autres imperfections). Néanmoins, les erreurs propres à la technique d'imputation sont d'autant plus faibles que l'on s'intéresse à des populations plus grosses. En effet, d'une part les erreurs de modélisation s'avèrent plus marquées sur des sous-populations spécifiques, d'autre part selon un principe statistique bien connu, l'effet de l'aléa d'imputation diminue avec le nombre d'individus concernés.

L'indicateur de résidence antérieure

Le traitement de la variable de lieu de résidence antérieure se heurte à un changement de questionnaire intervenu en 2011, précisément pour répondre au règlement européen. À cette occasion, la question n°5 du bulletin individuel « *Où habitiez-vous il y a cinq ans ?* » (ancienne nomenclature) est devenue « *Où habitiez-vous il y a un an ?* » (nouvelle nomenclature). Trois modalités restent offertes à l'enquêté : habitation dans le même logement qu'actuellement, habitation dans un autre logement de la même commune, ou habitation dans une autre commune (qu'il convient de préciser), voire à l'étranger. Dans les *Enquêtes annuelles de recensement* (*EAR*) 2010 et antérieures, la population recensée se répartie entre ces trois modalités respectivement à concurrence de 65 %, 10 % et 25 % environ au niveau national. Il a donc fallu concevoir une méthode pour imputer, pour chaque individu recensé au titre d'une *EAR* antérieure à 2011, une de ces trois modalités pour la variable « *Où habitiez-vous il y a un an ?* ». Le traitement ne distingue pas le type de ménage : la méthode est la même que l'individu réside en ménage ordinaire, en communauté (la communauté est alors considérée comme un seul et même ménage) ou qu'il ait un autre statut (sans domicile, marinier, etc.). Il ne distingue pas non plus les DOM de la métropole. Si on considère les *EAR* 2011 et 2012, on constate que les individus recensés se répartissent entre les trois modalités distinguées selon des proportions respectivement égales à 89 %, 3 % et 8 %. Ce sont les données du *RP* 2009, puis du *RP* 2010, qui ont permis de tester la méthodologie adoptée, étant entendu que c'est le *RP* 2010 qui fournira la production finale destinée à Eurostat.

Pour concevoir le processus d'imputation, comme il n'existe malheureusement aucune sous-population dans laquelle l'information soit collectée simultanément selon les deux nomenclatures en jeu, on commence par rechercher les variables disponibles naturellement explicatives de la mobilité géographique. En la circonstance, les autres variables du bulletin individuel (BI) n'apparaissent pas (ou trop peu) en relation avec la mobilité géographique et donc ne sont pas informatives. En revanche, la feuille de logement (FL) pose la question suivante (question n° 8) « *En quelle année avez-vous emménagé dans ce logement ?* », information qu'il est bien naturel d'exploiter dans notre contexte. Évidemment, il faut accepter l'erreur d'observation substantielle qui affecte cette variable, puisque elle-même s'avère imputée pour environ 24 % des logements chaque année (4 % pour cause de non-réponse, 20 % pour cause d'incohérence). Au demeurant, on ne peut en aucun cas échapper à ce type d'erreur et la cohérence des informations définitives au niveau individu / logement, cohérence dont la garantie reste essentielle, ne s'en trouve pas affectée.

La mobilité géographique est une caractéristique de nature individuelle. Néanmoins, il est clair que des corrélations existent entre les situations des différents membres d'un même ménage. Le questionnement du recensement ne permet hélas pas de détecter ces corrélations, lesquelles ne sont finalement prises en compte qu'au travers de quelques règles simples, que l'on considère comme logiques mais qui comportent nécessairement une part d'arbitraire. Dans cet esprit, le premier principe que nous avons adopté postule qu'il ne peut pas y avoir plus d'un changement de domicile au cours de cinq années consécutives. C'est une hypothèse qui bien entendu peut s'avérer fausse sur le terrain (en particulier pour certaines sous-populations, comme par exemple les étudiants et les jeunes actifs). En vertu de ce principe, tout individu qui était présent dans le logement il y a cinq années se trouve *de facto* considéré comme présent dans le logement il y a une année. De même, par hypothèse, on ne peut pas passer d'une situation où on résidait dans une autre commune il y a 5 ans à une situation où on résidait dans un autre logement de la même commune il y a 1 an.

La variable de la FL précisant l'année d'emménagement est une variable de niveau logement : par convention, si tous les occupants du logement recensé n'ont pas emménagé au même moment, on porte la date d'emménagement du premier arrivé, ce qui limite beaucoup la pertinence de son utilisation dans notre contexte. Dans ces conditions, nous avons considéré que l'année d'emménagement dans le logement

pouvait être exploitée seulement lorsqu'aucun individu du ménage recensé ne se déclarait présent dans le logement il y a cinq ans. Dans de telles circonstances, à défaut de tout indice permettant d'être plus précis, l'année d'emménagement donnée par la FL a été attribuée uniformément à tous les membres du ménage. Dans ce cas spécifique, l'imputation est donc déterministe : par exemple, si dans un ménage recensé en 2009 aucun des individus ne déclarait être présent dans le logement au 1er janvier 2004, on considère que tous les individus habitaient dans ce logement au 1er janvier 2008 dès lors que l'année d'emménagement déclarée est comprise entre 2004 incluse et 2007 incluse. En revanche, si l'année d'emménagement déclarée est 2008, voire 2009, on reporte la modalité recensée pour chaque individu (en conservant la commune d'origine si l'individu déclare avoir habité une autre commune il y a cinq ans). L'arbitraire de cette règle disparaît évidemment pour tous les ménages constitués d'une seule personne. Ces règles déterministes ont une importance fondamentale de par leur fréquence : en effet, elles concernent la plupart des individus, à tel point que les imputations plus ou moins complexes qui sont décrites dans toute la suite du texte portent sur moins de 15 % de la population recensée. On s'intéresse désormais aux cas qui ne relèvent pas des règles déterministes qui viennent d'être décrites.

L'approche de base, par modélisation simple

Il reste donc à préciser le principe d'imputation pour les individus qui ne résidaient pas il y a cinq ans dans le logement recensé mais qui appartiennent à un ménage dans lequel au moins une personne résidait il y a cinq ans dans le logement recensé. Dans ces circonstances, la méthode dite « de base » s'appuie sur une nouvelle hypothèse, que l'on considère comme étant la plus naturelle à formuler étant donnée l'information dont on dispose : elle consiste à postuler une loi uniforme entre zéro et cinq ans pour l'ancienneté d'emménagement de chaque individu dans le logement recensé. L'uniformité de la loi traduit l'absence d'indice permettant de dater l'emménagement à telle période plutôt qu'à telle autre. Plus précisément, la collecte d'une *EAR* ayant lieu « en moyenne » à la fin de la première semaine de février, on compte $5 \times 52 + 5 = 265$ semaines écoulées entre le 1er janvier n-5 et la date de collecte en moyenne, contre $52 + 5 = 57$ semaines écoulées entre le 1er janvier n-1 et la date de collecte. Cette correction, dont la précision est peut-être illusoire mais qui reste très simple, permet d'estimer la probabilité d'emménagement après le 1er janvier n-1 à la valeur $57 / 265 = 0{,}215$. Concrètement, on procède ainsi : on considère un individu du ménage qui n'était pas dans le logement au 1er janvier n-5 et on génère au niveau du ménage un nombre aléatoire X entre 0 et 1. Si X est supérieur à 0,215 on impute la modalité « résidence il y a un an dans le même logement que maintenant », sinon on reporte mécaniquement la modalité déclarée pour le lieu de résidence au 1er janvier n-5. On notera que X est le même pour tous les individus du ménage, ce qui est une façon simple d'introduire une corrélation (on pourrait dire une coordination) naturelle entre les comportements de mobilité des différents individus d'un même ménage.

Un traitement spécifique un peu complexe est appliqué aux enfants nés entre le 1er janvier n-5 et le 1er janvier n-1 : ces enfants ont initialement un indicateur de résidence antérieure non renseigné mais il faut évidemment leur en affecter un dans le nouveau contexte. On applique alors la règle suivante. Un filtre amont traite immédiatement le cas des enfants des ménages dans lesquels tous les (autres) individus déclarent après imputation habiter le logement au 1er janvier n-1 : ces enfants sont systématiquement considérés comme résidant dans le logement au 1er janvier n-1 (soit 81 % des enfants). Si ce n'est pas le cas, en premier lieu et seulement s'il s'agit d'un ménage ordinaire, on vérifie s'il y a dans le ménage au moins une femme dont l'âge est compris entre 20 et 50 ans. Si c'est le cas (concerne 18 % des enfants), on considère que l'une des femmes de cette tranche d'âge est la mère de l'enfant et on attribue à ce dernier, suite à l'imputation préalable des adultes, la modalité de rang minimum parmi les modalités prises par les femmes éligibles (les modalités du lieu de résidence il y a 1 an étant en effet ordonnées, de 1 à 3 dans le BI). Si aucun adulte de sexe féminin ne peut être considéré comme la mère de l'enfant ou s'il s'agit d'une communauté, et si aucun adulte n'était présent dans le logement il y a 1 an (0,9 % des enfants), on choisit d'imputer à l'enfant la modalité de code maximum repérée parmi tous les adultes, tout en récupérant la commune (voire le pays antérieur) de l'adulte associé à la modalité de rang maximum. En revanche, si au moins un adulte était présent il y a 1 an, avec une chance sur deux on considère que l'enfant résidait dans le logement il y a 1 an et avec une chance sur deux on lui affecte la modalité maximale des adultes.

Dans ce deuxième cas de figure, on impute alors la commune (voire le pays) d'origine associé(e) à l'adulte (ou à « un » adulte au hasard s'il y en a plusieurs…) qui possède la modalité maximale. Cet ultime contexte est très rare (0,1 % des enfants). Cet algorithme permet d'éviter les incohérences qui résulteraient d'une éventuelle imputation aléatoire « indépendante » des enfants, comme par exemple obtenir une situation non crédible où les parents n'étaient pas présents dans le logement il y a 1 an mais où l'enfant serait (seul…) résident dans ce logement. Le tableau 1 résume les différents cas de figure rencontrés.

Après imputation, les seuls individus recensés pour lesquels la variable définitive est non renseignée sont les enfants nés le 2 janvier n-1 et ultérieurement, les personnes qui ne résident ni en ménage ordinaire ni en communauté, ainsi que quelques cas exceptionnels autres, qui font figure d'anomalie. Le tableau 2 fournit, à partir des données relatives aux trois *EAR* collectées en ancienne nomenclature (2008 à 2010), France entière (donc y compris DOM), les effectifs individuels non pondérés associés aux croisements des modalités respectivement avant et après imputation. Les individus pour lesquels l'indicateur de résidence antérieure n'est pas renseigné avant imputation (1 785 000 individus, essentiellement les enfants de moins de 5 ans) ne sont pas pris en compte.

Les tableaux 3 (version non pondérée) et 4 (version pondérée avec les poids du *RP* 2010, expurgée des valeurs non renseignées) permettent de visualiser les différences de traitement selon l'année de collecte, c'est-à-dire selon l'*EAR*. On constate que l'imputation dans les *EAR* 2008 à 2010 conduit à une structure proche, mais néanmoins distincte, de celle des *EAR* 2011 et 2012 (laquelle prétend approcher la réalité du terrain), les écarts de profil ayant des origines multiples, comme *a priori* l'erreur de notre modèle d'imputation (modèle de loi uniforme) dont on peut penser qu'il s'agit du risque majeur, mais aussi l'aléa d'imputation que nous avons introduit, l'erreur d'échantillonnage des *EAR* et l'influence complexe de la pondération lorsqu'elle est prise en compte (la pondération est celle du recensement et non celle des *EAR*), l'erreur d'imputation des variables d'origine dans les *EAR* et peut-être en sus, mais ce n'est

Tableau 1
Type d'imputation selon la situation de l'individu 5 ans auparavant

Situation de l'individu	
Il y a 5 ans	*Il y a 1 an*
Présent dans le logement	Présent dans le logement
Dans un autre logement, vivant seul (actuellement)	Imputation déterministe, selon l'année d'emménagement (voir feuille de logement)
Dans un autre logement, actuellement dans un ménage où personne n'occupait le logement il y a 5 ans	Imputation déterministe, selon l'année d'emménagement (voir feuille de logement)
Dans un autre logement, actuellement dans un ménage où au moins une personne occupait le logement il y a 5 ans	Imputation aléatoire
Enfant de moins de 5 ans	Algorithme complexe spécifique

Tableau 2
Approche de base : situation définitive croisant « avant » et « après » imputation

		Résidence antérieure après imputation			
		Même logement	Autre logement, même commune	Autre commune	Total
Résidence antérieure avant imputation	Même logement	16 760 000	0	0	16 760 000
	Autre logement, même commune	1 760 000	645 000	0	2 405 000
	Autre commune	4 515 000	0	2 150 000	6 665 000
	Total	23 035 000	645 000	2 150 000	25 830 000

Lecture : avant imputation, dans le fichier constitué par les EAR *2008, 2009 et 2010, 1 760 000 individus recensés déclaraient résider (5 ans avant la collecte) dans un autre logement de la même commune et sont déclarés après imputation résider (1 an avant la collecte) dans le même logement qu'au moment de la collecte.*
*Champ : France entière, ensemble des individus recensés qui ont déclaré une résidence antérieure dans l'*EAR.
Source : EAR *2008, 2009 et 2010 (non pondéré).*

pas décelable sans une étude spécifique, une évolution dans le temps de l'ampleur de l'erreur de mesure.

C'est pour tenter de réduire ces (modestes) écarts de profil qu'une méthode alternative, dite « par calage » a été envisagée.

La méthode par calage

La méthode dite « de base » est simple à comprendre et à programmer, ce qui est un atout majeur. Néanmoins, elle possède un inconvénient *a priori* pénalisant, à savoir le biais que génère le modèle d'imputation simplifiant la réalité et qui est susceptible d'apparaître lorsqu'on compare la distribution de la variable imputée à la distribution de référence. Cette dernière est donnée par les *EAR* en nouvelle nomenclature qui, par construction, devraient traduire correctement la réalité du terrain, à l'erreur d'échantillonnage près, cette dernière étant toutefois très faible au niveau national. En particulier, le biais d'imputation, s'il existe, rendra plus délicate la communication qui accompagne la diffusion des estimations issues de recensements successifs. En effet, dans n'importe quelle grande commune par exemple, le *RP* 2011 sera constitué de deux *EAR* imputées (*EAR* 2009 et 2010) et de trois *EAR* en nouvelle nomenclature : lorsqu'on comparera les résultats du *RP* 2012 et ceux du *RP* 2011 on substituera, d'une certaine façon, l'année 2014 à l'année 2009, c'est-à-dire qu'une année de collecte en nouvelle nomenclature remplacera une année de collecte ayant donné lieu à imputation. Si un biais existe et que rien n'est fait, il faut s'attendre à des variations (peut-être) significatives dues à l'imputation et

Tableau 3
Approche de base : imputation définitive selon l'année de collecte, sans pondération

EAR	Résidence antérieure après imputation									
	Non renseignée		Même logement		Autre logement, même commune		Autre commune		Total	
	Effectif	%	Effectif	%	Effectif	%	Effectif	%	Effectif	%
2008	146 000	1,6	8 007 000	87,2	233 000	2,5	794 000	8,7	9 180 000	100
2009	142 000	1,5	8 038 000	87,5	234 000	2,5	777 000	8,5	9 191 000	100
2010	142 000	1,5	8 110 000	87,7	235 000	2,5	765 000	8,3	9 252 000	100
2011	203 000	2,2	8 077 000	86,9	296 000	3,2	715 000	7,7	9 291 000	100
2012	142 000	1,5	8 098 000	87,3	301 000	3,2	738 000	8,0	9 279 000	100
Total	775 000	1,7	40 330 000	87,3	1 299 000	2,8	3 789 000	8,2	46 193 000	100

Lecture: l'EAR 2010, après imputation, apporte au RP *2010 une contribution de 8 110 000 individus déclarant vivre dans le même logement qu'au 1er janvier 2009, ce qui représente 87,7 % des individus recensés en 2010.*
Champ: France entière, ensemble des individus.
Source: RP *2010 (non pondéré).*

Tableau 4
Approche de base : imputation définitive selon l'année de collecte, avec pondération

EAR	Résidence antérieure après imputation							
	Même logement		Autre logement, même commune		Autre commune		Total	
	Effectif	%	Effectif	%	Effectif	%	Effectif	%
2008	10 730 000	88,8	350 000	2,9	1 004 000	8,3	12 084 000	100
2009	10 780 000	88,9	354 000	2,9	987 000	8,2	12 121 000	100
2010	10 770 000	89,1	349 000	2,9	967 000	8,0	12 086 000	100
2011	11 270 000	88,7	484 000	3,8	947 000	7,5	12 701 000	100
2012	11 280 000	88,5	488 000	3,8	976 000	7,7	12 744 000	100
Total	54 830 000	88,8	2 025 000	3,3	4 881 000	7,9	61 736 000	100

*Lecture : l'*EAR *2010, après imputation et en utilisant les poids du* RP *2010, apporte au* RP *2010 une contribution de 349 000 individus déclarant vivre au 1er janvier 2009 dans un autre logement de la même commune, ce qui représente 2,9 % des individus recensés et imputés en 2010.*
Champ : France entière, ensemble des individus imputés.
Source : RP *2010.*

il faudra expliquer aux utilisateurs l'origine d'une partie des à-coups dans les séries impliquant cette variable de résidence antérieure. Ces derniers comprendront certainement que des changements de méthode créent des perturbations, mais il est possible en théorie de les éviter. C'est le principe de la méthode d'imputation alternative dite « par calage », dont l'objectif est de se protéger contre un éventuel biais d'imputation. Cette approche n'utilise pas la variable de la FL « *Année d'emménagement* ». Elle traite les enfants ayant entre un et cinq ans exactement comme la méthode de base et elle reste d'inspiration aléatoire, c'est-à-dire que le statut final des individus en matière de résidence antérieure est toujours régi par des règles basées sur une loterie, donc sur un nombre généré au hasard. Mais, fait nouveau, ces règles intègrent un principe de calage afin que le comportement en matière de résidence au 1er janvier n-1 après imputation coïncide « en espérance » avec le comportement en matière de résidence au 1er janvier n-1 tel qu'il ressort des seules *EAR* qui sont effectivement collectées en nouvelle nomenclature. Précisons cette démarche.

Plaçons-nous dans une grande commune. À partir des *EAR* collectées en ancienne nomenclature, la résidence au 1er janvier n-5 y est décrite par la ventilation des individus entre les trois modalités de la variable « *Où habitiez-vous le 1er janvier n-5 ?* », modalités notées respectivement A, B et C.

A : dans le même logement que maintenant

B : dans un autre logement de la même commune

C : dans une autre commune, ou un autre pays

Cela conduit, compte tenu des pondérations du recensement, à estimer des effectifs notés respectivement $\widehat{N}_A^5$, $\widehat{N}_B^5$ et $\widehat{N}_C^5$. Ce sont donc des effectifs estimés à partir des seules *EAR* à imputer. Les *EAR* disponibles en nouvelle nomenclature permettent pour leur part d'estimer la structure-cible de population associée à la variable « *Où habitiez-vous le 1er janvier n-1 ?* » : on en déduit, pour l'échantillon des seules *EAR* à imputer, les effectifs N_A^1, N_B^1 et N_C^1 faisant office d'effectifs-cibles. Il reste alors à trouver un jeu de probabilités organisant le changement de statut – on parlera désormais de probabilités de transition – et qui, partant de la structure $\widehat{N}_A^5$, $\widehat{N}_B^5$ et $\widehat{N}_C^5$ conduise en espérance à la structure N_A^1, N_B^1 et N_C^1. Il est possible pour cela de se contenter de deux probabilités θ et μ faisant office de paramètres (cf. tableau 5).

La probabilité égale à 1 organisant le passage systématique de la modalité A vers la modalité A (imputation déterministe) n'est pas surprenante et traduit bien le principe exposé supra : on ne change de logement qu'au plus une fois durant cinq années consécutives. On trouvera dans l'encadré 1 une description du processus de détermination mathématique des deux probabilités θ et μ.

L'expression mathématique des probabilités de transition ne garantit pas hélas que ces dernières soient toujours positives. Ce risque provient d'une éventuelle incohérence entre, d'une part la « vraie » structure-cible constituée par le triplet (N_A^1, N_B^1 et N_C^1) et d'autre part le triplet ($\widehat{N}_A^5$, $\widehat{N}_B^5$ et $\widehat{N}_C^5$). Intuitivement, on peut s'attendre à ce que, dans la grande commune considérée,

Tableau 5
Modèle de transition pour les *EAR* à imputer

		Après Imputation			Effectif total concerné avant imputation
		Modalité A (il y a 1 an)	Modalité B (il y a 1 an)	Modalité C (il y a 1 an)	
Avant imputation	Modalité A (il y a 5 ans)	1	0	0	$\widehat{N}_A^5$
	Modalité B (il y a 5 ans)	θ	$1-\theta$	0	$\widehat{N}_B^5$
	Modalité C (il y a 5 ans)	μ	0	$1-\mu$	$\widehat{N}_C^5$
Effectif total concerné après imputation (structure cible)		N_A^1	N_B^1	N_C^1	$\widehat{N}$

Lecture : si l'on considère un individu d'une EAR *à imputer ayant déclaré une modalité B (habite dans un autre logement de la même commune) au 1er janvier n–5, on lui affectera au 1er janvier n–1 aléatoirement la modalité A (habite dans le même logement) avec une probabilité θ ou la modalité B avec la probabilité complémentaire $1-\theta$.*
Champ : individus des EAR *à imputer.*

$\widehat{N}_B^5$ (respectivement $\widehat{N}_C^5$) soit à peu près cinq fois plus grand que N_B^1 (respectivement N_C^1), car le nombre et la structure des mobilités devraient être à peu près réguliers d'une année sur l'autre ; évidemment, on trouvera des situations spécifiques où un phénomène local a précipité des départs et des arrivées d'individus à certaines périodes, mais ce n'est probablement pas le scénario le plus courant. Lorsque le ratio entre $\widehat{N}_B^5$ et N_B^1 est très atypique, on ne peut pas trouver de probabilité de transition respectant en espérance la structure cible et pour procéder à l'imputation, la probabilité est forcée à la valeur zéro. Cette situation extrême survient d'autant moins fréquemment que la structure-cible est basée sur un plus grand nombre d'*EAR*, donc se trouve elle-même estimée avec une meilleure précision : on rappelle que l'application en vraie grandeur, parce qu'elle est basée sur le *RP* 2010, s'appuie sur une structure-cible construite à partir de deux *EAR* (2011 et 2012).

Dans le cas des petites communes, la même méthode peut être appliquée mais cette fois ce ne peut être que sur un regroupement de petites communes et de telle manière que l'on puisse construire des structures-cibles qui ont un sens. Un groupe de petites communes est *a priori* constitué par l'ensemble des petites communes d'un département (cette définition peut être facilement modifiée dans le programme informatique). Si on augmente la taille du groupe, on risque d'augmenter l'hétérogénéité des communes et donc l'ampleur du biais de modèle. Du fait de l'exhaustivité du recensement dans les petites communes, la diminution de la taille du groupe est plutôt souhaitable mais cela conduit à des travaux de typologie lourds et probablement peu rentables. Le traitement des petites communes est sensiblement plus risqué que celui des grandes communes : il est facile d'imaginer des situations, sur le terrain, où la mobilité géographique affectant une petite commune donnée se distingue de celle constatée en moyenne dans l'ensemble de son groupe.

Nous produisons une application de cette méthode sur la région Auvergne (cf. tableau 6) puis sur la commune de Clermont-Ferrand (cf. tableau 7). La structure de référence de la région Auvergne, toutes communes confondues est la suivante : 88,7 % des individus en modalité A, 3,6 % en modalité B et 7,7 % en modalité C. La structure de référence de la commune de Clermont-Ferrand est : 79,8 % des individus en modalité A, 8,5 % en modalité B et 11,7 % en modalité C. Les structures présentées sont toutes pondérées avec les poids du recensement (et non le poids des *EAR*). L'examen de ces

Encadré 1

DÉTERMINATION DES PROBABILITÉS DE TRANSITION

On note s l'échantillon constitué par les EAR en ancienne nomenclature (au nombre de λ, $\lambda = 3$ pour le RP 2010), hors individus qui ne sont pas à imputer. Soit w_i le poids au recensement de l'individu i. Par construction $\sum_{i \in s} w_i = \widehat{N}_A^5 + \widehat{N}_B^5 + \widehat{N}_C^5 = N_A^1 + N_B^1 + N_C^1$, noté $\widehat{N}$. Noter que l'interprétation de $\widehat{N}$ est déroutante, s'agissant d'une grandeur proche de $\lambda / 5$ fois la taille totale de la population de la grande commune. On va raisonner « en moyenne » par rapport à l'aléa d'imputation pour définir les probabilités θ et μ permettant de se rapprocher de la structure cible.

Soit $1_{i \in A}$ la variable indicatrice aléatoire formalisant l'imputation éventuelle en modalité A, valant 1 si i se voit imputer la modalité A et 0 sinon. Cette variable suit une loi de Bernoulli, de paramètre θ ou μ selon que i vérifie avant imputation la modalité B ou C. On s'intéresse à la variable aléatoire :

$$\widehat{N}_A^1 = \sum_{\substack{i \in A \\ i \in s}} w_i + \sum_{\substack{i \in B \\ i \in s}} w_i \cdot 1_{i \in A} + \sum_{\substack{i \in C \\ i \in s}} w_i \cdot 1_{i \in A}$$

les domaines de sommation étant relatifs à la situation cinq ans auparavant. $\widehat{N}_A^1$ a pour espérance par rapport à la loi traduisant le mécanisme d'imputation :

$$E\left(\widehat{N}_A^1\right) = \sum_{\substack{i \in A \\ i \in s}} w_i + \sum_{\substack{i \in B \\ i \in s}} w_i \cdot \theta + \sum_{\substack{i \in C \\ i \in s}} w_i \cdot \mu = \widehat{N}_A^5 + \theta \cdot \widehat{N}_B^5 + \mu \cdot \widehat{N}_C^5,$$

grandeur qui, dans la situation idéale et compte tenu de la stratégie préconisée, doit être égale à N_A^1, effectif cible obtenu à partir des EAR (en nombre $5 - \lambda$) qui sont collectées en nomenclature définitive et qui forment l'échantillon noté $RP - s$. Par le même raisonnement, $(1-\theta) \cdot \widehat{N}_B^5$ doit être égal à N_B^1 et $(1-\mu) \cdot \widehat{N}_C^5$ doit être égal à N_C^1. On vérifie facilement que les trois équations ainsi formées sont liées. Il suffit évidemment de choisir

$$\theta = 1 - \frac{N_B^1}{\widehat{N}_B^5} = 1 - \frac{\sum_{\substack{i \in B \\ i \in RP-s}} w_i}{\sum_{\substack{i \in B \\ i \in s}} w_i} \cdot \frac{\sum_{i \in s} w_i}{\sum_{i \in RP-s} w_i} \quad \text{et} \quad \mu = 1 - \frac{N_C^1}{\widehat{N}_C^5}$$

tableaux laisse penser que la méthode calée est un peu plus pertinente que l'approche de base, en particulier pour reproduire les modalités les moins fréquentes. Néanmoins, les améliorations d'ensemble restent modestes sur le plan numérique, même au niveau d'une commune donnée.

Le système ainsi contraint réduit en théorie le biais d'imputation mais ne contrôle pas la variance d'imputation. Dans aucune des deux méthodes il n'y a eu de tentative pour réduire la variance, pour les raisons suivantes. En premier lieu, dans la méthode par modèle, il n'existe pas de contrainte naturelle à imposer au niveau des communes ou des groupes de communes pour limiter la variabilité de l'imputation, donc d'une certaine façon la question ne se pose pas (toutes proportions gardées, c'est exactement l'esprit dans lequel on traite le phénomène de non-réponse dans les enquêtes, assimilé à une loterie qui n'est soumise à aucune contrainte). C'est différent dans l'approche par calage parce qu'il y a des effectifs à respecter et qu'un échantillonnage équilibré aurait pu être utilisé pour y parvenir de manière exacte, et non en moyenne. Au-delà du principe, la programmation aurait été sensiblement plus compliquée (en grande partie à cause des probabilités de transition négatives et du grand nombre de communes et groupes de communes à gérer) et cela n'aurait rien changé sur le plan numérique. En effet, au niveau où l'information sera exploitée, c'est-à-dire à un niveau très agrégé, le nombre d'individus imputés est tellement grand que la variance d'imputation est négligeable – étant entendu que d'éventuelles études porterant à un niveau localisé sur cette variable ou sur sa liaison avec d'autres variables seraient

Tableau 6
Comparaison des deux méthodes d'imputation, région Auvergne : structures en %

En %

EAR	Approche de base				Méthode par calage			
	Même logement	Autre logement, même commune	Autre commune	Total	Même logement	Autre logement, même commune	Autre commune	Total
2008	88,6	2,8	8,6	100	88,3	3,8	7,9	100
2009	88,5	2,8	8,7	100	88,1	3,8	8,1	100
2010	88,9	2,8	8,3	100	88,4	3,8	7,8	100
2011	88,7	3,7	7,6	100	88,7	3,7	7,6	100
2012	88,6	3,6	7,8	100	88,6	3,6	7,8	100
Total	88,6	3,1	8,3	100	88,4	3,7	7,9	100

Lecture : le sous-échantillon EAR *2010, pondéré avec les poids du* RP *2010, conduit, en approche de base, à une estimation finale après imputation de 88,9 % des individus qui déclarent résider dans le même logement qu'au 1er janvier 2009. Avec la méthode de calage, cette proportion devient 88,4 %.*
Champ : ensemble des individus recensés et imputés en région Auvergne.
Source : RP *2010.*

Tableau 7
Comparaison des deux méthodes d'imputation, commune de Clermont-Ferrand : structures en %

En %

EAR	Approche de base				Méthode par calage			
	Même logement	Autre logement, même commune	Autre commune	Total	Même logement	Autre logement, même commune	Autre commune	Total
2008	82,5	5,4	12,0	100	80,1	8,4	11,5	100
2009	80,1	6,4	13,5	100	79,1	8,7	12,2	100
2010	81,6	6,0	12,4	100	79,2	9,3	11,5	100
2011	79,4	8,3	12,3	100	79,4	8,3	12,3	100
2012	80,4	8,6	11,0	100	80,4	8,6	11,0	100
Total	80,7	7,0	12,3	100	79,6	8,7	11,7	100

Champ : ensemble des individus recensés et imputés dans la commune de Clermont-Ferrand (Puy-de-Dôme).
Source : RP *2010.*

de toute façon extrêmement fragiles. Secundo, les statistiques qui seront produites à partir des données imputées seront de diverses natures, la plupart d'entre elles concernant des sous-populations complexes définies par l'utilisateur pour ses besoins. Enfin, dès lors qu'elle reste faible, même au niveau local, la variance d'imputation n'est pas en soi un ennemi et le respect d'une structure en moyenne peut suffire, voire même paraitre plus réaliste dans notre contexte. Pour conclure sur ce point, il faut rappeler que l'écueil essentiel que la méthode par calage doit contourner est le biais susceptible d'être produit par la technique de modélisation, et rien de plus. Et on pourra rajouter que, dans les petites communes essentiellement parce que la contrainte est définie à un niveau supra-communal et que les contextes locaux peuvent être fortement dépendants de la commune, il est toujours possible de pratiquer un tirage dit rejectif, c'est-à-dire que l'on peut toujours procéder à une nouvelle imputation (aléatoire) si la loterie conduit à certaines imputations communales manifestement inadmissibles.

Que peut-on en penser ?

L'utilisation d'un modèle est une opération ingrate en ce sens où l'appréciation de sa pertinence ne peut s'effectuer sérieusement qu'en présence d'une information de même nature provenant d'une autre source. Autrement dit, si l'on ne dispose pas d'une référence, on ne peut pas apprécier quantitativement la qualité de l'imputation. En la circonstance et malheureusement, aucune source ne s'est avérée raisonnablement exploitable pour permettre d'affiner le modèle simple de base…si ce n'est justement les *EAR* 2011 et 2012 elles-mêmes, ce qui relève précisément de la stratégie de calage. La meilleure appréciation de l'efficacité de l'imputation se perçoit donc au travers des tableaux 6 et 7, où l'on constate clairement l'avantage comparatif de la méthode par calage. Cela provient du fait que l'hypothèse du modèle de base est évidemment simplificatrice et insuffisamment adaptée à certaines sous-populations, au premier rang desquelles les étudiants et les jeunes actifs. Ces derniers font preuve d'une plus grande mobilité et donc l'hypothèse formulée les place, un an auparavant, plus fréquemment dans le logement recensé qu'ils ne le sont en réalité – ce qui est visible si on examine les colonnes « Autre logement, même commune » et « Autre commune » des tableaux 6 et 7 pour l'approche de base. L'approche par calage semble mieux adaptée, mais ce n'est pas la méthode qui a été finalement retenue. Quels sont alors les arguments plaidant pour l'approche par modèle de base – dont on rappelle qu'elle ne porte que sur une partie minoritaire de la population ? Trois arguments peuvent être avancés. Primo, affiner le modèle de mobilité passe techniquement, d'une façon ou d'une autre, par l'estimation de la fonction de répartition de la variable « durée de résidence dans un logement » pour les sous-populations les plus mobiles, et cela a paru hors de portée, même avec les enquêtes répétées dans le temps dont on dispose à l'Insee (lesquelles sont inadaptées et insuffisantes pour cette mesure, qui requiert des suivis individuels longs et souffre d'effets de censure forts). Secundo, le phénomène est vraisemblablement très sensible à de multiples paramètres environnementaux (aspects législatifs, prix des logements, politique de l'éducation, marché local de l'emploi, etc.), donc instable dans le temps et dans l'espace, ce qui obère les chances de succès d'aboutir à un modèle satisfaisant pour résumer une situation locale. Au demeurant, la complexité du phénomène de mobilité est loin d'être totalement appréhendée, à en juger par les irrégularités régulièrement mises en évidence par certaines statistiques de mobilité régionale issues des recensements passés. Enfin, mais cela relève cette fois d'une considération entièrement stratégique, il n'a pas été jugé utile d'introduire des contraintes de cohérence pour cette variable, autrement dit se sont bien les statistiques nationales du tableau 4 qui ont constitué, seules, la préoccupation en matière de qualité, et en la circonstance elles ont été jugées satisfaisantes. Il n'en demeure pas moins que l'exploitation des données imputées, limitée à des sous-populations spécifiques, soit particulièrement mobiles, soit localisées dans de petites zones géographiques, est périlleuse et ne doit s'effectuer qu'avec la plus grande prudence.

La période d'achèvement de la construction

La feuille de logement du recensement (question n°2) contient l'information sur la période d'achèvement du logement (plus exactement de l'immeuble dans lequel se situe le logement). Les logements recensés sont tous les locaux destinés à l'habitation à l'exception de ceux qui abritent les communautés et les habitations mobiles. On ne prend pas non plus en compte les logements fictifs des personnes sans abri. En revanche, toutes les catégories de

logement sont ici concernées : logements principaux, secondaires, occasionnels et vacants. Le contexte motivant l'imputation de la période d'achèvement du logement est similaire à celui de la variable de résidence antérieure : le questionnaire de l'*EAR* 2011 adopte une nouvelle nomenclature. Aussi, connaissant la période d'achèvement selon l'ancienne nomenclature (*EAR* 2010 et antérieures), on veut prédire une période d'achèvement selon la nouvelle. Le tableau 8 explicite les anciennes et les nouvelles tranches, avec leur codification par des lettres (ancienne nomenclature) ou des chiffres (nouvelle) afin de pouvoir suivre plus clairement la démarche adoptée pour cette prédiction.. À titre préliminaire, il faut signaler que, comme pour l'indicateur de résidence antérieure, la variable période d'achèvement est entachée d'une erreur assez conséquente : chaque année, à peu près 18,5 % des logements ne renseignant pas cette question, auxquels il convient d'y ajouter près de 2 % de corrections pour incohérence.

Le traitement des logements construits en 1999 ou après est immédiat, puisque l'année d'achèvement est précisée dans l'ancienne nomenclature. Pour ce qui est des logements construits avant 1999, les distorsions de tranche concernent essentiellement les deux plus anciennes. En effet, les trois tranches de l'ancienne nomenclature relatives aux années 1975 à 1998 (tranches C, D et E) sont (presque) incluses dans les tranches de la nouvelle nomenclature, à l'exception notable de l'année 1990 qui vient perturber cette relation d'inclusion. Pour ce qui concerne les tranches qui ne bénéficient pas d'une relation d'inclusion (A, B, E), le processus d'imputation ne peut pas tirer bénéfice de l'information contenue par ailleurs dans la feuille de logement, aucune variable n'apparaissant un tant soit peu explicative de la période d'achèvement. Dans ces conditions, deux pistes sont envisageables. Ou bien on cherche à mobiliser une source externe pour tenir compte, d'une façon ou d'une autre, d'une distribution estimée des dates d'achèvement, ou bien on s'en tient à l'exploitation de la structure estimée des périodes d'achèvement à partir des seules *EAR* qui donnent lieu à une collecte en nouvelle nomenclature, considérant qu'il s'agit là de la meilleure référence dont on puisse disposer pour refléter la réalité du terrain.

L'approche de base : une modélisation qui exploite l'enquête *Logement*

Une difficulté particulière tient au fait que l'ancienneté de la construction immobilière est manifestement très dépendante de la commune : selon que l'on considère une ville nouvelle ou une ancienne cité médiévale, le contexte sera évidemment sans commune mesure. Cette hétérogénéité du terrain nous incite à rechercher en priorité une source exhaustive exploitable au niveau communal. Les recensements exhaustifs du passé ne sont évidemment d'aucun secours et la seule source qui pourrait *a priori* répondre à cet objectif est constituée par les fichiers annuels de suivi de la construction neuve (permis de construire, déclarations de fin des travaux). Malheureusement, le système d'information associé ne permet pas de remonter suffisamment loin dans le temps pour couvrir les tranches concernées par l'imputation. Ce défaut est rédhibitoire, puisque nous avons constaté que la composante délicate de l'imputation concerne presque exclusivement les logements construits avant 1975. La source alternative qui s'impose est celle de l'enquête *Logement* de 2006, mais s'agissant d'une enquête par sondage elle n'a bien entendu aucune pertinence à un niveau local. Aussi n'a-t-il pas semblé opportun d'utiliser l'enquête *Logement* 2006 (42 700 logements répondants) autrement qu'à titre d'ultime recours ; le traitement des

Tableau 8
Ancienne et nouvelle nomenclatures des périodes d'achèvement

Ancienne nomenclature (jusqu'à 2010)		Nouvelle nomenclature (à partir de 2011)	
Définition de la tranche	Code	Définition de la tranche	Code
Avant 1949	A	Avant 1919	1
De 1949 à 1974	B	De 1919 à 1945	2
De 1975 à 1981	C	De 1946 à 1970	3
De 1982 à 1989	D	De 1971 à 1990	4
De 1990 à 1998	E	De 1991 à 2005	5
1999 ou après (année précisée)	F	2006 ou après (année précisée)	6

logements achevés avant 1949 va néanmoins en donner l'occasion. En effet, dans toute tranche bornée de manière explicite, c'est-à-dire pour tout logement achevé (selon l'ancienne nomenclature) entre 1949 et 1998, nous avons considéré que l'approche la plus naturelle consistait à utiliser un principe de distribution uniforme de l'année d'achèvement entre les différentes années constituant la tranche. Par exemple, considérant la tranche de construction « De 1949 à 1974 » couvrant exactement 26 années, on postule qu'un logement recensé dans cette tranche a 1 chance sur 26 d'être achevé durant chacune des années constitutives de la tranche, donc finalement 22 chances sur 26 d'être *in fine* affecté à la nouvelle tranche « De 1946 à 1970 » et 4 chances sur 26 d'être affecté à la nouvelle tranche « De 1971 à 1990 ».

Ce principe simple permettrait d'aller au terme de l'imputation si on ne se heurtait pas au problème posé par la présence (bien entendu inévitable) d'une tranche inférieure non bornée, à savoir la tranche « Avant 1919 » de la nouvelle nomenclature. C'est à ce niveau qu'intervient l'enquête Logement, qui permet pour sa part d'estimer la distribution de l'année d'achèvement du logement selon dix tranches : les trois tranches les plus anciennes, qui sont les seules qui nous intéressent, distinguent « Avant 1871 », « De 1871 à 1914 » et « De 1915 à 1948 ». Même dans cette enquête en face-à-face, il serait totalement vain de chercher à obtenir une année d'achèvement, que la plupart des enquêtés ne connaissent pas – ou très approximativement – dès lors que l'immeuble n'est pas récent. Le fait qu'une des bornes de la nomenclature associée à l'enquête *Logement* soit l'année 1948, c'est-à-dire exactement la borne de la tranche A qui pose un problème d'imputation, constitue un atout majeur. En effet, considérant les proportions estimées de logements caractérisant les trois premières tranches associées à l'enquête *Logement*, et toujours sur la base d'une hypothèse de répartition uniforme parmi les années constituant chaque tranche, on peut ventiler aléatoirement les logements recensés dans l'ancienne nomenclature entre les tranches «Avant 1871 », « De 1871 à 1914 » et « De 1915 à 1948 ». Les deux premières modalités envoient systématiquement le logement dans la nouvelle tranche « Avant 1919 ». La troisième modalité (couvrant 34 années) offre trois cas de figure : avec 4 chances sur 34 le logement est déclaré construit entre 1915 et 1918 (tranche finale imputée « Avant 1919 »), avec 27 chances sur 34 il est déclaré construit entre 1919 et 1945 (tranche finale imputée « De 1919 à 1945 »), et avec 3 chances sur 34 il est déclaré construit entre 1946 et 1948, auquel cas on l'affecte à la tranche finale « De 1946 à 1970 ».

Pour procéder concrètement à l'imputation, on a tiré pour chaque immeuble un nombre au hasard compris entre 0 et 1 et on a imputé à chaque logement de l'immeuble une modalité fonction du positionnement de ce nombre par rapport aux probabilités associées aux tranches. Reprenant l'exemple ci-dessus d'un logement préalablement affecté à la tranche « De 1915 à 1948 » à l'issue de la première phase (aléatoire) de l'imputation, si le nombre au hasard est inférieur à 4/34 le logement est déclaré construit entre 1915 et 1918, s'il est compris entre 4/34 et 31/34 il se trouvera en tranche « De 1919 à 1945 », et s'il est supérieur à 31/34 il se trouvera en tranche « De 1946 à 1970 ». S'agissant d'une approche par modèle et pour la raison qui a été donnée au sujet de l'indicateur de résidence antérieure, aucune procédure de réduction de la variabilité de l'imputation n'a été mise en place. Le nombre au hasard généré est le même pour chaque logement de l'immeuble, ce qui permet de respecter une logique : tous les logements d'un même immeuble devraient être imputés de la même façon. En réalité, le traitement standard du recensement n'assurant pas lui-même cette logique, l'imputation ne rectifie pas la situation, mais du moins peut-on garantir qu'elle ne la dégrade pas davantage.

La structure des logements selon leur date d'achèvement, telle qu'elle ressort de l'enquête *Logement* 2006, ne peut pas raisonnablement s'apprécier sur l'ensemble des communes, car l'hétérogénéité du terrain est considérable. On a donc constitué en amont une typologie de communes, assez grossière parce qu'il faut conserver une taille d'échantillon totale répondante issue de l'enquête suffisamment grande dans chaque groupe de communes. De fait, une classification hiérarchique préalable a permis de distinguer trois groupes de communes à partir de la structure communale des logements selon leur période d'achèvement issue du recensement 2008 (s'agissant du recensement le plus récent offrant une homogénéité parfaite de la nomenclature de collecte de la période de construction). Dans le premier groupe de communes, 65 % environ des logements sont construits avant 1949 (communes à habitat essentiellement ancien), dans le second on trouve 50 % des logements construits après 1975 (communes à habitat essentiellement récent) et le troisième groupe est caractérisé par une proportion de 40 % de logements achevés entre 1949 et 1974.

Chaque commune de France, quelle que soit sa taille et y compris dans les DOM, étant affectée à l'un de ces trois groupes, l'imputation utilise la structure estimée issue de l'enquête *Logement* 2006 spécifique à ce groupe (impliquant seulement les trois premières tranches de l'enquête). Le cas des DOM, qui possèdent extrêmement peu de constructions anciennes (avant 1915), a été traité à part de la métropole : l'enquête *Logement* 2006 offre un échantillon de 5 700 logements répondants dans les DOM, ce qui a permis d'appliquer aux logements des DOM une structure estimée à partir de ce sous-échantillon spécifique (mais sans faire de distinction entre les quatre DOM).

Le tableau 9 fournit, pour chaque *EAR* constituant le *RP* 2010, en utilisant le poids du *RP* 2010 (et non le poids des *EAR*), la structure du parc de logements après imputation selon la (nouvelle) période d'achèvement - en se limitant aux logements construits avant 2006 afin qu'il n'y ait pas l'effet perturbateur des constructions nouvelles, qui pèseraient bien entendu très différemment selon l'*EAR* considérée.

À titre de validation des ordres de grandeur, les résultats après imputation peuvent être utilement rapprochés de ceux de l'enquête *Logement* 2006, malgré les différences de définition des tranches (cf. tableau 10).

Au niveau national, on constate un léger décalage de structure entre, d'une part les deux *EAR* en nouvelle nomenclature, et d'autre part les trois *EAR* en ancienne nomenclature. Cela n'est pas du tout surprenant puisque l'imputation relève d'une approche par modèle, autrement dit d'une hypothèse portant sur la distribution de la variable année d'achèvement. D'autres causes doivent y contribuer (erreur d'échantillonnage, variance d'imputation, rôle des poids, erreur de mesure). Néanmoins, le résultat d'ensemble apparaît tout à fait satisfaisant compte tenu des hypothèses assez audacieuses qui ont été faites pour traiter le cas des tranches inférieures – même si, au niveau local, on imagine facilement que des écarts plus sensibles puissent survenir dans certaines circonstances. On pourrait donc s'en satisfaire. Cela étant, il est possible d'explorer une méthode alternative qui devrait réduire encore un peu ces écarts de structure. Dans cet esprit, on trouvera en annexe 1 un développement basé sur une technique de calage, à l'image de ce qui a été fait pour la variable de résidence antérieure.

En conclusion de cette partie, on rappelle que l'appréciation de la pertinence de la modélisation requiert des statistiques jugées fiables et de même nature que celles que l'on produit. Au niveau local, pour les petites communes, de tels éléments n'existent pas : aucune opération

Tableau 9
Structure du parc de logements après imputation, selon la période d'achèvement

En %

EAR	Période d'achèvement de la construction (nouvelle nomenclature)					
	Avant 1919	1919-1945	1946-1970	1971-1990	1991-2005	Total
2008	20,0	11,0	25,0	28,1	15,9	100
2009	19,9	11,0	25,2	28,2	15,7	100
2010	19,7	10,8	25,3	28,5	15,7	100
2011	19,0	10,8	23,5	30,4	16,3	100
2012	19,0	10,9	23,5	30,5	16,1	100
Total	19,5	10,9	24,5	29,1	16,0	100

Lecture : le sous-échantillon EAR *2010 pondéré avec les poids du* RP *2010, après imputation, estime à 19,7 % la proportion des logements déclarés construits avant 1919.*
Champ : France entière, ensemble des logements recensés.
Sources : RP *2010 et enquête* Logement *2006.*

Tableau 10
Structure du parc de logements d'après l'enquête *Logement* 2006

En %

Avant 1915	1915-1948	1949-1967	1968-1989	1989-2006	Total
16,8	13,1	17,2	35,4	17,5	100

Lecture : l'enquête Logement *2006 estime à 16,8 % la proportion des logements construits avant 1915.*
Champ : France entière, ensemble des logements ordinaires.
Source : enquête Logement *2006.*

statistique suffisamment récente menée dans les communes de moins de 10 000 habitants recensées en 2008, 2009 ou 2010 n'a collecté une information permettant d'estimer la structure des logements selon la nomenclature adoptée à partir de 2011. C'est évidemment différent dans les grandes communes, parce qu'il existe deux *EAR* (2011 et 2012) qui permettent d'obtenir l'estimation en question. Néanmoins, au niveau d'une commune, l'appréciation des biais d'imputation reste délicate et nécessiterait en toute rigueur d'estimer les erreurs d'échantillonnage associées au cumul des deux *EAR* de référence. Une appréciation à vue des décalages numériques peut néanmoins paraître suffisante, au moins dans les plus grosses communes. Mais la méthode d'appréciation de la qualité la plus convaincante reste encore celle qui s'appuie sur la statistique agrégée, au niveau national, voire éventuellement régional. La variable « Période d'achèvement » n'ayant fait l'objet d'aucune contrainte de cohérence au cours de cette opération, il n'est pas étonnant que des écarts à la structure issue des *EAR* 2011 et 2012 existent, à tous les niveaux géographiques. C'est le tableau 9 qui résume le mieux la situation de ce point de vue, et c'est lui qui fournit donc la meilleure conclusion sur l'appréciation de la qualité. Partant de là, en l'absence de seuil d'erreur, chaque utilisateur appréciera la situation par rapport à son objectif.

L'activité au sens du Bureau International du Travail

L'activité au sens du Bureau international du Travail (BIT) relève d'une définition complexe et très précise qu'il est exclu d'appréhender en mobilisant l'information du bulletin individuel (BI). Il est donc nécessaire de recourir, d'une façon ou d'une autre, à la seule source de données qui traite de l'activité au sens du BIT, à savoir l'enquête *Emploi*. Il s'agit d'une enquête par sondage à échantillonnage complexe, effectuée selon un schéma rotatif au auprès de la seule population vivant en ménage ordinaire. Dès lors qu'ils ont un âge au moins égal à 15 ans au 31 décembre de l'année, tous les occupants de chaque logement échantillonné en métropole sont interrogés six trimestres consécutifs, en face-à-face le premier et le sixième trimestre, et par téléphone durant les quatre trimestres intermédiaires. Dans les DOM, jusqu'au début 2013, l'enquête de tout logement échantillonné a lieu une fois par an durant trois années consécutives. On trouvera dans Givord (2003) un descriptif de cette enquête. Eurostat a demandé à ce que l'activité au sens du BIT constitue une des informations produites par le recensement de 2011 – la situation française conduisant à traiter le *RP* 2010 et à procéder à une adaptation ultérieure des poids. En la circonstance, il ne s'agit plus de gérer un changement de variable comme dans les deux cas précédents, mais de construire intégralement une information pour chaque individu de 15 ans ou plus au moment de la collecte (ce champ d'imputation a été fixé par l'équipe en charge du recensement), quel que soit le type de ménage dans lequel il se trouve. Pour des raisons de recherche de cohérence entre les sources, l'Insee a souhaité que la procédure d'imputation permette de s'approcher le plus possible, pour la population vivant en ménage ordinaire, des taux de chômage BIT par sexe de chaque région, en métropole comme dans les DOM (la Corse a été regroupée avec la région PACA). Cet objectif a constitué la seule contrainte préalable à l'opération mais nous l'avons traduit en cherchant à respecter au niveau région / sexe les effectifs respectifs d'actifs occupés, de chômeurs et d'inactifs, appelés par la suite « effectifs-cibles ».

Dans le recensement, on pose à toute personne de 14 ans ou plus la question « *Quelle est votre situation principale ?* » (question n° 10) et les modalités offertes permettent de distinguer, après regroupement, les actifs occupés, les chômeurs et les inactifs. On qualifiera désormais cette variable « d'activité spontanée ». Elle produit des statistiques très différentes de celles qui résultent du traitement de l'enquête *Emploi*, qui distingue les mêmes modalités, mais cette fois au sens du BIT. Le principe général de la méthode consiste à estimer des probabilités de passage d'un statut individuel construit à partir d'un ensemble de variables du BI (variables à déterminer) vers le statut d'activité (prédit) au sens du BIT, en mobilisant l'information disponible dans l'enquête *Emploi*. Cette phase essentielle doit s'appuyer sur des variables explicatives de la situation d'activité BIT : on se place donc dans une optique de modélisation. Ensuite, il convient d'appliquer ces probabilités de passage (on parlera désormais de probabilités de transition) aux individus recensés. Cela suppose qu'il existe au moins une variable suffisamment explicative de l'activité BIT qui soit commune au questionnaire de l'enquête *Emploi* et au BI du recensement : en effet, si le modèle se construit entièrement à partir des données issues de l'enquête *Emploi*, l'imputation s'applique *in fine* à l'intégralité du fichier du

recensement. En dernière étape, pour la population vivant en ménage ordinaire, il est nécessaire de trouver une technique qui permette d'obtenir des dénombrements aussi proches que possible des effectifs-cibles par croisement région / sexe issus de l'enquête *Emploi* : cela se fait par ajustement des probabilités de transition obtenues en phase précédente. Sur le plan temporel, en métropole, il a été décidé de former les effectifs-cibles de l'enquête *Emploi* en agrégeant les trimestres 3 et 4 de 2010 ainsi que les trimestres 1 et 2 de 2011. En effet, ces quatre trimestres agrégés permettent de former des estimateurs suffisamment fiables (un trimestre concerne environ 110 000 personnes répondantes, il y a un recouvrement de 5/6 de l'échantillon entre deux trimestres consécutifs mais les variables collectées sont propres au trimestre) et ils encadrent harmonieusement la date de référence qu'est le 1er janvier 2011. Pour les DOM, on utilisera les enquêtes annuelles *Emploi* de 2010, 2011 et 2012. Côté *Recensement*, c'est sur l'année 2010 que sera effectuée l'imputation. On rappelle que ce décalage d'une année relève d'une décision stratégique de l'Insee, dictée par le calendrier de mise à disposition du *RP* 2011.

Étape 1 : sélection des variables expliquant l'activité BIT (cas de la métropole)

On recherche dans cette première étape les variables explicatives de l'état de chômage, cette modalité s'affirmant comme primordiale. Pour le moment, on se limite à la métropole et on considère uniquement les personnes de moins de 75 ans au 31 décembre de l'année de collecte et vivant en ménage ordinaire. L'enquête *Emploi* ne s'adressant qu'aux individus ayant au moins 15 ans au 31 décembre de l'année de collecte, lesquels forment par ailleurs un champ compatible avec le règlement européen organisant la mise à disposition des données, on limite systématiquement les procédures, dans toute la suite, aux personnes appartenant au champ de l'enquête *Emploi*. L'activité spontanée déclarée dans le BI trouve son équivalent, du moins en terme de formulation de la question, à la fin du questionnaire de l'enquête *Emploi* : qu'il s'agisse d'une collecte en face-à-face ou par téléphone, on demande à l'enquêté « *Quelle est ce mois-ci votre situation principale ?*», en lui offrant des modalités qui sont (presque) identiques à celles de la question du BI relative à la déclaration spontanée. Le tableau 11 compare les distributions métropolitaines de la variable de déclaration spontanée d'activité selon les sources, *EAR* 2011 d'une part et enquête *Emploi* du 1er trimestre 2011 d'autre part, afin de s'en tenir à des sources qui prétendent quantifier des phénomènes à la même période. On y ajoute pour information la répartition estimée des individus selon leur activité au sens du BIT.

Compte tenu de l'ordre de grandeur de l'erreur d'échantillonnage affectant les estimations d'effectifs dans l'enquête *Emploi*, les écarts entre les déclarations spontanées respectives de l'*EAR* et de l'enquête *Emploi* ne s'expliquent pas seulement pas des effets d'échantillonnage : il y a également un effet assez marqué du mode et du contexte de collecte. On constatera en outre un écart considérable entre l'approche spontanée et la définition BIT, en particulier lorsqu'on dénombre les chômeurs (25 % d'effectifs en plus pour la déclaration spontanée). Néanmoins, considérant comme acceptables les écarts obtenus entre les effectifs (associés à la déclaration spontanée) donnés par l'*EAR* et ceux de l'enquête *Emploi*, nous avons retenu cette variable de déclaration spontanée comme

Tableau 11
Par statut d'activité, effectifs estimés selon le concept et selon la source

	Déclaration spontanée d'activité		Activité au sens du BIT
	Selon l'*EAR* 2011	Selon l'enquête *Emploi* (1er trimestre 2011)	Selon l'enquête *Emploi* (1er trimestre 2011)
Actifs occupés	24 228 000	25 450 000	25 550 000
Chômeurs	3 490 000	3 363 000	2 684 000
Inactifs	17 249 000	16 430 000	17 009 000
Total	44 967 000	45 243 000	45 243 000

*Lecture : selon l'*EAR *2011, on estime à 3 490 000 le nombre total de chômeurs lorsqu'on exploite la déclaration spontanée d'activité (poids de l'*EAR*). Cet effectif est porté à 3 363 000 si on utilise l'enquête* Emploi*. Si on comptabilise des chômeurs au sens du BIT, l'enquête* Emploi *en estime l'effectif global à 2 684 000.*
Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Source : EAR *2011 et enquête* Emploi *du 1er trimestre 2011.*

facteur explicatif essentiel de l'activité au sens du BIT. Nous avons considéré qu'elle peut être utilisée, d'une part à partir de la source *Emploi* pour construire un modèle prédictif de l'activité BIT, d'autre part à partir de la source *RP* pour procéder à l'imputation proprement dite en tant que facteur prédictif principal. Il doit être clair qu'il ne s'agit en aucun cas, dans notre contexte, d'attribuer un caractère explicatif permettant de construire une interprétation socio-économique de l'activité, mais plutôt de rechercher des corrélations entre variables : autrement dit, ce qui compte ici est de s'assurer que si un individu se déclare spontanément chômeur dans une enquête, il est probable qu'*in fine* sa situation réelle au sens du BIT soit bel et bien un état de chômage. Il n'est donc pas rédhibitoire que la déclaration spontanée d'activité d'un individu ne fournisse aucune information sur son activité au sens du BIT. Il n'est pas nécessaire de comprendre le mécanisme qui sous-tend le passage d'un concept à un autre, mais il faut et il suffit que ce mécanisme s'affirme suffisamment régulier dans le temps et dans l'espace.

Une tentative a été faîte de construire une autre variable explicative de l'activité BIT, plus analytique mais aussi plus compliquée. Elle combinait, du coté du recensement, les variables n° 11 « *Travaillez-vous actuellement ?* » et n° 16 « *Cherchez-vous un emploi ?* ». Du coté de l'enquête *Emploi*, il s'agissait d'exploiter les trois questions centrales suivantes : « *Durant la semaine du lundi…au dimanche…, avez-vous effectué ne serait-ce qu'une heure de travail rémunéré ?* », ainsi que, en cas de réponse négative[2], « *Avez-vous cependant un (autre) emploi ?* », suivi de « *Pour quelle raison n'avez-vous pas travaillé dans le cadre de cet emploi ?* », laquelle question offrait 11 modalités de réponse. Nous avons abouti à la formation d'une variable synthétique en trois modalités qui s'avérait en théorie prometteuse pour jouer le rôle de variable explicative de l'activité BIT, ce que nous avons d'ailleurs pu vérifier concrètement dans la source *Emploi* parce que cette variable alternative est sensiblement mieux corrélée à l'activité BIT que ne l'est la déclaration spontanée. Mais côté *RP*, il y avait trop d'erreurs de mesure et trop de non-réponses pour que l'on puisse considérer, au vu des effectifs par modalité, la nouvelle variable comme comparable à celle construite à partir de la source *Emploi*. Cette tentative n'a donc pas eu de suite, la déclaration spontanée restant pour nous la seule variable explicative envisageable pour fonder nos prédictions.

Le tableau 12 donne, pour l'ensemble de l'échantillon *Emploi* du 1er trimestre de 2011, les effectifs non pondérés des croisements des modalités (regroupées) des variables respectives de déclaration spontanée et d'activité au sens du BIT. Comme on pouvait s'y attendre, la matrice est presque diagonale, même si la frontière entre les statuts de chômeur et d'inactif s'avère relativement perméable : d'assez nombreuses déclarations spontanées de chômage se transforment *in fine* en statut d'inactivité. Ce déséquilibre représente la difficulté essentielle puisque l'imputation au niveau de l'ensemble de la métropole doit se conclure dans le champ considéré par environ 800 000 chômeurs en moins par rapport à la situation que reflète la déclaration spontanée d'activité (cf. tableau 11)

Lors de la phase de mise au point du processus d'imputation, il s'est avéré impossible de trouver des probabilités de transition satisfaisantes entre activité spontanée et activité BIT : toutes

2. ou positive, dans le cas où l'enquêté précise ensuite qu'il s'agit d'un « petit boulot ».

Tableau 12
Corrélation entre activité spontanée et activité au sens du BIT

		Déclaration spontanée			
		Actif occupé	Chômeur	Inactif	Total
Activité au sens du BIT	Actif occupé	50 228	206	681	51 115
	Chômeur	136	4 982	354	5 472
	Inactif	576	1 669	35 084	37 329
	Total	50 940	6 857	36 119	93 916

Lecture : parmi les 93 916 individus répondants de l'échantillon Emploi *dans le champ considéré, 1669 se déclarent spontanément chômeurs mais s'avèrent être des inactifs au sens du BIT.*
Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Source : enquête Emploi *du 1er trimestre 2011 (non pondérée).*

les tentatives pour retrouver les effectifs-cibles de chômeurs par croisement sexe / région ont échoué. Plus précisément, ces probabilités de transition, dont l'expression analytique sera détaillée plus loin, sortaient en trop grand nombre de l'intervalle [0-1]. Elles étaient trop nombreuses à être négatives ou supérieures à 1 pour que l'on puisse accepter de poursuivre le processus sans adaptation. Ces valeurs aberrantes se sont avérées résulter d'une trop grande différence d'effectifs, au niveau des croisements considérés, entre l'activité spontanée selon l'enquête *Emploi* et l'activité spontanée selon le recensement : il était donc nécessaire de trouver un processus pour rapprocher encore les deux concepts. Nous avons donc mis au point une phase de correction qui a pris la forme d'une nouvelle variable (dite « variable intermédiaire ») construite dans l'enquête *Emploi* : il s'agissait de créer une variable qui simulerait la réponse des enquêtés de l'enquête Emploi comme s'ils remplissaient un bulletin individuel du recensement. Pour ce faire, on définit en amont et au niveau de l'ensemble du champ considéré un système de probabilités de transition de la déclaration spontanée d'activité au sens de l'enquête *Emploi* vers la variable intermédiaire d'activité que l'on veut créer, respectant en moyenne la distribution de la variable d'activité spontanée au sens du recensement (cf. encadré 2). On applique ensuite la technique traditionnelle : pour chaque individu enquêté, on génère un nombre au hasard dans la loi uniforme sur [0,1], tirage effectué indépendamment d'un individu à l'autre. Le positionnement de ce nombre par rapport aux valeurs des probabilités de transition permet d'affecter une modalité d'activité intermédiaire à l'individu considéré.

Afin d'améliorer la capacité prédictive du modèle, nous avons introduit en tant que facteurs explicatifs *a posteriori* et en sus de l'activité spontanée des variables sociodémographiques simples disponibles à la fois dans le BI et dans l'enquête *Emploi* : : sexe (2 modalités), tranche d'âge (5 modalités), et diplôme (4 modalités). Les modalités définitives de chacune de ces variables ont été fixées à partir

Encadré 2

DÉTERMINATION DE LA VARIABLE INTERMEDIAIRE D'ACTIVITÉ

On se place sur l'ensemble du champ traité et on paramètre les probabilités de transition ainsi pour chacun des sexes :

$1-\theta$ = probabilité de se déclarer actif occupé au *RP* sachant que l'on se déclare actif occupé à l'enquête *Emploi*.

$1-\mu$ = probabilité de se déclarer chômeur au *RP* sachant que l'on se déclare chômeur à l'enquête *Emploi*.

$1-\delta$ = probabilité de se déclarer inactif au *RP* sachant que l'on se déclare inactif à l'enquête *Emploi*.

Toutes les déclarations dont il est question ici sont les déclarations spontanées. Pour définir complètement le système probabiliste assurant les liaisons entre les deux variables, on postule que la probabilité de se déclarer chômeur au *RP* sachant que l'on se déclare actif occupé à l'enquête *Emploi* est égale à la probabilité de se déclarer inactif au *RP* sachant que l'on se déclare actif occupé à l'enquête Emploi. Cette probabilité commune est alors nécessairement égale à $\theta / 2$. Cet équilibre traduit notre manque total d'informations *a priori* sur ce qui différencie les comportements de déclaration spontanée entre les deux sources en jeu. On pratique de manière similaire pour les modalités chômeur et inactif à l'enquête *Emploi*. On note $N^{RP}_{actif\_occupé}$ le nombre total d'individus dans le champ (pour le sexe considéré) qui se déclarent spontanément actifs occupés au *RP* (les autres notations relèvent de la même logique, EEC renvoyant à l'enquête *Emploi*). L'expérience montre que ce schéma est satisfaisant, ce qui fait qu'on aboutit à une table de transition qui offre, en ce qui concerne les probabilités de transition, la structure reproduite dans le tableau ci-dessous.

Tableau
Schéma de transition pour construire la variable intermédiaire d'activité

$1-\theta$	$\mu/2$	$\delta/2$	$N^{RP}_{actif\_occupé}$
$\theta/2$	$1-\mu$	$\delta/2$	$N^{RP}_{chômeur}$
$\theta/2$	$\mu/2$	$1-\delta$	$N^{RP}_{inactif}$
$N^{EEC}_{actif\_occupé}$	$N^{EEC}_{chômeur}$	$N^{EEC}_{inactif}$	N

➔

des résultats d'une régression logistique qui a permis de regrouper en amont les modalités initiales les plus semblables. La situation définitive distingue, en dehors du sexe, trois classes d'âge (limites de tranches fixées respectivement à 49 ans et à 59 ans) et trois classes de diplôme (selon le positionnement par rapport au « brevet des collèges » et aux diplômes de niveau « bac plus deux ans »).

Étape 2 : calcul des probabilités de transition initiales

Les trois variables sociodémographiques sélectionnées permettent donc de distinguer dans un premier temps 18 sous-populations, au sein de chacune desquelles on peut quantifier, grâce aux données de l'enquête *Emploi*, la corrélation entre la variable intermédiaire d'activité spontanée et la variable d'activité BIT. Les probabilités de transition sont obtenues en utilisant la pondération de l'enquête *Emploi* (agrégeant les trimestres 3 et 4 de 2010 et les trimestres 1 et 2 de 2011), s'agissant par définition de ratios de deux effectifs estimés, définis sous-population par sous-population : par exemple, on va se placer dans une sous-population, on va estimer le nombre d'individus chômeurs au sens de la variable intermédiaire et le nombre d'individus à la fois chômeurs au sens de la variable intermédiaire et inactifs au sens du BIT : le ratio du second effectif estimé sur le premier effectif estimé constitue l'estimation de la probabilité de transition de l'état de chômeur selon le concept intermédiaire vers l'état d'inactif selon le concept BIT. Dans chaque sous-population, on a affaire à neuf probabilités de transition, puisqu'on distingue trois modalités d'activité pour chaque concept. Néanmoins, il faut absolument tenir compte de la taille de l'échantillon de l'enquête *Emploi* à partir de laquelle on estime ces probabilités, car une petite taille d'échantillon produit des estimations instables, c'est-à-dire très (trop) sensibles à la composition de l'échantillon. Il est préférable, lorsque la taille d'échantillon répondant apparaît inférieure à un certain seuil, d'effectuer des regroupements de sous-populations. Ces regroupements peuvent être décidés empiriquement, en tenant compte des probabilités d'inclusion initiales, même si leur fiabilité est un peu dégradée. Cette stratégie conduit à distinguer dix sous-populations (cf. tableau 13).

Encadré 2 (suite)

Le respect des contraintes portant sur les effectifs, au niveau de l'ensemble de la métropole mais en distinguant le sexe, se traduit par un jeu d'équations linéaires simple :

$$(1-\theta)\cdot N^{EEC}_{actif\_occupé} + \frac{\mu}{2}\cdot N^{EEC}_{chômeur} + \frac{\delta}{2}\cdot N^{EEC}_{inactif} = N^{RP}_{actif\_occupé}$$

$$\frac{\theta}{2}\cdot N^{EEC}_{actif\_occupé} + (1-\mu)\cdot N^{EEC}_{chômeur} + \frac{\delta}{2}\cdot N^{EEC}_{inactif} = N^{RP}_{chomeur}$$

$$\frac{\theta}{2}\cdot N^{EEC}_{actif\_occupé} + \frac{\mu}{2}\cdot N^{EEC}_{chômeur} + \frac{\delta}{2}\cdot N^{EEC}_{inactif} = N^{RP}_{inactif}$$

Comme

$$N^{EEC}_{actif\_occupé} + N^{EEC}_{chômeur} + N^{EEC}_{inactif} = N^{RP}_{actif\_occupé} + N^{RP}_{chomeur} + N^{RP}_{inactif} = N,$$

la troisième équation se déduit des deux premières. On peut alors tirer θ et μ en fonction du paramètre δ : cette dernière valeur paramètre donc à elle seule l'ensemble du processus de transition. On peut opter pour toute valeur comprise entre 0 et 1, sous condition expresse que θ et μ soient tous deux également compris entre 0 et 1, mais notre préférence va à une petite valeur numérique de δ (la plus proche possible de zéro) afin de limiter l'ampleur des changements de statut (donc rapprocher au maximum le tableau ci-dessus d'une matrice diagonale). La solution du système d'équations ci-dessus est :

$$\theta = \frac{N^{EEC}_{inactif}}{N^{EEC}_{actif\_occupé}}\cdot\delta - \frac{2}{3}\cdot\frac{1}{N^{EEC}_{actif\_occupé}}\cdot \left(N^{RP}_{chomeur} - N^{EEC}_{chomeur} + 2\cdot N^{RP}_{actif\_occupé} - 2\cdot N^{EEC}_{actif\_occupé}\right)$$

$$\mu = \frac{N^{EEC}_{inactif}}{N^{EEC}_{chomeur}}\cdot\delta - \frac{2}{3}\cdot\frac{1}{N^{EEC}_{chomeur}}\cdot \left(2\cdot N^{RP}_{chomeur} - 2\cdot N^{EEC}_{chomeur} + N^{RP}_{actif} - N^{EEC}_{actif}\right)$$

L'application numérique donne, pour les hommes (au niveau national, pour la métropole)

$$\theta = 0{,}529\cdot\delta + 0{,}049 \quad \text{et} \quad \mu = 4{,}497\cdot\delta + 0{,}119$$

et pour les femmes (au niveau national, pour la métropole)

$$\theta = 0{,}754\cdot\delta + 0{,}064 \quad \text{et} \quad \mu = 5{,}467\cdot\delta + 0{,}257$$

Si δ est proche de 0, comme on le souhaite, on vérifie que les probabilités θ et μ sont également assez proches de 0 (excepté pour le cas des femmes inactives), ce qui rend la matrice de transition proche d'une matrice diagonale et conforte l'idée qu'il y a, dans l'ensemble, une forme de similarité entre les deux variables de déclaration spontanée.

On donne dans le tableau 14 les valeurs des probabilités de transition estimées, pour les dix catégories de population distinguées – la déclaration spontanée issue de l'enquête *Emploi* étant ici la variable dite intermédiaire. Par exemple, considérant un individu de la sous-population 1 (un homme de 15 à 49 ans, sans diplôme ou ayant le brevet des collèges), si cet individu se déclare spontanément chômeur (au sens de la variable intermédiaire d'activité construite dans la source *Emploi* – mais, par construction, on suppose qu'il aurait la même attitude au recensement…), il y a 70,7 chances sur 100 pour qu'au sens du BIT il finisse par être classé comme chômeur – tandis qu'avec 11,7 chances sur 100 il sera classé actif occupé et avec 17,6 chances sur 100, inactif. Cette grille complète des probabilités de transition reflète toute l'hétérogénéité qui existe entre ces deux concepts, d'une part ce qui vient de la déclaration spontanée, d'autre part ce qui résulte d'un processus de classement assez sophistiqué, reposant sur des critères multiples et bien précis.

La ventilation en sous-populations – qui ne vise pas à être « optimale » – se justifie *ex post* par la diversité des valeurs numériques des probabilités de transition figurant dans le tableau 14.

Étape 3 : calcul des probabilités de transition définitives

L'étape précédente aboutit à la production de probabilités de transition qui fournissent des valeurs individuelles imputées tout à fait satisfaisantes lorsqu'elles sont appliquées à la variable de situation d'activité spontanée disponible dans

Tableau 13
Catégories de population utilisées en métropole pour l'estimation des probabilités de transition

Catégorie	Sexe	Tranche d'âge	Classe de diplôme
1	Homme	15 à 49 ans	Diplôme <= brevet
2	Homme	15 à 49 ans	Bac à bac +2
3	Homme	15 à 49 ans	Bac + 2 < diplôme
4	Homme	50 à 59 ans	Tous diplômes
5	Tous sexes	60 à 74 ans	Tous diplômes
6	Femme	15 à 49 ans	Diplôme <= brevet
7	Femme	15 à 49 ans	Bac à bac +2
8	Femme	15 à 49 ans	Bac + 2 < diplôme
9	Femme	50 à 59 ans	Diplôme <= brevet
10	Femme	50 à 59 ans	Bac < = diplôme

Tableau 14
Probabilités de transition initiales estimées par catégorie de population

En %

Activité spontanée « intermédiaire »	Activité BIT	Catégorie de population									
		1	2	3	4	5	6	7	8	9	10
Actif occupé	Actif occupé	96,9	98,7	99,1	98,5	97,0	92,3	96,2	97,5	96,4	98,1
Actif occupé	Chômeur	1,8	0,8	0,6	0,6	0,5	3,8	1,6	1,1	1,5	0,9
Actif occupé	Inactif	1,3	0,5	0,3	0,9	2,5	3,9	2,2	1,3	2,1	1,0
Chômeur	Actif occupé	11,7	25,0	32,1	25,7	23,7	14,4	29,3	39,2	26,5	39,9
Chômeur	Chômeur	70,7	59,4	54,9	52,0	20,0	63,0	53,0	46,9	44,6	43,7
Chômeur	Inactif	17,6	15,6	13,0	22,3	56,3	22,6	17,7	13,8	28,9	16,3
Inactif	Actif occupé	4,6	18,2	31,5	13,4	1,2	3,4	13,9	23,9	7,4	13,8
Inactif	Chômeur	2,8	4,9	7,3	3,2	0,2	3,9	6,6	6,8	3,3	3,8
Inactif	Inactif	92,6	76,9	61,1	83,4	98,6	92,8	79,5	69,3	89,3	82,4

Lecture : si un individu de la sous-population 1 (homme de 15 à 49 ans, sans diplôme ou dont le niveau de diplôme ne dépasse pas celui du brevet) se déclare spontanément chômeur au sens de la variable intermédiaire, il y a 11,7 chances sur 100 pour qu'il soit actif occupé au sens du BIT.
Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Source : enquête Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

la source recensement. En particulier, au niveau national (métropole), les effectifs estimés par modalité d'activité s'avèrent numériquement proches des estimations données par l'enquête *Emploi* (cumul de quatre trimestres). Cela étant, lorsqu'on descend au niveau d'un croisement région / sexe, les écarts d'estimation d'effectifs entre les deux sources sont parfois substantiels ; en tout cas, ils ne respectent pas suffisamment l'objectif fixé par l'Insee. C'est pourquoi il a fallu ajouter une ultime étape pour modifier ces probabilités de transition initiales. L'une des évolutions majeures a consisté à faire dépendre le jeu des probabilités définitives de la région – faute de quoi il devenait impossible de satisfaire les contraintes. Dans ce cadre, nous avons recherché un jeu de nouvelles probabilités aussi proches que possible des probabilités initiales issues de l'étape 2, mais qui permettaient de respecter en moyenne les effectifs-cibles régionaux par sexe. En particulier, les paramètres techniques introduits visent à élargir le champ de la stratégie et, si possible, à se rapprocher des effectifs-cibles. L'encadré 3 précise les grandes lignes de la méthode. Comme on pouvait s'y attendre, l'ampleur des évolutions numériques des probabilités de transition dépend de l'écart entre les effectifs-cibles et les effectifs imputés produits « en moyenne » par les probabilités initiales. Si cet écart est trop important, la solution mathématique du problème peut sortir de l'intervalle [0,1]. Dans ce cas, on force les probabilités concernées à la valeur admissible la plus proche, à savoir, selon les circonstances, soit 0, soit 1. La brutalité de cette ultime correction est l'une des explications des écarts résiduels entre effectifs-cibles et effectifs imputés au niveau région / sexe – une autre raison étant, bien entendu, la variabilité liée au processus d'imputation proprement dit, dont il faut rappeler qu'il est de nature aléatoire. Compte tenu du fait que l'on dénombre 6 catégories pour les femmes et 5 catégories pour les hommes, croisées avec 21 régions (on rappelle que la Corse est regroupée avec la région PACA), pour chacun des 9 types de transition on distingue $21\times(6+5) = 231$ valeurs de probabilité. L'ampleur des cas à problème est donnée par le tableau 15 qui précise les maximum et minimum des 231 valeurs de la distribution associée à chaque type de transition.

Au total, on a donc calculé 231×9 soit 2079 probabilités de transition, lesquelles ont remplacé les 90 probabilités de transition initiales. L'utilisation de nombres aléatoires tirés « au hasard » entre 0 et 1 et leur positionnement par rapport aux probabilités définitives achève le processus d'imputation. Comme cela a été expliqué pour l'indicateur de résidence antérieure, l'aléa d'imputation n'a pas été contrôlé ; au demeurant, il n'aurait pu l'être qu'aux niveaux région / sexe puisqu'il n'y a pas eu de contraintes portant sur d'autres croisements. Mais là encore, les tailles des populations imputées sont considérables par région / sexe et de ce fait l'aléa d'imputation génère une erreur négligeable par rapport au biais qui est attaché au modèle (et qui constitue le véritable problème). Les résultats qui suivent le confirment puisqu'on atteint sans mal les objectifs avec la technique aléatoire utilisée, quitte à effectuer quelques tirages successifs si on veut raffiner (technique du tirage dit « rejectif »). La complexité substantielle de

Tableau 15
Statistiques concernant les probabilités de transition définitives

Transition	Minimum	Maximum	Problème ?
Actif occupé → actif occupé	0,914	1,044	OUI
Chômeur → actif occupé	0,113	0,402	
Inactif → actif occupé	- 0,014	0,317	OUI
Actif occupé → chômeur	- 0,007	0,042	OUI
Chômeur → chômeur	0,200	0,710	
Inactif → chômeur	- 0,018	0,074	OUI
Actif occupé → inactif	- 0,049	0,059	OUI
Chômeur → inactif	0,129	0,563	
Inactif → inactif	0,610	1,012	OUI

Lecture : parmi les 231 probabilités de transition du statut inactif vers le statut chômeur, la valeur minimale de la distribution est - 0,018 et la valeur maximale est 0,074. Il y a donc un problème à traiter (probabilité négative).
Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Sources : RP *2010 et enquête* Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

Encadré 3

DETERMINATION DES PROBABILITÉS DE TRANSITION DÉFINITIVES

Le croisement de la région et du sexe est appelé « domaine » et repéré par l'indice *d*. Comme on distingue 21 régions (regroupement de la Corse avec PACA), il y a 42 domaines : c'est le niveau auquel se situe l'objectif de similitude numérique entre effectifs. La modalité de l'activité spontanée dans le *RP* est désignée par *x* et celle qui sera imputée, donc le statut final d'activité au sens du BIT, est repérée par *y*. On note *h* la sous-population à laquelle appartient l'individu (10 valeurs possibles, cf. tableau 13). Les probabilités de transition initiales, notées désormais μ_{xh}^{y} pour quantifier la transition de *x* vers *y* dans la catégorie *h*, ne dépendent pas du domaine *d* (cf. tableau 14) parce qu'elles ne dépendent pas de la région. On note :

$P_{xh}^{y}(d)$ = probabilité de transition (finale) de *x* à *y* quand on appartient à *h*, dans le domaine *d* ;

$\widehat{N}_{xh}^{d}$ = effectif estimé au *RP* relatif à la population de statut *x* appartenant à *h* et à *d* ;

$\widehat{N}_{h}^{d}$ = effectif estimé au *RP* relatif à la population appartenant à *h* et à *d* ;

$\widehat{N}^{d}$ = effectif estimé au *RP* relatif à la population appartenant à *d*.

Si l'individu *i* a un poids noté w_i dans le fichier du recensement, alors

$$\widehat{N}_{xh}^{d} = \sum_{\substack{i \in RP \\ i \in d \cap h \cap x}} w_i .$$

Les taux-cibles issus de l'enquête *Emploi* (trois dans chaque domaine) sont notés θ_{y}^{d}. Il s'agit de la proportion d'individus de la population totale du champ (qui, on le rappelle, regroupe les personnes âgées de 15 ans ou plus et de moins de 75 ans) appartenant au domaine *d* et classés en statut *y*.

$$\theta_{y}^{d} = \sum_{h}\sum_{x} P_{xh}^{y}(d) \cdot \Pr oba(i \in x \,|\, i \in h, i \in d) \cdot \Pr oba(i \in h \,|\, i \in d)$$
$$= \sum_{h}\sum_{x} P_{xh}^{y}(d) \cdot \frac{\widehat{N}_{xh}^{d}}{\widehat{N}_{h}^{d}} \cdot \frac{\widehat{N}_{h}^{d}}{\widehat{N}^{d}}$$

Par définition $\theta_{y}^{d} \cdot \widehat{N}^{d}$, noté $\tilde{N}^{d,y}$, représente l'effectif d'individus du domaine *d* imputés en catégorie *y* à partir du recensement. Il traduit l'effectif-cible par statut, pour chaque domaine. Comme on distingue trois statuts, on obtient un jeu de 42 . 3 = 126 valeurs-cibles. Les contraintes à respecter, notées (★), sont :

$$\sum_{h}\sum_{x} P_{xh}^{y}(d) \cdot \widehat{N}_{xh}^{d} = \tilde{N}^{d,y} \qquad \forall d,\ \forall y$$

$$\sum_{y} P_{xh}^{y}(d) = 1 \qquad \forall x,\ \forall h,\ \forall d$$

Pour éviter une trop grande complexité mathématique, les contraintes $0 \le P_{xh}^{y}(d) \le 1$ sont gérées *ex post*. On cherche des probabilités de transition respectant ces contraintes et il est naturel qu'elles soient proches des probabilités de transition initiales μ_{xh}^{y}. On peut se placer dans un domaine *d* et minimiser la distance

$$\sum_{h}\sum_{x}\sum_{y}\left(P_{xh}^{y}(d) - \mu_{xh}^{y}\right)^2$$

sous les contraintes (★). Il existe une unique solution, valant $\forall d,\ \forall x,\ \forall h,\ \forall y$:

$$P_{xh}^{y}(d) = \mu_{xh}^{y} + \widehat{N}_{xh}^{d} \cdot \frac{\tilde{N}^{d,y} - \sum_{h}\sum_{x} \mu_{xh}^{y} \cdot \widehat{N}_{xh}^{d}}{\sum_{h}\sum_{x}\left(\widehat{N}_{xh}^{d}\right)^2}$$

Le second terme du membre de droite est un terme correcteur, qui peut être positif ou négatif. Il est nul lorsque $\tilde{N}^{d,y} = \sum_{h}\sum_{x} \mu_{xh}^{y} \cdot \widehat{N}_{xh}^{d}$, c'est-à-dire lorsque la grille des probabilités de transition initiales permet, en moyenne et sans aucune modification, d'estimer exactement les effectifs-cibles (situation idéale). Cette approche a l'inconvénient d'accorder la même importance, dans la gestion des écarts entre probabilités initiales et probabilités cibles, à toutes les probabilités quelles que soient leurs valeurs numériques. Cela est critiquable et peut être responsable du fait qu'une partie des solutions du programme se retrouve en dehors de l'intervalle 0 - 1. Il est possible aussi que ce soit l'origine des écarts relatifs plus grands que l'on constate dans la catégorie des chômeurs lorsqu'on compare *in fine* par domaine les effectifs estimés issus de l'imputation et les effectifs-cibles issus de l'enquête *Emploi*. Aussi, on peut s'intéresser à une alternative qui consiste à modifier la fonction objectif à minimiser afin que les écarts entre petites probabilités ne disparaissent pas dans la masse. Une façon de faire consiste à pondérer les écarts associés aux probabilités initiales μ_{xh}^{y} par une fonction décroissante de ces mêmes valeurs. Dans cet esprit, nous avons introduit une pondération en $1 \Big/ \left(\mu_{xh}^{y}\right)^{\alpha}$ dans la fonction objectif, où α est un réel au choix compris entre 0 et 1. Il est également possible d'ajouter une modulation de la fonction objectif en donnant plus d'importance aux catégories dont l'effectif est plus petit, dans l'optique essentielle de réduire les écarts d'estimation entre enquête *Emploi* et recensement pour ce qui concerne la catégorie des chômeurs – qui est la moins nombreuse. Pour cela, on va introduire un coefficient parfaitement contrôlé fonction du domaine noté $\varphi_y(d)$ qui quantifie l'importance donnée *a priori* à la catégorie *y*. Ce coefficient, qui dépend donc de la région, est compris entre 0 et 1 : plus il est petit, plus la catégorie associée prend de l'importance dans le critère à minimiser. Il reste alors à résoudre, sous les contraintes (★) :

Minimum de $\sum_{h}\sum_{x}\sum_{y} \frac{1}{\varphi_y(d)} \cdot \frac{1}{\left(\mu_{xh}^{y}\right)^{\alpha}} \left(P_{xh}^{y}(d) - \mu_{xh}^{y}\right)^2$.

Ce programme conduit à une expression complexe, programmée mais non reproduite ici.

mise en œuvre d'échantillonnages équilibrés dans chaque région / sexe, pour gérer chacun des trois statuts, ne se justifiait donc pas dans ces conditions.

Les résultats obtenus en métropole

En métropole, les effectifs-cibles sont issus de quatre enquêtes *Emploi* trimestrielles centrées sur janvier 2011, alors que le recensement fournissant les données individuelles à imputer est le *RP* 2010, lequel représente la situation de janvier 2010. Pour créer la cohérence temporelle nécessaire au bon déroulement du programme de calcul des probabilités de transition définitives, il a fallu modifier très légèrement la pondération des individus de l'échantillon de l'enquête *Emploi* afin que les effectifs globaux des individus du champ concerné coïncident exactement avec ceux du *RP* 2010 lorsqu'on croise la région et le sexe. L'annexe 2 donne l'ampleur de cette correction au niveau national (métropole) ainsi que l'écart aux effectifs-cibles (qui ne sont pas affectés par cette repondération) quand on applique les probabilités de transition initiales. Si l'on accepte que l'enquête *Emploi*, une fois repondérée, fournisse les effectifs de référence, on constate que l'imputation initiale dans le *RP* surestime les inactifs au détriment des actifs. Les écarts en jeu sont cependant faibles au niveau national (métropole).

Au-delà de cette opération initiale de mise en concordance, il a fallu effectuer un choix de paramètres, puisque nous avons vu que le programme d'optimisation sous contraintes intègre plusieurs paramètres techniques intervenant dans la définition de la fonction à minimiser (cf. encadré 3). À l'usage, de manière empirique et après avoir testé de nombreuses combinaisons parce qu'il n'y avait pas d'indices permettant d'orienter *a priori* le choix de ces paramètres, il est apparu que les résultats les plus satisfaisants apparaissent lorsqu'on choisit le paramètre α égal à 1,5 (cf. encadré 3) et le coefficient δ égal à 0 (cf. encadré 2). Les modulations par modalité permises au niveau régional par les coefficients $\varphi_y(d)$ n'apportent rien (plusieurs scénarios testés), ce qui est un peu surprenant, si bien que l'on a opté pour des $\varphi_y(d)$ uniformément égaux à 1. Afin d'estimer au travers d'un indicateur simple la qualité de toute l'opération, pour chaque sexe et chacune des trois modalités d'activité, on calcule les effectifs estimés par région, d'une part à partir de l'enquête *Emploi*, d'autre part à partir des imputations d'activité dans le recensement, puis on forme les erreurs relatives à partir de leurs différences[3]. L'indicateur de qualité associé à chaque modalité d'activité et à chaque sexe est égal à la somme des valeurs absolues de ces erreurs relatives sur les 21 régions distinguées en métropole. L'indicateur est fourni en pourcentage : par exemple lorsqu'il vaut 42, il faut comprendre qu'« en moyenne », dans une région donnée et pour le sexe considéré, l'effectif total estimé après imputation diffère de 42 / 21, soit 2 %, de l'effectif-cible. C'est cet indicateur de qualité qui a été utilisé pour rechercher le meilleur scénario en termes de paramètres techniques dans le programme d'optimisation.

Afin de disposer d'une norme, on se réfère aux valeurs de l'indicateur de qualité que l'on obtient lorsqu'on utilise les probabilités de transition initiales, c'est-à-dire avant la prise en compte des contraintes de calage sur les effectifs-cibles (cf. tableau 16). Pour un sexe donné, chaque ligne correspond à une opération d'imputation complète. Effectuer quelques opérations successives exactement dans les mêmes conditions (ce qui explique l'utilisation du terme de « simulation » (tableau 16)) permet de juger de la variabilité des erreurs résultant de l'aléa intrinsèque à l'imputation aléatoire et d'apprécier ainsi la stabilité de la méthode.

Une erreur relative de 170 % pour les chômeurs (par exemple) correspond à une erreur relative moyenne par sexe / région de l'ordre de 8 % : c'est considéré comme trop élevé et cela justifie l'adaptation des probabilités de transition initiales (étape 3). Le tableau 17 fournit les erreurs relatives obtenues en utilisant cette fois les probabilités de transition finales. Il est obtenu en s'en tenant aux paramètres définitifs de l'optimisation, en reproduisant l'imputation six fois, de manière indépendante et dans les mêmes conditions à chaque fois.

À l'évidence, l'opération de calage a considérablement augmenté la qualité de l'imputation pour chacun des trois statuts d'activité, justifiant cette lourde phase du processus. Considérant les simulations effectuées, la sortie numéro 5 apparaît globalement parmi les meilleures. On présente en annexe 3 ce que produit, en terme d'estimation des effectifs par sexe et par statut d'activité, cette même sortie au niveau national (métropole).

Sur le plan opérationnel, pour utiliser à notre avantage la variabilité de l'imputation, le programme

3. Si l'effectif-cible est A et que l'estimation imputée est B, l'erreur relative est définie par (B-A) / A.

informatique gérant toute l'opération a été lancé à quelques reprises, jusqu'à obtenir :

(i) une *erreur relative sur le nombre de chômeurs* par sexe inférieure ou égale à 10 % au niveau national, c'est-à-dire une erreur qui ne dépasse pas « en moyenne » 0,5 % par région x sexe (par rapport à des effectifs issus de l'enquête *Emploi* repondérée) ;

(ii) une *erreur relative sur le nombre d'actifs occupés* par sexe ne dépassant pas significativement 1,5 % au niveau national;

Tableau 16
Erreur relative de l'imputation obtenue avec les probabilités de transition initiales : différentes simulations

En %

Numéro de simulation	Sexe	Erreur relative		
		Actifs occupés	Chômeurs	Inactifs
1	Homme	44,9	161,6	87,4
2	Homme	45,4	164,5	87,4
3	Homme	45,7	164,3	87,0
4	Homme	44,8	161,0	87,3
5	Homme	45,8	164,6	87,2
1	Femme	45,3	179,1	68,2
2	Femme	46,8	179,0	68,4
3	Femme	46,5	180,3	68,7
4	Femme	45,4	179,7	68,2
5	Femme	46,4	179,9	68,7

Lecture : la simulation 3, pour les femmes, conduit à une erreur relative de l'imputation de 180,3 % pour l'estimation de l'effectif métropolitain de chômeurs dans le champ considéré, soit en moyenne 8,6 % par région.
Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Source : RP *2010 et enquête* Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

Tableau 17
Erreur relative de l'imputation obtenue avec les probabilités de transition finales : différentes simulations

En %

Numéro de simulation	Sexe	Erreur relative		
		Actifs occupés	Chômeurs	Inactifs
1	Homme	1,0	7,1	1,4
2	Homme	0,6	7,3	1,0
3	Homme	0,5	8,8	1,3
4	Homme	0,4	7,3	1,5
5	Homme	0,6	7,2	1,5
6	Homme	0,4	9,3	1,5
1	Femme	1,3	11,3	2,4
2	Femme	1,4	10,3	2,3
3	Femme	1,4	11,6	2,1
4	Femme	1,1	11,2	1,9
5	Femme	1,5	8,6	2,2
6	Femme	0,9	9,6	1,5

Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Source : RP *2010 et enquête* Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

(iii) une *erreur relative sur le nombre d'inactifs* par sexe ne dépassant pas significativement 2 % au niveau national;

(iv) des taux de chômage régionaux par sexe dont l'erreur **absolue** par rapport à l'enquête *Emploi* considérée comme la référence ne dépasse pas 0,1 point de pourcentage - cet écart maximum devant par ailleurs concerner essentiellement des croisements région / sexe où l'effectif issu de l'enquête Emploi reste modeste en taille.

Les indicateurs de qualité par sexe de l'imputation définitive au niveau national (métropole) sont donnés dans le tableau 18.

Il reste *in fine* à traiter les deux sous-populations (importantes en nombre) qui n'ont pas du tout été concernées par le processus qui vient d'être décrit, à savoir :

- toutes les personnes de 75 ans et plus, qui sont systématiquement imputées comme inactives au sens du BIT car les données de l'enquête *Emploi* montrent que moins de 0,5 % de cette population est active occupée et qu'elle ne comprend pas du tout de chômeurs ;

- toutes les personnes ne résidant pas en ménage ordinaire, pour lesquelles il n'existe hélas aucune source à laquelle se raccrocher pour apprécier le statut au sens du BIT et pour lesquelles nous nous sommes donc résolus à imputer brutalement la situation spontanée déclarée au recensement à titre d'activité BIT.

Une fois ces compléments apportés, la version définitive du fichier du *RP* 2010 imputé présente les caractéristiques fournies par le tableau 19, respectivement en version non pondérée et pondérée (cf. tableau 19). On précise que ce fichier

Tableau 18
Imputation définitive : erreurs relatives nationales par sexe, pour chaque statut d'activité (en %)

Sexe	Erreur relative		
	Actifs occupés	Chômeurs	Inactifs
Homme	0,8	9,2	1,3
Femme	1,6	8,3	2,2

Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Source : RP *2010 et enquête* Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

Tableau 19
Impact de l'imputation définitive sur les changements de statut dans le *RP* 2010

A - Version non pondérée

		Activité imputée BIT			
		Actifs occupés	Chômeurs	Inactifs	Total
Déclaration spontanée (*RP* 2010)	Actifs occupés	17 240 000	204 252	244 456	17 690 000
	Chômeurs	543 432	1 268 304	432 539	2 244 275
	Inactifs	714 578	279 512	15 800 000	16 800 000
	Total	18 500 000	1 752 068	16 480 000	36 730 000

B - Version pondérée

		Activité imputée BIT			
		Actifs occupés	Chômeurs	Inactifs	Total
Déclaration spontanée (*RP* 2010)	Actifs occupés	23 960 000	288 259	354 148	24 600 000
	Chômeurs	820 993	1 911 936	661 400	3 394 329
	Inactifs	1 060 783	418 998	21 720 000	23 200 000
	Total	25 840 000	2 619 193	22 730 000	51 190 000

Lecture : A : dans le fichier RP *2010 de métropole, 543 432 individus se déclarant spontanément chômeurs sont finalement imputés avec le statut d'actifs occupés au sens du BIT. B : d'après le fichier* RP *2010, on estime à 820 993 le nombre d'individus se déclarant spontanément chômeurs et qui sont finalement déclarés actifs occupés au sens du BIT.*
Champ : ensemble des individus résidant en métropole, ayant 15 ans ou plus au moment de la collecte EAR *(tous types de ménages).*
Source : RP *2010, non pondéré (A) pondéré (B).*

contient 8 484 000 valeurs manquantes, correspondant aux jeunes de moins de 15 ans au moment de la collecte.

En revenant au champ constitué par les personnes vivent en ménage ordinaire, ayant 15 ans ou plus mais moins de 75 ans en fin d'année de collecte (cette restriction sur l'âge ne modifie pas le dénombrement des chômeurs, et de manière très peu sensible celui des actifs occupés), le tableau 20 fournit les taux de chômage régionaux par sexe issus de l'imputation et les

Tableau 20
Taux de chômage par région / sexe, comparaison de la source *RP* imputé à la source enquête *Emploi*

Rang	Région	Sexe	Taux de chômage BIT selon le *RP* 2010 après imputation	Taux de chômage BIT selon l'enquête *Emploi*	Écart des taux après arrondi
1	Auvergne	Homme	7,38	7,48	**0,1**
2	Haute-Normandie	Homme	8,58	8,48	**0,1**
3	Picardie	Homme	9,66	9,58	**0,1**
4	Midi-Pyrénées	Femme	8,36	8,44	0
5	PACA et Corse	Femme	11,40	11,48	**0,1**
6	PACA et Corse	Homme	8,95	9,02	0
7	Champagne-Ardenne	Homme	9,69	9,75	**0,1**
8	Bourgogne	Femme	10,67	10,61	**0,1**
9	Poitou-Charentes	Femme	7,64	7,58	0
10	Champagne-Ardenne	Femme	10,20	10,26	**0,1**
11	Franche-Comté	Femme	9,71	9,77	**0,1**
12	Picardie	Femme	11,52	11,57	**0,1**
13	Lorraine	Homme	9,95	9,90	0
14	Limousin	Homme	8,12	8,17	**0,1**
15	Lorraine	Femme	10,28	10,32	0
16	Languedoc-Roussillon	Homme	13,40	13,36	0
17	Centre	Homme	7,33	7,29	0
18	Île-de-France	Femme	8,81	8,85	0
19	Nord-Pas-de-Calais	Femme	13,62	13,59	0
20	Auvergne	Femme	9,28	9,31	0
21	Bretagne	Femme	7,63	7,66	**0,1**
22	Aquitaine	Homme	7,75	7,78	0
23	Limousin	Femme	6,89	6,86	0
24	Pays de la Loire	Femme	9,83	9,80	0
25	Poitou-Charentes	Homme	8,97	9,00	0
26	Languedoc-Roussillon	Femme	12,67	12,64	**0,1**
27	Basse-Normandie	Homme	8,96	8,98	0
28	Pays de la Loire	Homme	8,60	8,58	0
29	Centre	Femme	8,54	8,53	0
30	Basse-Normandie	Femme	8,62	8,61	0
31	Haute-Normandie	Femme	10,14	10,13	0
32	Nord-Pas-de-Calais	Homme	12,03	12,04	0
33	Alsace	Homme	7,20	7,19	0
34	Aquitaine	Femme	9,84	9,83	0
35	Rhône-Alpes	Homme	6,86	6,85	0
36	Île-de-France	Homme	8,64	8,64	0
37	Bourgogne	Homme	9,45	9,45	0
38	Alsace	Femme	9,72	9,72	0
39	Franche-Comté	Homme	7,61	7,61	0
40	Bretagne	Homme	6,86	6,86	0
41	Midi-Pyrénées	Homme	7,88	7,88	0
42	Rhône-Alpes	Femme	9,00	9,00	0

Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Source : RP *2010 et enquête* Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

compare aux taux issus de l'enquête *Emploi*, dont on rappelle qu'ils constituent notre objectif initial privilégié. Il classe les croisements région / sexe par ordre décroissant d'erreur relative et montre que, pour tout croisement région / sexe, le taux de chômage BIT estimé à partir du *RP*, lorsqu'il est arrondi de façon à ne conserver que la première décimale, ne s'éloigne jamais de plus de 0,1 point de pourcentage du taux de référence issu de l'enquête *Emploi*. *In fine*, cet écart n'est d'ailleurs effectif que dans 12 croisements sur 42 (données en gras) ; pour les autres croisements, il y a égalité parfaite après arrondi ! On rappelle que les effectifs servant de référence sont calculés en mobilisant quatre fichiers trimestriels de l'enquête *Emploi* et après calage des poids pour assurer une cohérence avec la population du *RP* 2010 : le lecteur ne retrouvera donc dans aucun document diffusé ce qui est ici dénommé « taux de chômage BIT selon l'enquête *Emploi* ». Les croisements concernés par les plus gros écarts absolus (supérieurs à 0,06 points de pourcentage) ne concernent d'ailleurs que des populations modestes en taille.

On trouvera en annexe 4, par croisement région / sexe, les effectifs par statut d'activité BIT issus du *RP* imputé comparés à ceux estimés à partir de l'enquête *Emploi* (repondérée). Tous ces résultats sont (très) satisfaisants et donnent lieu à des erreurs relatives (très) faibles. En particulier, les erreurs relatives associées au nombre d'actifs occupés sont systématiquement inférieures ou égales (en valeur absolue) à 0,2 % si l'on excepte les femmes de Midi-Pyrénées (- 0,5 %). Même situation pour les inactifs, si l'on excepte là encore les femmes de Midi-Pyrénées (+ 0,9 %) et les hommes du Limousin (+ 0,3 %). Les écarts sont un peu plus importants pour les chômeurs, mais il n'y a que deux croisements où l'on dépasse 1 % en valeur absolue (femmes de Midi-Pyrénées et hommes du Limousin), ce qui reste excellent. Les écarts avec les effectifs-cibles sont probablement (un peu) plus importants, mais ils sont à mettre au compte du décalage temporel d'une année et ne sont en rien imputables à la méthode mise en œuvre.

Pour achever cette présentation, revenons à la question de la qualité. L'appréciation de cette dernière s'avère difficile à un niveau géographique infra régional. En effet, comme mentionné dans l'introduction, une telle appréciation nécessite de s'appuyer sur des données externes. De telles données n'existent pas sur des sous-populations plus restreintes que celles sur lesquelles ont porté les contraintes de compatibilité entre sources que l'on s'est données. En effet, le processus a été construit pour caler presque parfaitement les statistiques du recensement sur les statistiques d'activité agrégées issues de l'enquête *Emploi* au niveau régional. De ce point de vue, l'objectif semble atteint. Mais que dire des niveaux infra régionaux, là où l'enquête *Emploi* ne peut plus être une référence ? Même s'il n'a jamais été question d'en faire une cible, il existe des statistiques départementales officielles sur l'activité, mais elles sont elles-mêmes produites par modélisation. Un calage à ce niveau se ramènerait donc à un processus « modèle contre modèle ». Aussi nos capacités d'appréciation de l'imputation s'arrêtent là. En conclusion, en deçà du niveau régional et hélas sans que l'on soit en mesure de détecter le phénomène, il faudra compter sur l'absence d'effets locaux trop marqués : des effets qui concerneraient non la structure de l'activité elle-même, mais la relation modélisée qui lie la déclaration spontanée d'activité du recensement à la réalité de l'activité BIT.

L'annexe 5 précise la méthode appliquée et les résultats obtenus dans les départements d'outre mer.

* *
*

Les opérations d'imputation ont produit, pour les trois variables traitées, des résultats numériques conformes aux attentes formulées et qui permettent à la France de répondre dans les temps à la demande européenne. Pour les deux variables qui ont été modifiées, et en l'absence de priorité donnée à la statistique localisée, les méthodes dites de calage ont été perçues *in fine* comme plutôt décevantes en terme de rapport coût / avantage, ce qui a poussé finalement la décision vers les méthodes par modèle, plus simples, pour la production des données. Pourtant, l'idée d'une exploitation des *EAR* en nouvelle nomenclature ne pouvait pas ne pas être explorée parce qu'elle correspond mieux à l'esprit de la statistique publique qui tend à éviter la dépendance des estimations envers des modèles, c'est-à-dire envers des hypothèses dont on ne peut jamais vérifier la pertinence sans un recours à des sources externes. Dans le même ordre d'idée, une autre approche en théorie envisageable aurait été d'utiliser une technique de redressement des poids d'échantillon, comme le font traditionnellement les statisticiens d'enquête, et de se caler sur les structures

données par les *EAR* en nouvelle nomenclature, mais cela aurait conduit à une modification des pondérations du *Recensement* et ce n'est pas admissible en pratique. Pour sortir de ce dilemme, il faudrait produire pour chaque grande commune et chaque groupe de petites communes les imputations associées aux deux méthodes et choisir de manière pragmatique, au cas par cas, celle qui a la meilleure allure. Encore resterait-il à contourner le problème majeur des valeurs manquantes que l'on récupère en traitant, par la méthode dite de calage, la variable « période d'achèvement ».

Quant à l'imputation de l'activité au sens du BIT, il s'agit à l'évidence d'une « machinerie » complexe qui peut donner des idées redoutables : finalement, en adoptant une attitude résolument provocatrice et en poussant à bout la logique utilisée, le recensement étant réalisé à rythme annuel avec une collecte d'activité spontanée qui semble pérenne, si tant est que le modèle soit suffisamment stable dans le temps, la mise en œuvre du processus d'imputation chaque année permettrait presque de se passer, au moins de temps à autre, de l'enquête *Emploi* pour produire des taux de chômage annuels au sens du BIT… □

BIBLIOGRAPHIE

Baccaïni B. (2001), « Les migrations internes en France de 1990 à 1999 : l'appel de l'ouest », *Économie et statistique*, n° 344, pp. 39-79.

Baccaïni B. (2007), « La mobilité géographique d'un recensement à l'autre », dans « L'État de la France » 2007-2008, *La Découverte*, pp. 318-324, La Documentation française.

Baccaïni B. (2007), « Les flux migratoires interrégionaux en France depuis 50 ans » , *Population*, vol. 62, n° 1, pp. 143-160, Ined.

Baccaïni B. et Levy D. (2009), « Recensement de la population de 2006. Les migrations entre départements : le sud et l'ouest toujours très attractifs », *Insee Première*, n° 1248.

BIT (2013), « Résolution concernant les statistiques du travail, de l'emploi et de la sous-utilisation de la main-d'œuvre », 19e conférence internationale des statisticiens du travail.

Blanchet D. et Marchand O. (2003), « Mesurer l'emploi et le chômage : nouvelle enquête, débats anciens », *Économie et statistique*, n° 362.

Briant P., Donzeau N., Marpsat M., Pirus C. et Rougerie C. (2010), « Le dispositif statistique de l'Insee dans le domaine du logement : état des lieux et évaluation comparée des sources », *Document de travail Insee-DSDS*, n° F1002.

Chardon O. et Goux D. (2003), « La nouvelle définition européenne du chômage BIT », *Économie et statistique*, n° 362.

Cnis (2014), Commission emploi, qualifications et revenus du travail, compte rendu de la séance du 10 avril 2014, n° 69/H030, pp. 13-18.

Commission européenne, (2000), Règlement (CE) N° 1897/2000 de la Commission européenne du 7 septembre 2000 portant application du règlement (CE) N° 577/98 du Conseil relatif à l'organisation d'une enquête par sondage sur les forces de travail dans la Communauté en ce qui concerne la définition opérationnelle du chômage.

Concialdi P. (2014), « Quand les statisticiens du travail définissent le travail », *Chronique internationale de l'Ires*, n° 145.

Daudin V. et Rivière J (2014), « Enquête *Emploi* 2013 à La Réunion », *Insee Informations rapides Réunion*, n° 300, avril 2014.

Foucauld J.-B., Reynaud M. et Cézard M. (2008), « Emploi, chômage et précarité », Rapport d'un groupe de travail Cnis, n° 108, Insee.

Givord P. (2003), « Une nouvelle enquête Emploi », *Économie et statistique*, n° 362, pp 59-66.

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

Gonzalez-Demichel C. et Nauze-Fichet E. (2003), « Les contours de la population active : aux frontières de l'emploi, du chômage et de l'inactivité », *Économie et statistique*, n° 362.

Guggemos F. et Vidalenc J. (2014), « Une photographie du marché du travail en 2013 », *Insee Première*, n° 1516, septembre.

Guillemot D. (1996), « La population active : une catégorie difficile à cerner », *Économie et statistique*, n° 300.

Inspection générale des finances, Inspection générale des affaires sociales (2007), « Rapport sur les statistiques d'estimation du chômage ».

Insee (2014), communiqués de presse, septembre 2013 et mars 2014 : http://www.insee.fr/fr/indicateurs/ind14/20140306/communique_presse_chomage_060314.pdf.

Insee Résultats (2013), Séries longues sur le marché du travail, *Société*, n° 149.

Journal officiel de l'Union européenne (2008), « Règlement (CE) N° 763/2008 du parlement européen et du Conseil du 9 juillet 2008 concernant les recensements de la population et du logement ».

Toulemon L. (2012), « Évolution des situations familiales à travers les recensements français de 1962 à 2009 », *Population*, vol. 67, n° 4, octobre-décembre, pp. 657-681.

Treyens P.-E. (2014), « Enquête *Emploi* 2013 en Guadeloupe », *Insee Premiers Résultats*, n° 103, avril 2014.

Treyens P.-E. (2014), « Enquête *Emploi* 2013 en Guyane », *Insee Premiers Résultats*, n° 104, avril 2014.

Treyens P.-E. (2014), « Enquête *Emploi* 2013 en Martinique », *Insee Premiers Résultats*, n° 105, avril 2014.

ANNEXE 1

SECONDE MÉTHODE D'IMPUTATION POUR LA VARIABLE « PÉRIODE D'ACHÈVEMENT DE LA CONSTRUCTION »

Comme pour la variable de lieu de résidence antérieure, cette seconde approche permet en théorie de limiter les à-coups qui surviennent dans les séries chronologiques lorsqu'une *EAR* en ancienne nomenclature disparaît du recensement et se trouve remplacée par une *EAR* en nouvelle nomenclature. L'idée centrale consiste à considérer une tranche de l'ancienne nomenclature et à déterminer un ensemble de probabilités de passage vers les différentes tranches de la nouvelle nomenclature de telle façon qu'en moyenne la structure imputée soit identique à la structure des *EAR* collectant les données en nouvelle nomenclature. Techniquement, il s'agit donc de trouver des probabilités conditionnelles – on parlera désormais de probabilités de transition – *a priori* plus pertinentes que ne le sont les lois uniformes caractérisant la première méthode (c'est-à-dire tenant compte des données collectées d'après la nouvelle variable, qui plus est à un niveau géographique pouvant aller selon les cas jusqu'à la commune, autrement dit tenant compte du terrain). Certaines transitions sont évidentes, grâce aux inclusions de tranches anciennes dans des tranches nouvelles : par exemple, un logement recensé dans la tranche ancienne « De 1982 à 1989 » sera imputé dans la tranche nouvelle « De 1971 à 1990 » avec une probabilité égale à 1. Comme nous l'avons déjà constaté avec la première méthode, les difficultés essentielles surgissent lors de l'affectation des logements des tranches « Avant 1949 » et « De 1949 à 1974 », en rajoutant le cas de la tranche « De 1990 à 1998 » où l'on peut craindre quelques perturbations puisqu'une seule année (1990) est responsable d'une situation de non-inclusion. À partir des données du *RP* 2010, le calcul des probabilités de transition s'appuie sur un ensemble de contraintes et ne pose aucune difficulté particulière (cf. encadré) ; en revanche, le contexte

Encadré

DÉTERMINATION DES PROBABILITÉS DE TRANSITION ENTRE TRANCHES

En se limitant à la construction antérieure à 2006, l'échantillon recensé selon la nouvelle nomenclature est noté s. On note w_i le poids du logement i dans le recensement considéré. On fait désormais référence aux codes donnés dans le tableau 8. A partir de s, pour chaque tranche k (k varie de 1 à 5) exprimée dans la nouvelle nomenclature, on estime la proportion $\widehat{P}_k = \sum_{i\in s, i\in k} w_i \Big/ \sum_{i\in s} w_i$ des logements achevés durant la période k. On a $\sum_{k=1}^{5} \widehat{P}_k = 1$. Toujours en se limitant à la construction antérieure à 2006, l'échantillon des logements à imputer, noté s', est structuré en ancienne nomenclature (tranches codées de A à F) selon $(\widehat{P}_A, \widehat{P}_B, ..., \widehat{P}_F)$ où $\sum_{I=A,...,F} \widehat{P}_I = 1$. On a par exemple $\widehat{P}_A = \sum_{i\in s', i\in A} w_i \Big/ \sum_{i\in s'} w_i$.

Notons X l'année d'achèvement réelle, dont on enregistre l'appartenance à une tranche selon l'ancienne nomenclature. Certaines transitions vers la nouvelle nomenclature sont possibles - voire même imposées - d'autres ne le sont pas. On décrit le système d'imputation par les probabilités conditionnelle suivantes :

$1 = P(X \in 5 \mid X \in F)$ $\qquad$ $\theta_{B3} = P(X \in 3 \mid X \in B)$

$\theta_{B4} = P(X \in 4 \mid X \in B)$

$\theta_{E4} = P(X \in 4 \mid X \in E)$

$\theta_{E5} = P(X \in 5 \mid X \in E)$ $\qquad$ $\theta_{A3} = P(X \in 3 \mid X \in A)$

$\theta_{A2} = P(X \in 2 \mid X \in A)$

$1 = P(X \in 4 \mid X \in D)$ $\qquad$ $\theta_{A1} = P(X \in 1 \mid X \in A)$

$1 = P(X \in 4 \mid X \in C)$

On a en sus :

$$\theta_{E4} + \theta_{E5} = 1 \qquad \theta_{B3} + \theta_{B4} = 1 \qquad \theta_{A1} + \theta_{A2} + \theta_{A3} = 1$$

On cherche à assurer une structure de s' identique à celle de s (en moyenne). Dans cette perspective, on calcule l'espérance mathématique des proportions par classe après imputation (nouvelle nomenclature), et on l'égale à la proportion constatée. Le système suivant relie toutes ces grandeurs :

$$\widehat{P}_5 = \widehat{P}_F + \theta_{E5} \cdot \widehat{P}_E$$

$$\widehat{P}_4 = \theta_{E4} \cdot \widehat{P}_E + \widehat{P}_D + \widehat{P}_C + \theta_{B4} \cdot \widehat{P}_B$$

$$\widehat{P}_3 = \theta_{B3} \cdot \widehat{P}_B + \theta_{A3} \cdot \widehat{P}_A$$

$$\widehat{P}_2 = \theta_{A2} \cdot \widehat{P}_A$$

$$\widehat{P}_1 = \theta_{A1} \cdot \widehat{P}_A$$

On obtient donc 8 équations à 7 inconnues, mais 7 équations sont indépendantes. La solution est

$$\theta_{A1} = \frac{\widehat{P}_1}{\widehat{P}_A} \qquad \theta_{A2} = \frac{\widehat{P}_2}{\widehat{P}_A} \qquad \theta_{A3} = \frac{\widehat{P}_A - (\widehat{P}_1 + \widehat{P}_2)}{\widehat{P}_A}$$

$$\theta_{B3} = \frac{\widehat{P}_1 + \widehat{P}_2 + \widehat{P}_3 - \widehat{P}_A}{\widehat{P}_B}$$

$$\theta_{B4} = \frac{(\widehat{P}_A + \widehat{P}_B) - (\widehat{P}_1 + \widehat{P}_2 + \widehat{P}_3)}{\widehat{P}_B}$$

$$\theta_{E4} = \frac{\widehat{P}_1 + \widehat{P}_2 + \widehat{P}_3 + \widehat{P}_4 - (\widehat{P}_A + \widehat{P}_B + \widehat{P}_C + \widehat{P}_D)}{\widehat{P}_E}$$

$$\theta_{E5} = \frac{\widehat{P}_5 - \widehat{P}_F}{\widehat{P}_E}$$

peut conduire à des valeurs numériques sortant de l'intervalle [0,1], ce qui n'est pas admissible. Ce problème survient lorsqu'une borne de la nouvelle nomenclature se rapproche trop d'une borne de l'ancienne : en scénario extrême, si les deux nomenclatures s'articulaient autour d'une borne identique (ce qui n'est pas le cas…), la réconciliation des structures nécessiterait que la proportion estimée de logements construits antérieurement à cette borne dans les *EAR* utilisant l'ancienne nomenclature pour la collecte soit identique à la proportion estimée de logements construits antérieurement à cette même borne dans les *EAR* collectant en nouvelle nomenclature – ce qui ne peut pas se produire en réalité puisqu'il s'agit de deux échantillons différents. De fait, les risques sont majoritairement dus au positionnement des années 1989 / 1990, qui rapprochent considérablement les deux nomenclatures et rendent donc particulièrement périlleuse la réconciliation des structures.

L'imputation proprement dite, comme dans la première méthode, s'effectue à partir d'un nombre aléatoire généré au niveau de l'immeuble et tiré au hasard entre 0 et 1. À l'expérience, il est apparu impossible d'éviter que certaines probabilités de transition ne sortent de l'intervalle [0,1]. Lorsque cela survient, on peut s'en remettre en ultime recours à la première approche, qui pour sa part fonctionne en toutes circonstances. Il est également possible de ramener les probabilités faiblement négatives à la valeur zéro. Ainsi, nous avons proposé le scénario qui consiste à adapter toute probabilité négative supérieure à - 0,05 lorsque le logement est construit avant 1949 et toute probabilité négative supérieure à - 0,15 lorsque le logement est construit entre 1990 et 1998, considérant que le risque est sensiblement plus fort dans cette dernière tranche à cause de l'année 1990 (la probabilité d'être achevé avant 1990 – tranches A, B, C, D réunies – sera très proche de la probabilité d'être achevé avant 1991 – tranches 1, 2, 3, 4 réunies). Dans les deux cas, la probabilité déficiente est forcée à la valeur 0 et l'on étalonne bien entendu les autres probabilités de transition pour que la somme reste égale à 1. Lorsque les valeurs des probabilités négatives sont trop importantes (ce qui exclut de les forcer à zéro), la période d'achèvement reste à valeur manquante. La fréquence de ces situations laisse penser qu'il y a trop souvent, au niveau local, une forme d'incompatibilité entre les échantillons recensés en ancienne nomenclature et les échantillons recensés en nouvelle nomenclature ; mais peut-être est-ce dû tout simplement à l'erreur d'échantillonnage et peut-être aussi a-t-on été trop exigeant… La variance d'imputation n'a pas été contrôlée, l'argumentaire étant exactement celui qui a été donné pour la variable de résidence antérieure.

Partant du *RP* 2010, on procède ainsi sur chaque grande commune, de manière autonome en considérant qu'on dispose à chaque fois d'une structure-cible de qualité acceptable, produite à partir de deux *EAR*. En revanche, on ne peut évidemment pas pratiquer de même avec les petites communes et il est nécessaire de les regrouper au préalable. La programmation standard a distingué des groupes croisant la région ou un groupe de régions avec une typologie des communes construite selon la première méthode, donc sur la base des données du *RP* 2008. Concernant le nombre de groupes de petites communes à distinguer, qui reste paramétré, il semble que l'exhaustivité du recensement dans les petites communes donne une forte légitimité au regroupement mais il y a un équilibre à trouver à ce niveau car trop de groupes rend le gain de précision illusoire (il y a toujours *in fine* une phase d'imputation aléatoire…) et augmente le temps de traitement informatique (qui devient rapidement excessif), tandis que trop peu de groupes risque naturellement d'introduire davantage de biais. Le scénario standard est celui qui distingue dix groupes de petites communes dans la région traitée. La structure-cible est alors estimée groupe par groupe.

Cette méthode a été appliquée à la région Auvergne toutes communes confondues (cf. tableau A) et à l'ensemble des grandes communes de cette même région (cf. tableau B). Il est intéressant de comparer les résultats avec ceux que donne la première méthode : très léger avantage pour la seconde méthode lorsqu'il s'agit de la région dans son ensemble, avantage un peu plus marqué si on s'en tient aux grandes communes, mais sa valeur ajoutée reste discutable. Cela étant, il faut aussi prendre en compte l'effet négatif de la méthode calée sur la production de l'information individuelle : si les estimations globales de proportions ne semblent pas en souffrir, la méthode calée génère de nombreuses valeurs non renseignées au niveau du logement du fait des probabilités de transition « trop négatives », ce qui est évidemment catastrophique pour les estimations d'effectifs. En région Auvergne, sur un effectif global total (pondéré) de 760 000 logements, 140 000 ne sont *in fine* pas imputés (contre seulement 2 200 avec la première méthode) ! Si on s'en tient aux grandes communes, sur 235 000 logements estimés, 62 500 logements ne sont pas imputés. Pour la production standardisée des données mises à disposition d'Eurostat, cette méthode alternative n'est donc pas applicable. Elle peut par contre être exploitée dans d'autres circonstances si on souhaite produite des estimations de structures.

Dans le cas de l'Auvergne, sur les dix groupes de petites communes qui ont été distingués, la structure de référence est construite à partir d'un échantillon comprenant, selon le groupe considéré, entre 2 800 et 50 000 logements, la plupart des groupes s'appuyant sur un échantillon d'environ 20 000 logements.

Tableau A
Comparaison des méthodes, région Auvergne, toutes communes

En %

EAR	Méthode avec source externe						Méthode par calage					
	Avant 1919	1919-1945	1946-1970	1971-1990	1991-2005	Total	Avant 1919	1919-1945	1946-1970	1971-1990	1991-2005	Total
2008	30,2	14,4	20,5	23,4	11,5	100	29,2	15,1	19,2	24,4	12,1	100
2009	28,7	13,9	21,6	24,0	11,8	100	28,2	14,3	19,4	25,8	12,3	100
2010	28,6	14,1	22,1	23,9	11,3	100	28,0	13,9	20,0	26,4	11,7	100
2011	27,5	12,8	21,1	26,0	12,6	100	27,5	12,8	21,1	26,0	12,6	100
2012	27,6	14,1	19,9	26,6	11,8	100	27,6	14,1	19,9	26,6	11,8	100
Total	28,5	13,8	21,1	24,8	11,9	100	28,0	14,0	20,0	25,9	12,1	100

Lecture : dans le sous-échantillon EAR *2010, avec la pondération du* RP *2010, après imputation par la méthode de calage, 28,0 % des logements sont déclarés construits avant 1919.*
Champ : région Auvergne, ensemble des logements recensés.
Sources : RP *2010 et enquête* Logement *2006.*

Tableau B
Comparaison des méthodes, région Auvergne, ensemble des grandes communes

En %

EAR	Méthode avec source externe						Méthode par calage					
	Avant 1919	1919-1945	1946-1970	1971-1990	1991-2005	Total	Avant 1919	1919-1945	1946-1970	1971-1990	1991-2005	Total
2008	17,4	12,9	32,5	26,0	11,2	100	12,0	16,0	31,2	27,8	13,0	100
2009	16,4	12,1	33,8	26,6	11,1	100	11,8	15,5	32,3	28,2	12,2	100
2010	16,6	13,6	35,6	25,9	8,3	100	12,3	15,7	35,0	28,6	8,5	100
2011	13,3	13,1	33,3	28,0	12,3	100	13,3	13,1	33,3	28,0	12,3	100
2012	12,8	15,4	31,3	29,4	11,1	100	12,8	15,4	31,3	29,4	11,1	100
Total	15,3	13,4	33,1	27,1	11,1	100	12,6	14,9	32,5	28,5	11,5	100

Champ : région Auvergne, ensemble des logements recensés dans les grandes communes (> 10 000 habitants).
Sources : RP *2010 et enquête* Logement *2006.*

Tableau
Ampleur de la repondération dans l'enquête *Emploi* pour mise en cohérence avec les effectifs *RP* 2010 et qualité de l'imputation avec les probabilités de transition initiales

	Actifs occupés	Chômeurs	Inactifs	Estimation de la taille de la population totale
A : Effectif extrapolé enquête *Emploi*, avant calage sur la population recensée par région X sexe (= effectif-cible)	25 720 000	2 605 000	16 818 000	45 143 000
B : Effectif extrapolé enquête *Emploi* obtenu après calage sur la population recensée par région X sexe	25 635 183	2 596 752	16 760 762	44 992 707
C : Effectif imputé dans le *RP* 2010 en utilisant les probabilités de transition initiales	25 597 177	2 575 761	16 819 769	44 992 707
Erreur relative due à l'imputation par rapport aux effectifs-cibles (en %) = (C - A) / A	- 0,48	- 1,12	ε	- 0,33
Erreur relative due à l'imputation par rapport aux effectifs *Emploi* corrigés (en %) = (C - B) / B	- 0,15	- 0,81	+ 0,35	0

Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Sources : RP *2010 et enquête* Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

ANNEXE 3

Tableau A
Estimation d'effectifs par statut d'activité : résultats de l'imputation (cas de la simulation n° 5)

Sexe	Actifs occupés BIT		Chômeurs BIT		Inactifs BIT	
	RP	Enquête *Emploi*	*RP*	Enquête *Emploi*	*RP*	Enquête *Emploi*
Homme	13 473 050	13 474 152	1 283 326	1 283 239	7 247 521	7 246 505
Femme	12 157 585	12 161 028	1 312 608	1 313 516	9 518 619	9 514 262
Total	25 630 635	25 635 180	2 595 934	2 596 755	16 766 140	16 760 767

Lecture : l'imputation considérée ici conduit à une estimation de 12 157 585 femmes actives occupées dans le champ considéré alors que l'enquête Emploi en estime l'effectif à 12 161 028.
Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Source : RP *2010 et enquête* Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

Tableau B
Erreur d'estimation des effectifs par statut d'activité : résultats de l'imputation (cas de la simulation n° 5)

Sexe	Erreur actifs occupés	Erreur chômeurs	Erreur inactifs	Population totale *RP*	Population totale enquête *Emploi*
Homme	ε	ε	ε	22 003 897	22 003 896
Femme	ε	- 0,1 %	ε	22 988 812	22 988 806
Total	ε	ε	ε	44 992 709	44 992 702

Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire
Source : RP *2010 et enquête* Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

ANNEXE 4

EFFECTIFS ESTIMES PAR RÉGION ET SEXE SELON LE STATUT D'ACTIVITÉ BIT COMPARAISON IMPUTATION *RP* 2010 / ENQUÊTE *EMPLOI*

Région	Sexe	Actifs occupés *RP*	Actifs occupés Enquête *Emploi*	Chômeurs *RP*	Chômeurs Enquête *Emploi*	Inactifs *RP*	Inactifs Enquête *Emploi*
Île-de-France	Homme	2 669 713	2 670 337	252 636	252 399	1 239 328	1 238 943
Île-de-France	Femme	2 492 586	2 492 008	240 860	242 034	1 729 181	1 728 585
Champagne-Ardenne	Homme	275 505	275 331	29 555	29 739	166 877	166 868
Champagne-Ardenne	Femme	255 456	255 464	29 026	29 220	200 499	200 296
Picardie	Homme	415 224	415 670	44 410	44 032	218 061	217 993
Picardie	Femme	364007	363 589	47 382	47 588	282 114	282 326
Haute Normandie	Homme	386479	386 868	36 289	35 834	220 997	221 063
Haute Normandie	Femme	339513	339 781	38 310	38 289	296 308	296 060
Centre	Homme	540818	540 831	42 786	42 537	30 2999	303 234
Centre	Femme	501 265	501 359	46 830	46 749	367 270	367 258
Basse Normandie	Homme	298 252	298 190	29 345	29 427	185 797	185 777
Basse Normandie	Femme	275 535	275 724	26 003	25 992	225 459	225 282
Bourgogne	Homme	338 965	338 982	35 372	35 382	199 662	199 635
Bourgogne	Femme	310 150	310 377	37 038	36 831	242 168	242 146
Nord-Pas-de-Calais	Homme	813 605	813 430	111 286	111 395	489 795	489 861
Nord-Pas-de-Calais	Femme	689 607	690 190	108 772	108 538	683 751	683 402
Lorraine	Homme	523 858	523 959	57 900	57 544	257 248	257 503
Lorraine	Femme	445 879	446 121	51 101	51 316	366 318	365 860
Alsace	Homme	441 769	442 102	34 282	34 240	192 583	192 291
Alsace	Femme	369 287	369 103	39 750	39 746	275 402	275 589
Franche-Comté	Homme	258 730	258 786	21 318	21 328	136 847	136 782
Franche-Comté	Femme	227 442	227 086	24 456	24 589	168 709	168 931
Pays de la Loire	Homme	776 310	776 791	73 064	72 859	399 401	399 126
Pays de la Loire	Femme	685 941	685 806	74 818	74 485	517 035	517 501
Bretagne	Homme	661 960	662 038	48 749	48 763	409 187	409 094
Bretagne	Femme	620 303	620 493	51 274	51 470	469 980	469 593
Poitou-Charentes	Homme	356 095	356 069	35 099	35 203	223 960	223 882
Poitou-Charentes	Femme	344 646	345 090	28 523	28 312	265 243	265 009
Aquitaine	Homme	687 449	687 855	57 740	58 015	382 354	381 674
Aquitaine	Femme	629 095	629 254	68 628	68 618	489 653	489 503
Midi-Pyrénées	Homme	637 674	637 659	54 569	54 524	324 779	324 838
Midi-Pyrénées	Femme	595 172	598 049	54 317	55 105	394 987	391 322
Limousin	Homme	154 160	154 371	13 633	13 726	88 491	88 187
Limousin	Femme	144 106	144 185	10 664	10 627	110 040	109 996
Rhône-Alpes	Homme	1 406 440	1 406 074	103 535	103 468	687 543	687 977
Rhône-Alpes	Femme	1 232 454	1 232 089	121 903	121 894	918 087	918 463
Auvergne	Homme	293 844	293 511	23 399	23 724	158 929	158 938
Auvergne	Femme	257 278	257 403	26 328	26 439	201 122	200 887
Languedoc-Roussillon	Homme	486 845	486 857	75 348	75 062	346 861	347 135
Languedoc-Roussillon	Femme	446 829	446 754	64 821	64 618	461 175	461 452
PACA et Corse	Homme	1 063 597	1 063 208	104 535	105 427	624 301	623 798
PACA et Corse	Femme	945 342	944 809	121 613	122 556	866 174	865 764
Total		25 659 185	25 663 653	2 597 267	2 599 644	16 786 675	16 779 824

Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant en métropole et en ménage ordinaire.
Source : RP *2010 et enquête* Emploi *(3e et 4e trimestres de 2010 , 1er et 2e trimestres de 2011).*

ANNEXE 5

IMPUTATION DE L'ACTIVITÉ AU SENS DU BIT : LE CAS DES DÉPARTEMENTS D'OUTRE-MER

Dans les quatre DOM « historiques », la démarche a été similaire à celle appliquée en métropole. En premier lieu, les estimations d'effectifs par statut issues de l'*EAR* 2011 peuvent être comparées à celles issues de l'enquête *Emploi* annuelle 2011, en distinguant dans cette dernière la déclaration spontanée du concept BIT (cf. tableau A). On vérifie que l'incohérence entre les deux sources concernant la déclaration spontanée du statut d'activité est sensiblement plus importante dans les DOM qu'en métropole (voir aussi le tableau 11 du texte).

Tableau A
Par statut d'activité, dans les DOM, effectifs estimés selon le concept et selon la source

	Déclaration spontanée d'activité		Activité au sens du BIT
	Selon l'*EAR* 2011	Selon l'enquête *Emploi* de 2011	Selon l'enquête *Emploi* de 2011
Actifs occupés	548 000	541 000	557 000
Chômeurs	296 000	251 000	188 000
Inactifs	483 000	525 000	572 000
Total	1 327 000	1 317 000	1 317 000

*Lecture : selon l'*EAR *2011, on estime à 296 000 le nombre total de chômeurs lorsqu'on exploite la déclaration spontanée d'activité (poids de l'*EAR*). Cet effectif est réduit à 251 000 si l'on utilise l'enquête* Emploi*. Si l'on comptabilise des chômeurs au sens du BIT, l'enquête* Emploi *en estime l'effectif global à 188 000.*
Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant dans les DOM et en ménage ordinaire.
Sources : EAR *2011 et enquête* Emploi *de 2011.*

Le contexte des DOM a été plus favorable qu'en métropole à la construction de la variable explicative alternative à la déclaration spontanée qui a été succinctement décrite dans la partie consacrée à la métropole, lui donnant même un avantage sur la déclaration spontanée en terme de pouvoir explicatif (on peut penser que ce n'est pas sans rapport avec les résultats mitigés du tableau A). Néanmoins, l'avantage est modéré et il a été décidé de conserver la déclaration spontanée comme facteur explicatif majeur, l'homogénéité de traitement avec la métropole étant apparue comme un argument fort. Les trois variables sociodémographiques que sont le sexe, le diplôme (3 modalités) et la tranche d'âge (4 modalités) ont constitué les facteurs explicatifs venant en complément de la déclaration spontanée d'activité pour expliquer l'activité au sens du BIT.

Comme en métropole, il a fallu créer une variable intermédiaire d'activité, répondant aux contraintes suivantes (cf. encadré 2) :

$\theta = 0{,}765 \cdot \delta - 0{,}052$ et $\mu = 1{,}698 \cdot \delta - 0{,}150$ pour les hommes

$\theta = 0{,}794 \cdot \delta + 0{,}036$ et $\mu = 1{,}592 \cdot \delta - 0{,}018$ pour les femmes

Il apparaît, comme le laissaient supposer les résultats du tableau A, une différence de comportement significative avec la métropole, qui aura pour conséquence ultérieure d'imposer le choix d'un coefficient *delta* plus important qu'en métropole (les probabilités θ et μ devant prendre des valeurs positives). Ensuite, des probabilités de transition ont été estimées en scindant la population en dix sous-populations, selon le partitionnement détaillé dans le tableau B.

Tableau B
Catégories de population utilisées dans les DOM pour l'estimation des probabilités de transition

Catégorie	Sexe	Tranche d'âge	Classe de diplôme
1	Homme	15 à 49 ans	Diplôme <= brevet
2	Homme	15 à 39 ans	Brevet < diplôme
3	Homme	40 à 49 ans	Brevet < diplôme
4	Homme	50 à 59 ans	Tous diplômes
5	Homme	60 à 74 ans	Tous diplômes
6	Femme	15 à 39 ans	Diplôme <= brevet
7	Femme	15 à 39 ans	Brevet < diplôme
8	Femme	40 à 49 ans	Tous diplômes
9	Femme	50 à 59 ans	Tous diplômes
10	Femme	60 à 74 ans	Tous diplômes

Le partitionnement final apparaît légèrement différent de celui utilisé en métropole (cf. tableau 13 du texte). La source « enquête *Emploi* » représentant la situation au 1er janvier 2011 a été constituée par l'agrégation de trois enquêtes annuelles successives (2010, 2011 et 2012), le dispositif d'enquête *Emploi* dans les DOM étant distinct de celui de la métropole jusqu'à fin 2012. Le tableau C détaille les probabilités de transition initiales pour chaque sous-population distinguée.

Tableau C
Probabilités de transition initiales estimées par catégorie de population, dans les DOM

En %

Activité spontanée intermédiaire	Activité BIT	Catégorie de population									
		1	2	3	4	5	6	7	8	9	10
Actif occupé	Actif occupé	76,9	92,3	97,1	93. 8	65,3	45,5	85,9	88,2	87,7	51,1
Actif occupé	Chômeur	6,6	3,8	1,7	1,8	0,4	13,6	6,2	5,5	3,6	0,6
Actif occupé	Inactif	16,5	3,9	1,2	4,4	34,3	40,9	7,9	6,3	8,7	48,3
Chômeur	Actif occupé	11,1	19,2	31,0	22,9	9,9	7,5	23,9	26,1	29,1	10,8
Chômeur	Chômeur	53,1	61,4	50,8	42,4	7,8	43,0	54,1	49,2	35,4	3,4
Chômeur	Inactif	35,8	19,4	18,2	34,7	82,3	49,5	22,0	24,7	35,5	85,8
Inactif	Actif occupé	5,5	20,8	56,8	18,5	2,8	2,7	20,3	23,2	17,4	2,3
Inactif	Chômeur	6,0	12,6	13,9	6,8	0,3	6,1	14,5	14,3	6,9	0,3
Inactif	Inactif	88,5	66,6	29,3	74,7	96,9	91,2	65,2	62,5	75,7	97,4

Lecture : considérant un individu de la catégorie 1 (un homme de 15 à 49 ans, sans diplôme ou dont le niveau de diplôme ne dépasse pas celui du brevet) qui se déclare spontanément chômeur au sens de la variable intermédiaire, il y a 11,1 chances sur 100 pour qu'il soit actif occupé au sens du BIT.
Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant dans les DOM et en ménage ordinaire.
Source : enquêtes Emploi *2010, 2011, 2012.*

On rappelle que chaque DOM constitue une région administrative. Pour construire les probabilités de transition définitives à partir du jeu de probabilités initiales, après une phase de tâtonnement, le paramétrage finalement retenu diffère de celui de la métropole puisque, reprenant les notations introduites dans les encadrés 3 et 4, le paramètre δ a été fixé à 0,2 pour chaque sexe, et le paramètre α a été fixé pour sa part à 0,5. Le programme d'imputation a été lancé plusieurs fois jusqu'à obtenir :

- une *erreur relative sur le nombre de chômeurs* par sexe inférieure ou égale à 2 % sur l'ensemble des quatre DOM, c'est-à-dire une erreur qui ne dépasse pas « en moyenne » 0,5 % par DOM / sexe ;

- une *erreur relative sur le nombre d'actifs occupés* par sexe ne dépassant pas significativement 1 % sur l'ensemble des quatre DOM;

- une *erreur relative sur le nombre d'inactifs* par sexe ne dépassant pas significativement 1 % sur l'ensemble des quatre DOM;

- des taux de chômage par DOM / sexe dont l'erreur absolue par rapport à la référence qu'est l'enquête *Emploi* ne dépasse pas 0,3 point de pourcentage.

Ces objectifs conduisent à une qualité comparable à celle de la métropole, les taux de chômage étant sensiblement plus élevés dans les DOM qu'en métropole. L'imputation retenue conduit, sur l'ensemble des DOM, aux performances résumées dans le tableau D.

Tableau D
Erreurs par sexe, pour chaque statut d'activité

En %

Sexe	Erreur actifs occupés	Erreur chômeurs	Erreur inactifs
Homme	0,8	2,2	1,0
Femme	0,6	1,6	0,6

Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant dans les DOM et en ménage ordinaire.
Source : RP *2010 et enquêtes* Emploi *2010, 2011, 2012.*

Dans chaque croisement DOM / sexe, les effectifs estimés par statut d'activité BIT à partir du *RP* imputé sont très proches des effectifs-cibles (cf. tableau E), ce qui conduit à considérer le résultat de toute cette procédure comme très satisfaisant.

Tableau E
Estimation d'effectifs par DOM / sexe, par statut d'activité BIT : résultats de l'imputation

DOM	Sexe	Actifs occupés BIT		Chômeurs BIT		Inactifs BIT	
		RP	Enquête *Emploi*	*RP*	Enquête *Emploi*	*RP*	Enquête *Emploi*
Guadeloupe	Homme	62 270	62 386	16 370	16 250	54 256	54260
Guadeloupe	Femme	63 949	63 973	22 210	22 174	71 799	71 810
Martinique	Homme	62 883	62 697	15 002	14 958	52942	53 170
Martinique	Femme	66 910	66 861	19 397	19 569	70 310	70 187
Guyane	Homme	35 554	35 490	7 594	7 556	27 963	28 067
Guyane	Femme	27 864	27 794	9 289	9 239	38 224	38 344
La Réunion	Homme	129 550	129 670	49 167	48 805	104 444	104 685
La Réunion	Femme	108 567	108 372	48 109	48 118	150 621	150 807
Total		557 547	557 243	187 138	186 669	570 559	571 330

Lecture : à la Réunion, l'imputation considérée conduit à une estimation de 108 567 femmes actives occupées dans le champ considéré alors que l'enquête Emploi en estime l'effectif à 108 372.
Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant dans les DOM et en ménage ordinaire.
Sources : RP *2010 et enquêtes* Emploi *2010, 2011, 2012.*

Une procédure d'arrondi à la première décimale des taux de chômage BIT par DOM et par sexe conduit à un écart entre source *RP* d'une part et source *Emploi* d'autre part, qui ne dépasse pas 0,1 point d'indice (cf. tableau F) : on se situe donc nettement en deçà des seuils de qualité initialement fixés.

Tableau F
Taux de chômage par DOM / sexe, comparaison de la source *RP* imputé à la source *Emploi*

En %

Rang	Dom	Sexe	Taux de chômage BIT après imputation	Taux de chômage BIT selon l'enquête *Emploi*	Écart des taux après arrondi
1	Martinique	Homme	22,47	22,64	0,1
2	Guadeloupe	Femme	20,82	20,66	0,1
3	La Réunion	Homme	27,51	27,35	0,1
4	Guyane	Femme	17,60	17,55	0,1
5	Guyane	Homme	25,00	24,95	0,1
6	Guadeloupe	Femme	25,78	25,74	0,1
7	La Réunion	Homme	30,71	30,75	0
8	Martinique	Femme	19,26	19,26	0

Champ : ensemble des individus ayant 15 ans ou plus mais moins de 75 ans au 31 décembre de l'année concernée, résidant dans les DOM et en ménage ordinaire.
Sources : RP *2010 et enquêtes* Emploi *2010, 2011, 2012.*

Après avoir, comme en métropole, considéré comme inactives toutes les personnes de 75 ans et plus, et après avoir imputé à tout individu hors ménage ordinaire sa déclaration d'activité spontanée au *RP* en tant qu'activité BIT, la structure globale de l'activité après imputation peut être rapprochée de celle de la déclaration spontanée, en version non pondérée (cf. tableau G). Cela permet de juger de l'ampleur de la modification des déclarations d'activité spontanée nécessaire pour atteindre les objectifs fixés en matière de respect des taux de chômage BIT par croisement DOM / sexe.

Tableau G
Impact de l'imputation sur les changements de statut d'activité dans le *RP* 2010, version non pondérée

		Activité imputée BIT			
		Actifs occupés	Chômeurs	Inactifs	Total
Déclaration spontanée *RP* 2010	Actifs occupés	233 556	11 358	28 557	273 471
	Chômeurs	26 520	67 625	47 514	141 659
	Inactifs	20 699	13 083	261 638	295 420
	Total	280 775	92 066	337 709	710 550

Lecture : dans le fichier du RP *2010, 26 520 individus se déclarant spontanément chômeurs sont finalement imputés avec le statut d'actifs occupés au sens du BIT.*
Champ : ensemble des individus ayant 15 ans ou plus au moment de la collecte et résidant dans les DOM (tous types de ménages).
Source : RP *2010, non pondéré.*