LE SUIVI DE COLLECTE DE L'ENQUÊTE BUDGET DES FAMILLES 2017 : ASSURER LA REPRÉSENTATIVITÉ DES FAMILLES MONOPARENTALES

Oriane LAFUENTE (*), Thomas MERLY-ALPA (**)

(*) Ecole Normale Supérieure Paris-Saclay, département de sciences sociales
(**) Insee, Division Sondage

oriane.lafuente@hotmail.fr

Mots-clés : échantillonnage, collecte, classification hiérarchique par cluster, Tronc commun des enquêtes ménages, exploitation complémentaire du recensement

Résumé

La Direction de la recherche, des études, de l'évaluation et des statistiques du Ministère de la santé et des affaires sociales a utilisé l'édition 2011 de l'enquête BDF pour des études sur le coût de l'enfant et les échelles d'équivalence. Cependant, le faible nombre de familles monoparentales dans l'enquête a limité les possibilités d'exploitation et la précision des études sur ces familles particulières. La Drees et l'Insee ont donc conclu un partenariat afin de sur-échantillonner les familles monoparentales dans l'enquête BDF 2016 et ainsi obtenir un nombre suffisant de familles pour effectuer des études spécifiques sur cette catégorie de ménage. Du fait de l'absence d'information sur le type de ménage dans les enquêtes annuelles du recensement, un échantillon complémentaire de 2 000 familles monoparentales avec au moins un enfant mineur a été tiré dans les données de la Caisse nationale des allocations familiales.

Nous avons donc mis en place des outils de suivi de collecte permettant d'utiliser la structure des 6 vagues de 2 mois de l'enquête pour envisager des aménagements en cas de problèmes pour l'obtention du nombre de répondants et de représentativité.

Ce suivi de collecte a donc deux dimensions :

1) Assurer l'obtention de 1 500 familles monoparentales répondantes. Pour cela, nous avons proposé des objectifs mensuels de répondants, produits à partir des informations comprises dans les enquêtes annuelles du recensement permettant de repérer de potentielles familles monoparentales dans l'échantillon et des comportements de réponse à l'enquête BDF 2011. Nous avons également automatisé l'analyse des pots-CAPI des vagues de l'enquête.
2) Assurer la représentativité socio-démographique sur ce groupe. Nous avons réalisé une typologie des familles monoparentales grâce à une classification hiérarchique par cluster, appliquée à l'exploitation complémentaire du recensement. Nous avons ensuite automatisé la vérification de la représentation des profils types à chaque vague de l'enquête grâce aux informations recueillies dans le tronc commun des enquêtes ménages, exploitables environ un mois après la collecte.

Abstract en anglais

We present in this article the 2016 Family budget survey's collection follow up. One collection's goal was the obtention of 1 500 single-parents answering families. To achieve it, a 2 000 additional sample has been drawn from the National family allowances fund administrative files.
Collection follow up tools have been put in place, by using the survey's structure in 6 independant waves, to ensure the number goal and the respondents' representativity.

1. Introduction

L'enquête Budget des familles (BDF) est produite par l'Insee depuis 1956 et existe sous sa forme quinquennale actuelle depuis 1979. Elle vise à reconstituer la comptabilité des ménages en capturant l'ensemble de leurs dépenses et ressources. Ces données sont ensuite utilisées par la comptabilité nationale et pour la pondération des indices catégoriels de l'indice des prix, ainsi que par de nombreux acteurs extérieurs : services administratifs, chercheurs, bureaux d'étude... Du fait des forts enjeux de ses exploitations, l'enquête Budget des Familles est reconnue d'intérêt général et s'est vue conférer un caractère de réponse obligatoire.

La Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) du Ministère de la santé et des affaires sociales a utilisé l'édition 2011 de l'enquête pour des études sur le coût de l'enfant et les échelles d'équivalence (R.Hotte et H.Martin, 2015). Cependant, le faible nombre de familles monoparentales dans l'enquête a limité les possibilités d'exploitation et la précision des études sur ces familles particulières. La Drees et l'Insee ont donc conclu un partenariat afin de sur-échantillonner les familles monoparentales dans l'enquête BDF 2016 et ainsi obtenir un nombre suffisant de familles pour effectuer des études spécifiques sur cette catégorie de ménage. Du fait de l'absence d'information sur le type de ménage dans les enquêtes annuelles du recensement (EAR) utilisées pour le tirage des échantillons, un échantillon complémentaire de 2 000 familles monoparentales avec au moins un enfant mineur a été tiré dans les données de la Caisse nationale des allocations familiales (Cnaf) au 31/12/2015.

L'enquête BDF s'est déroulée sur une année de septembre 2016 à septembre 2017. Elle est échelonnée en 6 vagues de 8 semaines environ afin de tenir compte de la saisonnalité des comportements de consommation. Chaque vague est considérée dans sa gestion et son suivi comme une enquête à part entière. L'échelonnement permet un suivi de collecte rapproché et d'envisager des aménagements en cours d'année en cas de problèmes de représentativité. Nous avons alors mis en place des outils pour surveiller le déroulement de la collecte et en particulier l'obtention des 1500 familles monoparentales répondantes souhaitées par la DREES. Pour cela, nous avons utilisé les comportements de réponse à l'enquête BDF 2011, les remontées hebdomadaires des directions des enquêtes ménages régionales (DEM), les pots Capi des premières vagues et les bases brutes des troncs communs des enquêtes ménages (TCM) de la vague 1.

2. Obtenir 1 500 familles monoparentales répondantes

Notre premier objectif a été la mise en place d'outils pour assurer les objectifs numériques fixés par le contrat avec la Drees. Chaque vague comprenait environ 3 450 fiches adresses de l'échantillon principale et 330 de l'échantillon supplémentaire.

2.1. Estimer le nombre de familles monoparentales potentielles dans l'échantillon principal

Nous avons commencé par évaluer le nombre familles monoparentales présentes dans l'échantillon principal et le sous-échantillon CNAF. Ce travail comporte trois difficultés majeures :

- la complexité de définition des familles monoparentales
- les enquêtes annuelles de recensement à partir desquelles l'échantillon est tiré ne contiennent pas d'information sur le type de ménage
- le statut de famille monoparentale est peu durable, une partie des familles identifiées

comme telles dans l'enquête annuelle de recensement 2015 et la base CNAF auront changé de statut au moment de l'enquête

Une première difficulté est la définition du concept de famille monoparentale. En effet, les différentes sources n'utilisent pas forcément la même définition statistique, on trouve ainsi plusieurs définitions de la monoparentalité dans les enquêtes Insee. Au sens du tronc commun des enquêtes ménages, utilisé dans l'enquête BDF, un ménage est une famille monoparentale s'il est composé uniquement d'un parent ne vivant pas en couple et de son ou ses enfants. Si une tierce personne, quelle qu'elle soit, vit avec eux, alors le ménage est considéré comme complexe. Dans l'exploitation complémentaire du recensement, les ménages monoparentaux sont les ménages, avec comme famille principale une famille monoparentale. Certains ménages complexes peuvent donc être inclus dans cette définition et exclus de celle du tronc commun des enquêtes ménages.

Par ailleurs, les EARs dans lesquelles les échantillons sont tirés ne contiennent pas d'informations sur le type de ménage. Il est donc impossible de repérer de façon certaine les familles monoparentales de l'échantillon.

Quatre indicateurs pour prédire la monoparentalité des familles dans l'échantillon ont été testés :

- les ménages comprenant uniquement un majeur et avec une différence d'âge d'au moins 12 ans entre le majeur et l'enfant le plus âgé
- les ménages avec un seul majeur, une différence d'âge d'au moins 14 ans entre le majeur et l'enfant le plus âgé et où personne ne se déclare comme vivant en couple
- les ménages avec au moins un mineur, une différence d'âge de plus de 14 ans entre les deux habitants les plus âgées et où personne ne se déclare comme vivant en couple
- les ménages avec au moins un jeune de moins de 25 ans, dont les deux habitants les plus âgés ont au moins 14 ans de différence et où personne ne se déclare en couple

Le relâchement de la contrainte entre les différentes définitions nous permet d'arbitrer entre le repérage d'un plus grand nombre de familles effectivement monoparentales et l'augmentation des faux positifs.

Tableau 1 : Comparaison des indicateurs de familles monoparentales

Indicateur	Échantillon BDF 2011	Familles monoparentales répondantes	Échantillon BDF 2016
Un majeur uniquement	485	189	782
Un majeur, pas de couple	520	203	767
Au moins un mineur, pas de couple	926	302	1189
Au moins un jeune, pas de couple	1663	536	1555

Le relâchement la contrainte de l'unique majeur majeur, permet d'appréhender des ménages dont l'un des enfants déjà majeur vit encore avec son parent et ses frères et soeurs mineurs. Ce relâchement double le nombre de familles sélectionnées dans l'échantillon 2011. Cependant, l'indicateur perd également en précision, capturant davantage de ménages n'étant pas monoparentaux lors de la collecte. L'une des explications de cette hausse des faux positifs est l'utilisation de la variable "couple", qui peut-être mal renseignée et donc capturer des couples avec une différence d'âge importante et un enfant mineur comme ménage monoparental. Il se peut également que ces familles soient bien monoparentales au moment du recensement, mais changent davantage de statut que les familles avec uniquement un majeur dans les deux années séparant le recensement et l'enquête. En effet, le fait que l'un des enfants soit majeur laisse supposer que l'âge des autres enfants du ménage soit élevé et donc que l'ensemble des enfants puisse avoir décohabité au moment de l'enquête. Il est également possible que les parents de ces familles retrouvent plus facilement des conjoints du fait de l'âge de leurs enfants.

Nous avons finalement choisi d'utiliser l'indicateur autorisant la présence de plusieurs majeurs car il

permet de tenir compte de familles exclues par le premier indicateur et tout de même pertinentes pour l'analyse. Grâce à cet indicateur, nous avons pu estimer une première fois le nombre de familles monoparentales répondantes que l'on pourrait obtenir dans l'enquête BDF 2016. En appliquant les proportions trouvées en 2011, j'ai pu supposer que 364 des 1 189 familles trouvées avec l'indicateur seraient monoparentales et répondantes, auxquelles s'ajouteraient 297 familles non repérées, pour un total de 661 familles monoparentales répondantes dans l'échantillon principal.

2.2. Estimer le nombre de familles monoparentales potentielles dans le sur-échantillon Cnaf : la question de la durée de vie des familles monoparentales

Les familles monoparentales changent plus rapidement de statut que les autres types de famille. Une famille monoparentale identifiée comme telle à une date donnée a de fortes probabilités de changer de statut dans les mois suivants. Cette probabilité est fonction de la durée passée dans la monoparentalité. L'échantillon principal étant tiré dans la base de sondage du recensement 2015, c'est-à-dire en moyenne deux ans et trois mois avant la date d'enquête, la probabilité que les familles monoparentales supposées dans cet échantillon aient changé de statut entre le recensement et l'enquête est très importante. L'utilisation des données Cnaf devrait permettre de mieux capter les familles monoparentales les plus récentes, la base de sondage étant composée des familles présentes dans les registres Cnaf neuf mois avant le début de l'enquête.

Grâce aux informations issues de “Depuis combien de temps est-on parent de famille monoparentale” (G.Buisson, V.Costemalle et F.Daguet, 2015) j'ai produit un graphique estimant les taux de déperdition des familles monoparentales suivant leur durée dans ce statut au 31/12/2015. Pour ce calcul, j'ai commencé par appliquer à l'échantillon Cnaf les proportions des familles monoparentales selon l'ancienneté de la monoparentalité. Ainsi, 19 % des familles monoparentales ont moins d'un an, 39 % entre 1 et 5 ans et 25 % entre 5 et 10 ans.

J'ai ensuite appliqué à ces groupes, les probabilités de sortie de la monoparentalité en fonction du temps passé dans la monoparentalité à chaque vague de l'enquête BDF 2016.

Dans l'échantillon Cnaf, en moyenne 80 % des familles monoparentales identifiées au 31/12/2015 devraient l'être encore lors de la première vague contre 70 % à la fin de l'enquête. Ces chiffres varient entre 75 % pour les plus jeunes générations de famille monoparentales lors de la première vague et 87 % pour les familles monoparentales existants depuis quatre ou cinq ans. Par ailleurs, on ne trouvera plus que 60 % des plus jeunes familles en vague 6.

Cette estimation est très succincte et consiste uniquement en l'application de taux de variations mensuels moyens. Elle ne tient pas compte des possibles déménagements et des non-réponses. Elle ne permet donc pas d'estimer à elle seule le nombre de familles monoparentales obtenues grâce au sous-échantillon Cnaf, mais fournit une information que nous avons utilisé pour les estimations finales.

L'exploitation des pots-CAPI à la fin des vagues 1 et 2 a également permis d'étudier plus en détail cette question de durée de vie des familles monoparentales. Les fiches adresses du sous-échantillon Cnaf correspondaient à des familles monoparentales au 31/12/2015, les retours de collecte nous permettent de savoir lesquelles parmi les répondantes l'étaient encore 9 à 10 mois plus tard. En moyenne, 78.3 % des fiches adresses validées étaient encore monoparentales au moment de la vague 1, résultat cohérent avec les 81 % estimés précédemment. En vague 2, seul 72,2 % des fiches adresses CNAF étaient encore monoparentales contre 79,9 % estimés. Il y a d'importantes disparités régionales dans ces taux de maintien effectifs avec des taux inférieurs à 70 % dans certaines régions dès la vague 1.

La base de sondage CNAF est une compilation de bases de données départementales. Il se peut que ces fichiers soient de qualités inégales en fonction de la difficulté pour les caisses d'allocations familiales de les maintenir à jour. Les départements les plus problématiques se situent autour de

grandes agglomérations, ce qui est cohérent avec cette explication. En Île-de-France par exemple, un déménagement peut facilement entraîner un changement de département ou de région. Les CAF doivent donc probablement transférer davantage de dossiers, ce qui rend la gestion des fichiers plus complexe.

2.3. Production de scénarios de collecte

Ces travaux préliminaires permettent par la suite de produire une estimation du nombre de familles monoparentales répondantes dans l'enquête 2016. Les scénarios de prévision ont été déclinés au niveau régional. Ils peuvent être évalués à l'aune de l'objectif de 1500 familles répondantes et fournissent donc également une limite normative pour évaluer la collecte. La méthodologie diffère pour l'échantillon principal et le sous-échantillon CNAF, les informations disponibles étant différentes.

2.3.1. Dans l'échantillon principal

Afin d'estimer le nombre de familles monoparentales que l'on peut espérer obtenir à chaque vague dans l'échantillon principal, nous avons calculé trois proportions issues de l'enquête BDF 2011 afin de les appliquer à l'échantillon 2016 :

- le taux de réponse par région et par vague dans BDF 2011 ;
- le taux de familles monoparentales parmi les répondants par région dans BDF 2011 ;
- un taux d'actualisation fondé sur l'augmentation de la proportion de familles monoparentales, repérées par l'indicateur défini précédemment, dans l'échantillon 2016 par rapport à 2011 par région

Nous avons appliqué ces trois taux au nombre de ménages tirés dans l'échantillon 2016 à chaque vague et région et avons ainsi obtenu le nombre de familles monoparentales répondantes que l'on peut espérer obtenir effectivement dans la région à chaque vague.

$$FM_{2016,R,V} = FA_{2016,R,V} * \%Réponse_{2011,R,V} * \%FM_{2011,R} * Actualisation_{11-16}$$

L'utilisation du taux d'actualisation permet de tenir compte de l'augmentation du nombre de familles monoparentales au cours du temps. Nous avons calculé l'évolution de la proportion de familles monoparentales supposées par l'indicateur défini précédemment entre les échantillons 2011 et 2016. La proportion de familles repérées dans l'échantillon passe de 5.5 % à 6.4 % soit une augmentation de 16 % en moyenne.

Afin de tester la sensibilité des estimations, nous avons diversifié les sources pour la proportion de familles monoparentales régionales et fait varier les taux de réponses à l'enquête, qui peuvent avoir baissé du fait des nouvelles conditions d'emploi des enquêteurs et enquêtrices.

Ainsi nous avons également experimenté l'estimation :

1. Sans actualisation par rapport à 2011 : on applique exactement les proportions de réponse et de famille monoparentale de BDF 2011 à l'échantillon BDF 2016.
2. Avec les proportions de familles monoparentales par régions issues du recensement : on remplace les proportions trouvées dans l'enquête BDF 2011 par celles de l'enquête complémentaire du recensement 2013. Cette hypothèse est très optimiste et ne tient pas compte des taux de réponses différenciés ni de la contrainte d'au moins un enfant mineur.
3. En retirant 5 points de pourcentage aux taux de réponse dans les régions et 10 points en Île-de-France.

2.3.2. Dans le sur-échantillon Cnaf

Nous avons utilisé deux taux pour ce calcul :

- le taux de réponse par région et par vague issu de l'enquête BDF 2011
- le taux moyen de maintien des familles monoparentales à chaque vague (1.1.2)

$$FM.CNAF_{2016,R,V} = FA.CNAF_{2016,R,V} * \%Réponse_{2011,R,V} * \%Maintien_V$$

Ce modèle d'estimation néglige les possibilités de déménagement qui pourraient réduire le nombre de familles monoparentales répondantes si les familles partantes sont souvent remplacées par des familles classiques. Il fait également l'hypothèse d'un comportement de réponse similaire à la moyenne des ménages en 2011 alors que l'on peut supposer un taux de réponses inférieur. En effet, les familles monoparentales ne disposent généralement que d'un majeur susceptible de répondre à l'enquête, les possibilités de contact sont donc moins importantes, tout comme la disponibilité pour répondre au questionnaire.

2.3.3 Résultats : de l'estimation aux objectifs normatifs

Tableau 2 : Estimation du nombre de familles monoparentales lors de la collecte de BDF 2016

Scénarios	Actualisé	Non-actualisé	Taux de l'ECRP	Taux de réponse faibles	Moyenne
Familles monoparentales	1587	1501	1646	1428	1541
Dans l'échantillon principal	661	575	720	591	637
Dans l'échantillon CNAF	926	926	926	837	904

Le nombre de familles monoparentales espéré varie de 200 à la baisse ou à la hausse suivant les méthodes d'estimation. Dans la plupart des cas l'objectif de 1500 familles monoparentales est atteint avec plus ou moins de marge.

La méthode avec les données du recensement est la plus encourageante et pourrait paraître la plus fiable. Cependant, les familles monoparentales ne sont pas en elles-mêmes une strate du plan de sondage de l'échantillon principal et donc peuvent ne pas être représentées exactement de la même façon dans l'échantillon et dans le recensement du fait de l'aléa de sondage couplé à la structure de l'échantillon-maître et du recensement rotatif. De plus, il se peut que ces familles aient un comportement de réponse différent et soient donc moins représentées dans l'enquête qu'elles ne le sont dans la population totale.

La méthode paraissant la plus pertinente est celle utilisant les taux issus de BDF 2011 avec l'actualisation. Elle permet de tenir compte de l'accroissement du nombre de familles monoparentales entre les deux enquêtes et tient compte des comportements de réponse dans une enquête longue et complexe.

Il est inquiétant de constater la sensibilité des estimations aux taux de réponses. S'ils varient à la baisse du fait du Nouveau statut des enquêteurs et enquêtrices INSEE (NCEE), l'objectif de 1500 familles monoparentales interrogées pourrait ne pas être atteint. Les NCEE ont profondément impacté les taux de réponses, notamment en Île-de-France. Le suivi des premières vagues de collecte a confirmé le bien fondé de ces inquiétudes.

Pour le suivi de la collecte vague après vague, j'ai sélectionné trois scénarios afin de donner un intervalle dans lequel les résultats de la collecte devrait se trouver. Le maximum correspond au scénario avec les taux de familles monoparentales du recensement. La moyenne est la moyenne des

quatre scénarios. Le minimum correspond au scénario le plus bas ajusté pour obtenir les 1500 familles monoparentales souhaitées à la fin de l'enquête. Il permet de fournir une limite minimale normative pour les objectifs de collecte.

Tableau 3 : Estimation du nombre de familles monoparentales répondantes à chaque vague

	Minimum	Moyenne	Maximum
Vague 1			
Familles monoparentales	267	279	294
Dans l'échantillon principal	105	109	123
Dans l'échantillon CNAF	162	170	171
Vague 2			
Familles monoparentales	249	260	274
Dans l'échantillon principal	102	105	118
Dans l'échantillon CNAF	148	155	156
Vague 3			
Familles monoparentales	258	268	282
Dans l'échantillon principal	105	108	121
Dans l'échantillon CNAF	154	160	161
Vague 4			
Familles monoparentales	254	260	277
Dans l'échantillon principal	108	108	124
Dans l'échantillon CNAF	147	152	153
Vague 5			
Familles monoparentales	233	244	261
Dans l'échantillon principal	97	100	116
Dans l'échantillon CNAF	136	144	145
Vague 6			
Familles monoparentales	236	245	258
Dans l'échantillon principal	104	106	118
Dans l'échantillon CNAF	132	139	140

On prédit bien des difficultés pour la vague 2 -au moment des fêtes de fin d'année- et pour les dernières vagues, plus éloignées de la date de tirage de l'échantillon.

Ce sont les estimations régionales par vague qui sont utilisées pour le suivi de la collecte semaine après semaine.

2.4. Les outils du suivi de collecte

Ces estimations permettent de juger l'avancement de la collecte à chaque vague. Ainsi, on constate qu'il est normal d'observer des résultats plus faibles à certaines vagues ou dans certaines régions. La comparaison avec les résultats de collecte permet également de remettre en question les prédictions en regardant quelles hypothèses ne se sont pas réalisées.

2.4.1. Le suivi de collecte hebdomadaire

Du fait de la problématique particulière concernant les familles monoparentales, les Divisions enquêtes ménages régionales (DEM) ont fait remonter régulièrement des informations sur le statut des fiches adresses interrogées.

L'enquête budget des familles est une enquête longue et complexe. Elle nécessite deux visites à une semaine d'intervalle et le remplissage d'un carnet où le ménage note l'ensemble de ses dépenses durant la semaine. Cette charge importante pour le ménage explique en partie les difficultés à interroger les ménages. Pour le suivi de collecte, ce temps de passation important pose également des problèmes. En effet, les enquêteurs doivent également saisir informatiquement l'ensemble du carnet après la visite dans la famille. Les FA ne sont validées et remontées dans les suivis qu'une fois la deuxième visite réalisée et l'apurement effectué par les DEMs. Il y a donc un temps de latence important entre les résultats observés et les fiches adresses effectivement interrogées sur le terrain.

De ce fait, les remontées des informations de collecte arrivent tardivement et les résultats définitifs des vagues ne sont connus que deux à trois semaines après la fin officielle de la vague. Ainsi, pour la vague 1, l'île-de-France a vu son taux de collecte augmenter de 10 points de pourcentage dans les quinze jours suivant la fin de la vague.

J'ai mis en place des tables excel pour chaque vague reprenant par région :

- le nombre de FA sélectionnées et le nombre de FA répondantes attendues par simple application des taux de réponses de l'enquête BDF 2011 ;
- le nombre de FA CNAF sélectionnées et le nombre de FA monoparentales répondantes CNAF attendues ;
- le nombre de familles monoparentales attendues dans l'échantillon principal ;
- le nombre total de familles monoparentales répondantes ;

A chaque remontée des DEMs, les tableaux sont complétés par l'état actuel de la collecte à la date donnée. A la fin de la vague, un bilan est dressé, reprenant les objectifs estimés et les réalisations effectives. Il permet de juger de l'état d'avancement de la collecte et des possibilités de réalisations des objectifs. Sur la vague 1, le nombre de FA interrogées était légèrement supérieur au nombre attendu, signifiant qu'en moyenne les taux de réponses ont été meilleurs que pour BDF 2011 sur cette vague. Du point de vue des familles monoparentales, on obtient 5 familles répondantes de moins que la prédiction moyenne mais davantage que le minimum nécessaire pour assurer l'objectif de 1500 familles.

Ces résultats généraux masquent de grandes disparités régionales et entre les sous-échantillon. Ainsi, le nombre de FA monoparentales obtenues a été globalement supérieur aux prédictions dans l'échantillon principal, mais inférieur dans l'échantillon Cnaf. Les taux de réponses se sont améliorés dans la majorité des régions, ce qui explique les bons résultats dans l'échantillon principal. Par contre, il est probable que l'application des taux de réponses régionaux de BDF 2011 au sous-échantillon Cnaf surestime les taux de réponses réels. En effet, il est possible que les familles monoparentales concernées aient un profil de réponse différent de la moyenne régionale.

Les disparités régionales sont aussi très importantes. La Franche-Comté, le Poitou-Charentes et les Midi-Pyrénées ont vu leur taux de fiches adresses validées diminuer de plus de 10 points de pourcentage entre 2011 et 2016. D'autres régions comme le Limousin, PACA et la Bourgogne l'ont vu augmenter dans les mêmes proportions. La situation en Île-de-France est particulièrement inquiétante. En vague 1, la baisse a été contenue. Cependant, le taux de réponse en 2011 était déjà faible et cette région est donc la seule à avoir un taux de réponse inférieur à 50 % pour la vague 1 de 2016 en métropole. Ces fortes variations des taux de FA validées par rapport à 2011 impacte significativement la précision des prévisions de collecte, les régions avec de meilleurs taux que prévus ayant tendance à obtenir davantage de familles monoparentales et celles avec des taux plus faibles moins. Cette corrélation n'est pour autant pas automatique : la région Centre par exemple a vu son taux de réponse diminuer mais a obtenu davantage de familles monoparentales répondantes que prévu, notamment dans l'échantillon principal. L'échantillon était peut-être particulièrement riche en familles monoparentales à cette vague. De façon général, le nombre de familles monoparentales répondantes dans la région respecte les ordres de grandeur de la prédiction.

2.4.2. L’analyse des pots-CAPI

A la fin de chaque vague, les DEMs ont fourni leurs pots-CAPI, des bases de données reprenant pour chaque fiche adresse tirée les informations de collecte. Grâce à ces fichiers, on peut savoir quelles FA de l'échantillon étaient effectivement monoparentales au moment de la collecte. Ils permettent également d'analyser les déformations de l'échantillon induites par la non-réponse, notamment dans les régions à faible taux de réponses.

Ainsi nous avons pu analyser la répartition géographique des répondants, ainsi que les caractéristiques leurs caractéristiques grâce informations contenues dans la base de sondage aux caractéristiques de l'échantillon initial. Si les résultats sur les logements ou la taille de l'agglomération sont fiables, ceux sur les caractéristiques des habitants sont plus fragiles. En effet, il y a un délai de plus d'un an entre le recensement à l'origine de l’échantillon et le moment de l'enquête. Les habitants initiaux peuvent avoir déménagé ou changé de situation.

Les personnes seules paraissent plus difficiles à capter que les grands ménages, ce qui biaise légèrement la structure des répondants en ce sens.

Les caractéristiques du logement paraissent les plus discriminantes en termes de réponse. Ainsi, que ce soit à Paris ou dans l'ensemble de la métropole, les maisons sont sur-représentées parmi les répondants et les appartements sous représentés. Cet effet est lié aux problèmes d'accès au logement pour les enquêteurs. L'entrée dans les immeubles est souvent protégée par un ou plusieurs codes alors que l'accès est plus simple en maison. De même, les grands logements paraissent plus répondre que les plus petits et les répondants sont plus souvent propriétaires que la moyenne dans l'échantillon initial. Il y a donc un fort biais de sélection lié au logement qui pourrait influer sur les structures de consommation mesurées. Par ailleurs, la population rurale est également sur-représentée alors que les habitants de l'agglomération parisienne sont sous représentés.

Les caractéristiques sociales ne paraissent pas être l'objet d'un fort biais de sélection. Les différences en termes d'activité principale sont minimes et peu significatives, d'autant que ce statut est fortement susceptible de changer entre le recensement et l'enquête. Les non diplômés paraissent légèrement plus difficile à interroger que les autres. Les personnes détenant un CAP ou un BEP sont quant à elles légèrement plus représentées parmi les répondants que dans l'échantillon initial.

Les déformations de l'échantillon sont très faibles et habituelles. Il sera donc possible de corriger ces biais par la pondération.

3. Assurer la représentativité des familles monoparentales répondantes

3.1. La création d'une typologie des familles monoparentales à partir de l'enquête complémentaire du recensement

3.1.1. Méthodologie

Les travaux sur les familles monoparentales utilisent souvent des critères démographiques a priori pour déterminer différents types de familles. L'âge des enfants, le sexe du chef de ménage ou les raisons de la monoparentalité sont les caractéristiques les plus souvent utilisées pour affiner les analyses. Ces différenciations se justifient pour l'étude de publics particuliers visés par des politiques sociales. Cependant, elles sont décidées de façon a priori et appliquées à l'analyse. Il nous est paru intéressant d'utiliser une méthode inductive partant des caractéristiques socio-démographiques des familles pour faire émerger un espace social sans faire d'hypothèses préalables sur l'importance de certaines variables par rapport à d'autres ou sur la structure de leurs liens. Nous avons donc choisi d'utiliser une analyse des correspondances multiples, méthode qui identifie des associations entre les modalités de différentes variables et fait ressortir des facteurs marquants les associations les plus structurantes. Les facteurs sont ordonnées en fonction de leur importance dans la variance expliquée.

Pour définir des catégories arrêtées, nous avons réalisé une combinaison de deux méthodes, décrite dans "Principal component methods - hierarchical clustering - partitional clustering : why would we need to choose for visualizing data ?" (F.Husson, J.Josse, and J.Pagès, 2010), consistant en l'utilisation d'une analyse des correspondances multiples (ACM) en amont d'une classification hiérarchique sur composantes principales. L'ACM permet de transformer les variables qualitatives en variables continues via les composantes principales. Cela permet également de restreindre la classification aux dimensions les plus significatives et ainsi de réduire le biais et de rendre la classification plus stable.

Nous avons commencé par réaliser l'ACM sur les données de l'exploitation complémentaire du recensement (ECRP). L'ECRP a l'avantage d'interroger un nombre important de ménages et de contenir des informations sur les liens familiaux, ainsi que des caractéristiques socio-démographiques

détaillées sur la personne de référence. Nous avons ainsi pu réaliser une ACM et une classification hiérarchique sur composante principale sur une base de 798 948 ménages monoparentaux.

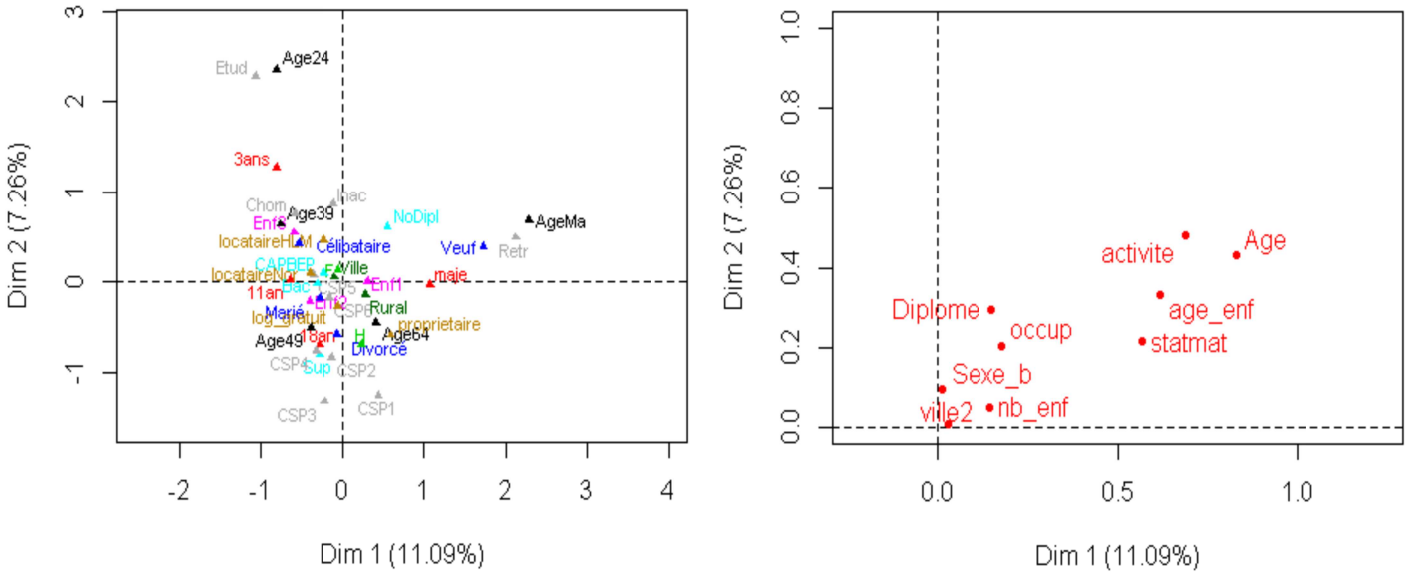
Pour réaliser l'ACM, nous avons sélectionné 9 variables :

- l'âge de la personne de référence en 5 groupes,
- son sexe,
- sa CSP ou son statut par rapport à l'emploi (inactif, chômeur, étudiant, retraité),
- son statut marital,
- le nombre d'enfants du ménage,
- l'âge de l'enfant le plus jeune
- le statut d'occupation du logement
- et l'environnement urbain.

Les variables choisies couvrent à la fois les caractéristiques du parent, du ménage et de l'environnement. Elles ont paru être les plus clivantes et suffisamment courantes pour être facilement réutilisables dans d'autres enquêtes.

Les résultats obtenus sont présentés dans des graphiques appelés plans factoriels. Chaque axe représente un facteur, une dimension expliquant une part de la variance.

Image 1 : Axes de l'analyse des correspondances multiples



Quatre variables expliquent une grande partie de l'inertie des deux premiers axes de l'ACM. Il s'agit de l'activité professionnelle, de l'âge, de l'âge de l'enfant et du statut matrimonial. L'axe 1 est également marginalement déterminé par le nombre d'enfant et le statut d'occupation du logement. Le diplôme et le statut d'occupation du logement expliquent quant à eux une part importante de l'inertie de l'axe 2.

Les modalités proches les unes des autres dans les plans correspondent souvent aux mêmes individus et deux individus aux caractéristiques similaires tendent à être proches dans les plans. Les modalités les plus communes se regroupent au centre du graphique vers l'origine des axes alors que les caractéristiques les plus clivantes s'en éloignent. On peut donc faire ressortir des pôles de similarités et d'oppositions en examinant la position des modalités et des individus de part et d'autres des axes.

On voit ainsi se dessiner des groupes de modalités proches. Aux extrêmes, le fait d'être à la retraite, veuf, d'avoir un enfant majeur et plus de 64 ans sont fortement corrélés. Et de même, être étudiant,

avoir moins de 25 ans et un enfant de moins de 4 ans sont fortement liés.

Les autres modalités sont moins clivantes et donc plus regroupées vers le centre de l'ACM. On peut tout de même interpréter le bas du graphique comme associant les diplômés du supérieur avec les CSP3 et CSP4, le fait d'être propriétaire et le sexe masculin. De même, le statut d'indépendant est corrélé à la propriété, au sexe masculin et à l'habitat en milieu rural.

Le statut marital semble être lié à l'âge : les célibataires semblent les plus jeunes et les veufs les plus âgés. Le haut centre du graphique associe les chômeurs et les inactifs avec le célibat, le fait de vivre en HLM et le fait d'avoir trois enfants ou plus. Ces regroupements sont intéressants, mais il est difficile d'établir une typologie précise regroupant les individus au cas par cas.

Pour définir précisément les catégories, nous avons utilisé les composantes principales issues de l'ACM pour réaliser une classification hiérarchique. Cette méthode agrège les individus en minimisant l'augmentation de la variance intra-catégorie à chaque étape. Cette agrégation est représentée par un arbre à partir duquel on peut choisir le nombre de clusters souhaités. Nous avons choisi d'en conserver six, correspondant aux catégories les plus archétypales et stables dans les différentes versions testées. Les profils sont très différents les uns des autres et cohérents en interne.

1) Les jeunes parents de moins de 25 ans, célibataires, vivant en ville avec un enfant de moins de quatre ans et sans activité professionnelle

2) Les familles nombreuses vivant en HLM, dont l'adulte est sans activité professionnelle et sans diplôme

3) Les mères célibataires de 25 à 50 ans, employées ou professions intermédiaires, diplômées du baccalauréat, avec deux enfants à charge en âge d'être scolarisés

4) Les adultes cadres ou des professions intellectuelles supérieures et intermédiaires, diplômés du supérieur, propriétaires dans les grandes villes, avec un ou deux enfants en âge d'être scolarisés dans le secondaire

5) Les hommes en milieu rural, agriculteurs, ouvriers ou indépendants, propriétaires de leur logement et avec un ou deux enfants en âge d'être scolarisés dans le secondaire

6) Les retraités veufs, non diplômés, cohabitant avec un enfant majeur

3.1.2. Limites de la méthode

La méthode est intéressante car elle part des corrélations entre les modalités des variables pour créer des groupes de façon inductive et non pas par imposition d'un modèle. Cependant, elle souffre de certaines limites.

La limite principale est le manque de clarté des catégories réalisées. Ainsi, les groupes ne sont pas déterminés par une unique caractéristique commune mais par leur proximité dans l'espace des modalités. Chaque groupe peut être défini par la modalité modale de chaque variables mais celles-ci est rarement partagée par une large majorité. Ce manque de précision rend complexe le classement d'individus à la combinaison de modalités non trouvée dans l'exploitation complémentaire du recensement et la communication sur les catégories. Par exemple, un homme ouvrier de moins de 25 ans avec deux enfants en bas âge et vivant en HLM ou une jeune femme en milieu rural n'auront pas forcément leur combinaison classée dans la typologie initiale et se trouvent à l'intersection de plusieurs catégories.

Ce manque de clarté s'associe à un problème de robustesse du classement des individus. Ainsi, j'ai testé ces travaux à de nombreuses reprises en modifiant certaines variables et sur d'autres échantillons, comme celui de l'enquête BDF 2011. Les catégories obtenues étaient souvent très

proches, faisant ressortir les mêmes dimensions clivantes pour leur description. En ce sens, la méthode et la typologie obtenue paraissent donc plutôt robustes. Cependant, les même individus ne se retrouvent pas toujours classés dans le même groupe suivant l'échantillon retenu. Il y a donc une grande porosité des catégories du fait de l'absence de caractéristiques discriminant précisément l'appartenance à l'une plutôt qu'aux autres.

L'exploitation complémentaire du recensement, grâce à son grand nombre d'individus, permet de couvrir un champs important des combinaisons de modalité, la classification paraît donc robuste. Cela permet également de limiter les classements manuels lorsqu'on applique la typologie à d'autres échantillons.

3.2. Utilisation pour le suivi de collecte

Cette typologie est utilisée pour le suivi de la collecte de l'enquête BDF 2016. BDF étant une enquête par vagues traitées comme six enquêtes indépendantes, il est possible d'exploiter sommairement les résultats des premières vagues grâce aux remontées anticipées des troncs communs des enquêtes ménages (TCM), qui contiennent l'essentiel des informations socio-démographiques. Nous avons ainsi pu vérifier si les familles monoparentales de chaque vague se répartissaient dans les catégories de la typologie dans les mêmes proportions que celles de l'exploitation complémentaire du recensement. Cette vérification permet de repérer rapidement d'éventuels déséquilibres ou trous de collecte.

Nous disposons initialement de peu d'informations sur les ménages du sur-échantillon CNAF. Il est probable que les familles présentes dans ces fichiers aient un profil particulier et ne soient pas représentatives de l'ensemble des familles monoparentales. L'exploitation des remontées anticipées permet de mesurer cet écart de profils, dont il faudra tenir compte pour les traitements post-collecte.

Grâce aux informations de la base de sondage et des TCM, on peut appliquer la typologie aux familles monoparentales présentes des premières vagues de BDF 2016. Cependant, le statut d'occupation du logement est disponible uniquement grâce aux informations de la base de sondage et n'existe donc pas dans les TCMs pour les ménages du sous-échantillon. Ainsi, on ne peut pas lui appliquer directement la typologie. Il y a également quelques valeurs manquantes pour 15 familles de l'échantillon principal qui empêchent leur inclusion dans la typologie.

Tableau 4 : Présence des catégories de la typologie dans la vague 1 de BDF 2016

Catégorie	ECRP	V1 BDF 2016
(1) Jeunes sans emploi avec un jeune enfant	2.8%	2.1%
(2) Familles nombreuses en HLM et sans activité professionnelle	14.1%	16.3%
(3) Femmes employées entre 25 et 49 ans avec un ou deux enfants scolarisés	35.4%	34.2%
(4) Adultes en CSP3 ou CSP4, propriétaire dans une grande ville	18.1%	19.5%
(5) Hommes ouvriers ou indépendants en milieu rural	8.4%	12.1%
(6) Retraités veufs sans mineurs	21.1%	15.8%
Total des familles monoparentales	798 948	190

Les ordres de grandeur sont similaires et on ne repère pas de trous de collecte. Les familles monoparentales avec des chefs de ménage âgés sont moins représentées. Cet écart peut venir d'une différence de définition de la monoparentalité entre l'ECRP et les TCM ou d'une probabilité de réponse plus faible des personnes plus âgées.

Les ménages tirées dans le sous-échantillon CNAF ne peuvent pas être inclus dans la typologie du fait de l'absence des informations sur leur statut d'occupation du logement et sur la taille de leur unité urbaine d'habitation, présentes dans la base de sondage EAR.

On dispose tout de même de la plupart des informations disponibles dans le TCM sur le ménage et les individus. On peut donc comparer ces caractéristiques dans le sous-échantillon CNAF à celles des

familles de l'ECRP et de l'échantillon principal. Cela nous permet d'évaluer le biais induit par l'utilisation des fichiers administratifs pour le sur-échantillonnage.

Les familles du sous-échantillon CNAF sont beaucoup plus nombreuses que la moyenne à avoir entre 25 et 40 ans et la personne de référence est plus souvent une femme. Les chefs de ménage sont plus souvent au chômage ou inactifs et ont deux fois plus souvent que les autres trois enfants ou plus. Leur enfant le plus jeune est pour 54 % d'entre elles en âge d'être scolarisé dans le primaire, mais elles ont également un peu plus souvent un enfant de moins de 4 ans. Ces déviations par rapport au reste de l'échantillon sont importantes et il faudra en tenir compte pour les analyses menées sur les familles monoparentales.

Ces déséquilibres peuvent s'expliquer de plusieurs façons. Tout d'abord, la Drees souhaite étudier les familles avec au moins un mineur et la base de sondage CNAF a été définie suivant cette contrainte. L'échantillon est donc biaisé en ce sens, ce qui peut expliquer le jeune âge des parents et des enfants. Il vaut donc mieux comparer les proportions obtenues dans l'échantillon CNAF à celles de l'ECRP restreinte aux familles monoparentales avec au moins un mineur. Les proportions obtenues sont plus proches mais des écarts importants demeurent, notamment concernant le sexe du chef de ménage, le rapport à l'emploi et la proportion de familles avec trois enfants ou plus.

Ensuite, la base de sondage est constituée des allocataires de la Cnaf, c'est-à-dire des personnes ayant demandé l'obtention d'allocations. De par ce biais de sélection, ils ne sont pas une copie conforme et représentative de l'ensemble de la population des familles monoparentales.

Le grand nombre d'enfant peut par exemple s'expliquer par l'ouverture de certaines aides à partir de deux enfants, ce qui sur-représenterait les familles avec deux enfants ou plus dans les fichiers de la Cnaf. De plus, les familles nombreuses ont un quotient familiale plus faible et potentiellement droit à davantage d'aides. L'absence d'emploi peut être relié au nombre important de jeunes enfants dans ces familles. Les parents isolés ont une contrainte temporelle très forte. L'activité professionnelle est mise en concurrence avec les frais de garde et l'ensemble des activités à accomplir pour les enfants, qui rendent le coût d'opportunité de l'emploi particulièrement important.

4. Conclusion

La collecte de Budget des familles se termine sur un taux de collecte métropolitain de 65% en moyenne, avec 15 944 ménages répondants en France entière hors Mayotte . Malgré les Ncee, les taux de réponses sont donc proches de ceux observés en 2011 et sont compréhensibles au vu de la complexité de l'enquête. Cependant, ces bons résultats cachent une grande hétérogénéité régionale. Ainsi, le taux de collecte n'est que de 50,4% en Ile de France, mais de 76,2% dans le Limousin. L'étude des pots-CAPI a permi de rassurer quant aux biais induits par la non réponse. En effet, elle est socialement bien répartie et a donc peu déformé la structure de l'échantillon, même en Ile-de-France. Seule la sixième vague a subi des déformations importantes quant à la structure des répondants par rapport à l'échantillon. Ses déformations pourront être corrigées lors des traitements post-collecte, au prix d'une perte de précision.

1 420 familles monoparentales ont répondu à l'enquête en métropole et 519 dans les DOMs. Ainsi, le partenariat avec la Drees est respecté, notamment grâce aux familles d'outre-mer. Les estimations fournies en début de collecte surestimaient les réponses des familles métropolitaines. Les erreurs d'estimations viennent essentiellement d'une surestimation des taux de réponse dans l'échantillon Cnaf, ainsi que d'une surestimation de la proportion de familles toujours monoparentales 9 mois à un an après le tirage de l'échantillon. On remarquait ainsi dès la seconde vague une chute brutale du nombre de familles monoparentales parmi les répondants de l'échantillon Cnaf.

Il est par ailleurs rassurant de constater que les familles monoparentales de l'échantillon principal

couvrent l'ensemble des situations repérées dans l'exploitation complémentaire du recensement de façon homogène et sont donc représentatives de la population générale. Les familles de l'échantillon Cnaf ont un profil plus particulier, signe d'un biais de sélection dans la base de sondage. Ces différences devraient être corrigées par une pondération adaptée lors des traitements posts-collecte.

Bibliographie

[1] Vanessa Bellamy. L'impact des enfants sur les budgets des ménages. Les familles monoparentales fragilisées. CNAF - Informations sociales, pages 46–52, 2007/1.

[2] Guillemette Buisson, Vianney Costemalle and Fabienne Daguet. Depuis combien de temps est-on parent de famille monoparentale. Insee Première, Mars 2015.

[3] Haut Conseil de la Famille. Le "coût de l'enfant". Rapport et propositions adoptés par consensus par le Haut Conseil de la Famille, Juillet 2015.

[4] Rozenn Hotte and Henri Martin. Mesurer le coût de l'enfant : deux approches à partir des enquêtes budget de famille. DREES - Dossiers solidarité et santé, Juin 2015.

[5] Jean-Michel Hourriez and Lucile Oulier. Niveau de vie et taille du ménage : estimations d'une échelle d'équivalence. Economie et Statistique, pages 65–94, Octobre 1998.

[6] Francois Husson, Julie Josse, and Jérôme Pagès. Principal component methods - hierarchical clustering - partitional clustering : why would we need to choose for visualizing data ? Technical Report - Agrocampus, Septembre 2010.

[7] Elodie Kranklader and Amandine Schreiber. Le sentiment d'aisance financière des ménages : stable au fil des générations, mais fluctuant au cours de la vie. Insee Références - Dossier, page 69, 2015