Compte-rendu intermédiaire

Projet ANR-16-CE41-0007

Big_Stat

Programme DS0806 2016

A IDENTIFICATION

Acronyme du projet	Big_Stat
Titre du projet	Big Statistical Data for a Mobile Society
Coordinateur du projet (société/organisme)	Laurent Toulemon
Date de début du projet Date de fin du projet	Mars 2017 Mars 2021
Labels et correspondants des pôles de compétitivité (pôle, nom et courriel du corresp.)	
Site web du projet, le cas échéant	https://big-stat.site.ined.fr/

Rédacteur de ce rapport	
Civilité, prénom, nom	M. Laurent Toulemon et Mme Giulia Ferrari
Téléphone	01 56 06 21 16
Courriel	Toulemon@ined.fr
Date de rédaction	31/07/2018
Période faisant l'objet du rapport d'activité	1 Mars 2017 - 30 Septembre 2018

B LIVRABLES ET JALONS

N°	Intitulé	Nature*	Date de fourniture			Partenaires (souligner le responsable)
			Prévue initialement	Replanifiée	Livrée	
1	Accès aux données par le CASD	Jalon	3-2017		3-2017	
2	Ajouts de membres et de sources (ajouts acceptés par le Comité du secret statistique les 9-6-2017, 13-10-2017, 2-2-2018, 29-6-2018	Jalon			Divers	
3	Réunion de lancement	Séminaire	10-2016		10-2016	
4	Réunion du WP « Structures familiales »	Séminaire	10-2017		10-2017	
5	Réunion du WP « jeunes adultes »	Séminaire	10-2017		02-2018	
6	Réunion de l'advisory Board	Jalon	6-2017	11-2018		
7	Présentation à des conférences, articles de recherche (voir liste ci-dessous)	Jalon				

C RAPPORT D'AVANCEMENT

C.1 OBJECTIFS INITIAUX DU PROJET

Centré sur l'utilisation des données administratives françaises en sciences sociales, le projet comporte trois aspects.
Tout d'abord, des recherches sur des sujets importants en sociodémographie, pour lesquels les données administratives viennent compléter les données d'enquête. Trois sujets forment le cœur des projets de recherche. 1) évaluation des doubles comptes dans les enquêtes et le recensement, description de la situation familiale des habitants en tenant compte des personnes recensées ou enquêtées deux fois car elles ont deux logements habituels. 2) formation et rupture des couples par les jeunes adultes. 3) analyse des situations familiales et socioéconomiques des enfants de parents séparés qui partagent leur temps entre les deux domiciles parentaux.
Ensuite, une validation des données, en collaboration avec les producteurs, fondée d'une part sur une comparaison entre sources et une confrontation avec des données d'enquêtes sociologiques (y compris des entretiens qualitatifs) pour analyser les situations concrètes qui se cachent derrière des situations familiales mal recensées ou mal identifiées par la statistique publique et, d'autre part, par un retour aux données initiales en cas de résultat étrange. Ces validations peuvent conduire, comme c'est déjà le cas pour l'Échantillon démographique permanent (EDP) de l'Institut de la statistique et des études économiques (Insee), à des enrichissements ou des corrections des fichiers.
Enfin, un effort de diffusion et de mise à disposition des données administratives qui se traduit par la réalisation d'une liste de diffusion du projet, de sites internet dédiés à chaque source, et d'actions de formation des utilisateurs de ces données.

C.2 TRAVAUX EFFECTUES ET RESULTATS ATTEINTS SUR LA PERIODE CONCERNEE

Le projet a rassemblé dans un premier temps des travaux portant sur l'Échantillon démographique permanent de l'Insee, qui regroupe les données du recensement et de l'état civil depuis 1968, et dont l'enrichissement récent aux données sociales et fiscales en fait un fichier de données extrêmement riche, mais dont la documentation complète et la validation gagnent à tirer bénéfice de retours des utilisateurs. Benjamin Marteau, Laurent Toulemon et Sébastien Durier ont estimé la proportion et le nombre d'individus en double compte au recensement et ils décrivent les situations familiales et sociales de ces individus. Julien Boelaert a utilisé des méthodes de *machine learning* pour distinguer les vrais couples de même sexe au recensement à partir de l'enquête Famille. Giulia Ferrari et Laurent Toulemon, en utilisant les données socio-fiscales appariées avec les EAR, ont analysé l'évolution des situations conjugales et les transitions entre les différents états conjugaux en France en 2010 et 2015. Giulia Ferrari, Carole Bonnet et Anne Solaz ont étudié la mobilité résidentielle suivant un divorce ou une rupture de PACS, avec un focus particulier sur les parents et sur le rôle du type de garde des enfants. Marie-Caroline Compans a exploré l'état civil pour étudier la fécondité.
De très nombreux projets de recherche ont été lancés (voir annexe 1).

Ces travaux ont donné lieu à des retours vers l'Insee, producteur de l'EDP (voir annexe 2). Les travaux de validation des sources se poursuivront.

Nous avons créé un site web du projet https://big-stat.site.ined.fr/ où nous présentons les sources potentiellement disponibles pour la recherche et les moyens d'y accéder, les travaux effectués à partir de ces données et financés par ce projet. Le site regroupe également des références techniques et théoriques sur l'utilisation des données administratives et des liens vers les expériences similaires, en France et à l'étranger.

Ensuite, nous avons créé un site participatif pour les utilisateurs de l'EDP (https://utiledp.site.ined.fr), où ils peuvent consulter la documentation des données (qui n'est pas disponible par ailleurs) et contribuer à son enrichissement en proposant des codes de variables construites par eux-mêmes et exploitables par d'autres utilisateurs, sur le modèle du site des utilisateurs de la cohorte d'enfants Elfe. Le site contient ainsi un ensemble cohérent de métadonnées : documentation officielle produite par les producteurs, variables ajoutées par les utilisateurs, notes sur les fichiers. Cet ensemble est mis à jour et corrigé en permanence.

Des sites similaires sont en préparation pour les enquêtes annuelles de recensement, le fichier regroupant les « troncs communs » des enquêtes de l'Insee auprès des ménages, les données de la Caisse national des allocations familiales (Cnaf) qui vont être mises à disposition en octobre 2018. D'autres sont envisagés pour les enquêtes européennes EU-Silc et le fichier des données fiscales Fidéli.

Nous avons organisé au printemps une formation aux méthodes d'analyse des données massives et aux routines utilisables en langage R, et participé au financement d'une formation de l'Ined sur les données EU-Silc. Une école d'été sur l'Échantillon démographique permanent sera organisée à l'été 2019.

Le tout est conforme au plan initial.

C.3 DIFFICULTES RENCONTREES ET SOLUTIONS

Nous n'avons pas rencontré de difficulté.

C.4 FAITS ET RESULTATS MARQUANTS

L'analyse de l'Échantillon démographique permanent, en collaboration entre l'Insee et l'Ined, a permis de mesurer (pour la première fois depuis l'adoption des enquêtes annuelles de recensement en 2004) la fréquence des doubles comptes au recensement à 2%, et de discuter de la précision et de la portée de ce résultat avec les responsables du recensement, notamment en termes d'estimation des situations familiales complexes souvent associées à des doubles comptes (enfants de parents séparés, jeunes adultes plus ou moins partis de chez leurs parents).

Le site du projet https://big-stat.site.ined.fr , en français et en anglais, regroupe un ensemble d'informations utiles pour les chercheurs souhaitant utiliser des données administratives ou des données ouvertes ou des données contextuelles. Un ensemble cohérent d'articles théoriques sur les données administratives et les données massives en sciences humaines et sociales est présenté, et complété par de nombreux exemples de travaux utilisant de telles données, en France et ailleurs. Le lien vers www.data.gouv est complété par des liens vers les principales bases de données dans les domaines de la démographie, de la santé, de l'équipement et des services territoriaux, de l'économie et des transports. De même les principales bases de données contextuelles sont référencées.

C.5 TRAVAUX SPECIFIQUES AUX ENTREPRISES (LE CAS ECHEANT)

C.6 REUNIONS DU CONSORTIUM (PROJETS COLLABORATIFS)

Date	Lieu	Partenaires présents	Thème de la réunion
19/10/2016	INED	Tous les membres du projet (18 participants)	Réunion de lancement du projet
04/10/2017	INED	14 participants	Réunion des participants à l'axe 2, jeunes adultes
09/02/2018	INSEE	13 participants	Réunion des participants à l'axe 1, comptage de la population et situations familiale

C.7 COMMENTAIRES LIBRES

Commentaires du coordinateur

La collaboration efficace entre les membres du projet a permis de corriger ou de compléter les données de l'Échantillon démographique permanent (EDP). Les allers-retours entre l'Insee, producteur de l'EDP, et les utilisateurs sont tout à fait conformes à l'ambition initiale du projet, fondée sur le constat que les données administratives ou les données complexes comme l'EDP ne peuvent être parfaitement validées et documentées par les producteurs. Le projet a permis des recherches collaboratives et a conduit les utilisateurs à valider les éléments du fichier nécessaires à leur recherche, puis à faire des retours vers l'Insee, qui a pu ainsi améliorer le fichier et sa documentation. Une note de l'Insee, reproduite en annexe, en atteste.

Le projet regroupe au Centre d'accès sécurisé aux données (CASD) pas moins de 38 chercheurs. Alors que le modèle du Comité du secret est plutôt fondé sur des petits projets déconnectés les uns des autres (pour garantir au mieux le respect de la confidentialité), le projet Big_Stat valorise la mise en commun des expériences et les retours vers les producteurs, offrant ainsi un autre modèle de recherche, plus collaboratif et associant pleinement l'Insee, producteur des données.

Outre l'Ined, les membres du projet appartiennent au CNRS, à l'Inserm ou à diverses universités en France (Bordeaux, Strasbourg, Nanterre, Paris Descartes, Paris Sorbonne) et à l'étranger (Penn University, université d'Anvers). Centré à l'origine sur l'EDP, le projet s'élargit à de nombreuses autres sources, soit déjà disponibles (Fichiers démographiques sur les logements et les individus (Fidéli), Enquête emploi en continu et Enquêtes Formation qualification professionnelle de l'Insee, enquêtes européenne EU-SILC) soit qui le seront bientôt (Tronc commun des enquêtes auprès des ménages de l'Insee, enquêtes annuelles de recensement avec leur pondération spécifique, fichiers de la Caisse nationale des allocations familiales, causes de décès enrichissant l'EDP).

De nouveaux membres sont venus se joindre au projet. Aux vingt-deux membres initiaux se sont ajoutés dix doctorants, cinq post-doctorants, six chercheurs et trois enseignants-chercheurs en France. Des contacts ont été pris avec des responsables de projets similaires en Belgique (université d'Anvers), en Allemagne (MPIDR de Rostock), au Canada (McGill à Montréal) et aux États-Unis (Penn University ; Madison, Wisconsin). Un chercheur et un post-doctorant de l'université d'Anvers ont également rejoint le groupe.
En rajoutant trois administratifs et neuf inscriptions pour information, la liste de diffusion comprend actuellement 60 membres actifs ou proches du projet.

Commentaires des autres partenaires

Question(s) posée(s) à l'ANR

D VALORISATION ET IMPACT DU PROJET DEPUIS LE DEBUT

D.1 PUBLICATIONS ET COMMUNICATIONS

Liste des publications monopartenaires (impliquant un seul partenaire)		
International	**Revues à comité de lecture**	1. Toulemon Laurent. 2017. Undercount of young children and young adults in the new French census, Statistical Journal of the IAOS, Vol 33, p. 311–316. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji1054
	Communications (conférence)	1. Giulia Ferrari et Laurent Toulemon « Is the French PACS similar to cohabitation or marriage? New data on family situations and transitions ». Présenté à la conférence annuelle de la PAA (Population Association of America) à Denver (USA) en avril 2018. 2. Giulia Ferrari et Laurent Toulemon « Is the French PACS similar to cohabitation or marriage? New data on family situations and transitions ». Présenté à la conférence Européenne de Population (EPC) à Bruxelles en juin 2018. 3. Giulia Ferrari, Carole Bonnet et Anne Solaz « Who keeps the children will keep the house? Residential mobility after divorce and civil partnerships' dissolution ». Présenté à un workshop dédié aux différences internationales dans la mobilité résidentielle après le divorce à Saint Andrews (Écosse) en mai 2017. 4. Giulia Ferrari, Carole Bonnet et Anne Solaz « Who keeps the children will keep the house? Residential mobility after divorce and civil partnerships' dissolution ». Présenté à Divorce Conference à Anverse en octobre 2017. 5. Laurent Toulemon, Sébastien Durier et Benjamin Marteau "Two homes, two families? People counted twice in the French rotating census". Présenté à XXVIII IUSSP International Population Conference, Session 257: Improving collection and quality of demographic data in censuses and surveys, Cape Town (South Africa), 29 October - 4 November 2017. 6. Laurent Toulemon, Sébastien Durier et Benjamin Marteau, "Towards a Precise and Accurate Analysis of the Situation of Two-Home Children in France from New Demographic Panel Data". Présenté à la European Population Conference 2018, European Association for Population Studies (EAPS), Brussels (Belgium) en juin 2018 .
France	**Communications (conférence)**	1. Laurent Toulemon, Sébastien Durier et Benjamin Marteau, « Au recensement, 2,3 % de doubles comptes, d'après l'échantillon démographique permanent ». Présenté aux 13es Journées de Méthodologie Statistique de l'Insee (JMS 2018), Paris (France), 12-14 juin.

D.2 AUTRES ELEMENTS DE VALORISATION

Liste des éléments. Préciser les titres, années et commentaires	
Autres (sites web)	1. Site du projet : https://big-stat.site.ined.fr/ 2. Site des utilisateurs de l'EDP : https://utiledp.site.ined.fr

D.3 POLES DE COMPETITIVITE (PROJET LABELLISES)

Collaboration du projet avec le(s) pôle(s) ayant labellisé

Activités financées par le complément de pôle (laboratoires publics uniquement)

D.4 PERSONNELS RECRUTES EN CDD (HORS STAGIAIRES)

Identification				Avant le recrutement sur le projet			Recrutement sur le projet			
Nom et prénom	Sexe H/F	Adresse email (1)	Date des dernières nouvelles	Dernier diplôme obtenu au moment du recrutement	Lieu d'études (France, UE, hors UE)	Expérience prof. antérieure (ans)	Partenaire ayant embauché la personne	Poste dans le projet (2)	Date de recrutement	Durée missions (mois) (3)
Ferrari Giulia	F	giulia.ferrari@ined.fr		Doctorat	UE	6	Ined	Post-Doc	01/03/2017	24

D.5 ÉTAT FINANCIER

Nom du partenaire	Crédits consommés (en %)	Commentaire éventuel
INED	35%	Recrutement d'un post-doc et abonnements au CASD conformes aux prévisions. Dépenses de personnel : 76 344 € Dépenses de fonctionnement : 19 841 € Total : 96 185 € (sur 269 986 € attribués au total)

E ANNEXES EVENTUELLES

E.1 ANNEXE 1. PRESENTATION TRES SYNTHETIQUES DES PROJETS DE RECHERCHE

Le travail est collaboratif, et de nombreux projets de recherche sont inclus dans le programme. Cette annexe les présente très brièvement sous forme de liste.

- Benjamin Marteau et Marie Bergström s'intéressent à étudier le lien entre précarité professionnelle et instabilité conjugale des jeunes adultes avec l'enquête SRCV et les données fiscales.
- À partir des données des EAR de 2004 à 2016, Antoine Robin examine les conditions de logements et l'insertion professionnelle des jeunes natifs des Dom en métropole.
- Yajna Govind étudiera les inégalités des distributions de revenus dans les DOM et la comparera à la situation en Metropole.
- Louise Caron, Myriam Khlat et Lidia Panico utiliseront l'EDP pour décrire l'évolution des conditions de vie des enfants nés de parents immigrés, et leur impact sur les profils sociodémographiques et la mortalité à l'âge adulte.
- Julien Boelaert, Giulia Ferrari et Benjamin Marteau proposent de retravailler la question des inégalités de trajectoires dans le passage à la vie adulte à partir de l'EDP et des données sociales et fiscales en employant des réseaux de neurones convolutifs.
- Baptiste Coulmont a repéré des couples mariés du même sexe dans le fichier détail du recensement et dans l'EDP. Avec Gaëlle Meslay ils étudieront leurs caractéristiques sociodémographiques et la présence d'enfants.
- Baptiste Coulmont a travaillé sur le fichier électoral de l'EDP, pour décrire la population des non-inscrits et de personnes "inscrites ailleurs". Il compte poursuivre cette étude en lien avec l'exploitation de l'enquête Participation électorale 2017.
- Marie-Caroline Compans compte étudier les déterminants du report et du rattrapage de la fécondité à des âges tardifs.
- John Tomkinson travaillera sur les enfants de moins de 5 ans non recensés ainsi que sur l'agrandissement de famille à partir des données fiscales et panel DADS.
- Alessandra Trimarchi utilisera l'EDP pour étudier la formation des couples, l'homogamie éducative, socio-économique et d'âge en France.
- Cyril Jayet travaillera sur les variables mesurant l'origine sociale de l'enquêté.
- Cécile Flammant termine une thèse sur les orphelins en France. Elle a utilisé les données de la Cnaf et va mettre en ligne un site Internet facilitant l'accès aux fichiers statistiques auxquels elle a eu accès et qui seront disponibles en octobre.

E.2 ANNEXE 2. REPRODUCTION D'UNE NOTE DE L'INSEE

Note n° 2018_12063_DG75-F170 du 20 juillet 2018
Objet : Apports des échanges avec les utilisateurs à la production et à la documentation de l'Échantillon démographique permanent
Isabelle Robert-Bobée ; personne chargée du dossier : Sébastien Durier

La division Enquêtes et études démographiques (EED) de l'Insee prépare chaque année une nouvelle base étude de l'échantillon démographique permanent (EDP), qui est un panel d'individus, en y ajoutant des données plus récentes. Au départ le panel EDP était centré sur la compilation de données extraites de trois sources : des données des enquêtes annuelles de recensement, des données d'état civil et des données du fichier électoral. L'EDP s'est ensuite enrichi de deux nouvelles sources. Depuis l'enrichissement de l'échantillon démographique permanent par les données du panel « tous salariés » dans la base Études 2013 (diffusion début 2015) et les données fiscales (Fidéli et FiLoSoFi) dans la base Études 2014 (diffusion début 2016), la communauté des utilisateurs de l'EDP s'est grandement élargie. La possibilité d'accéder à l'EDP via le CASD a aussi fortement contribué à cette expansion. Pour accompagner les utilisateurs, la division EED a décidé de mettre en place depuis novembre 2015 un groupe des utilisateurs. Elle a aussi ouvert la possibilité aux utilisateurs d'échanger par mail avec les producteurs. Les échanges avec les utilisateurs ont aussi lieu grâce à la participation de la division EED aux réunions sur les comparaisons de sources utilisant l'EDP, pilotées par l'Ined dans le cadre de l'ANR « des données massives pour une société mobile ». Cela contribue aussi à des contacts plus proches entre producteurs et utilisateurs, et une meilleure diffusion de l'EDP, la documentation de la base étant disponible sur un site internet accessible à tous.
Si l'objectif initial de ces échanges est évidemment d'aider les utilisateurs dans la réalisation de leur projet d'études ou de recherche, ceux-ci en retour peuvent contribuer à l'amélioration de l'EDP. On propose dans cette note de présenter les apports des utilisateurs dans trois domaines : les enrichissements par de nouvelles données ou variables, les corrections ou améliorations de la documentation et enfin les corrections d'erreurs dans les données de l'EDP.

1- Enrichissements de l'EDP suite aux demandes ou aux questions des utilisateurs

Un des objectifs du groupe des utilisateurs est de pouvoir recueillir les besoins des utilisateurs. En particulier, l'EDP récupérant des données de cinq sources différentes, des choix ont été opérés par les producteurs parmi la masse des informations disponibles. Ces choix peuvent ne pas s'avérer finalement les meilleurs au regard des besoins des utilisateurs.

a- Ajout de variables brutes

Les sources que l'EDP mobilisent sont des données initialement transversales, qui proposent souvent des variables redressées (notamment par imputation de la non-réponse). Pour l'EDP, les variables redressées peuvent être utiles, mais dans le cas très fréquent d'une mobilisation en panel, disposer des données brutes s'avère essentiel pour les utilisateurs, pour reconstituer eux-mêmes des données manquantes en tenant compte des différentes réponses déjà apportées par le passé ou dans d'autres sources intégrées à l'EDP. La stratégie des producteurs consiste donc à fournir autant que possible les variables à la fois dans leurs versions brutes et dans leurs versions redressées. Cette stratégie générale n'avait en pratique pas toujours été adoptée pour chacune des variables. Deux oublis de variables brutes ont ainsi été corrigés dans la base études 2016 (BE2016) suite à des demandes d'utilisateurs :
- ajout de la variable IRAN_X, qui donne la réponse à la question du recensement « Où étiez-vous un an auparavant ? », et qui permet de compléter la variable IRAND dans laquelle les non-réponses à la variable IRAN_X sont imputées (étude sur les double-comptes au recensement, INED)
- rang de naissance : remplacement des valeurs redressées des variables CTX_NAV_PREC_DATE, donnant la date de la naissance précédente, et CTX_MERE_VIVANT_ENF_PREC_NBR, donnant le nombre de naissances précédentes, par les valeurs brutes issues du bulletin de naissance. (étude sur les naissances tardives, INED)

b- Variables apparues dans les sources

Les sources mobilisées par l'EDP ont évolué au cours du temps. En particulier des variables nouvelles peuvent faire leur apparition dans ces sources. La veille opérée par les producteurs peut s'avérer insuffisante. Par exemple, une demande d'une chargée d'étude de la Drees sur la disponibilité de deux « cases » de la déclaration fiscale (sommes versées pour la garde d'enfant à domicile et sommes versées pour l'emploi de personnes à domicile) a permis aux producteurs de constater l'apparition des deux variables correspondantes dans les fichiers sources (GARDEMM et SERVDOMM) et de les ajouter à la base études 2016.

c- Récupération de données anciennes

L'EDP compile des données depuis le recensement de 1968. Ces données anciennes sont en théorie complètes et bien documentées. Il peut cependant arriver que des données anciennes puissent être récupérées et ajoutées à la base études. C'est le cas pour les recensements 1990 et 1999. Dans ces deux recensements les identifiants de l'îlot du recensement sont disponibles dans l'EDP, mais sans possibilité de faire le lien entre eux. Des questions d'un utilisateur ont permis aux producteurs de connaître l'existence d'une table de passage entre les îlots 1990 et les îlots 1999 qui sera ajoutée à la base études 2017 (étude sur la mobilité des immigrés, Paris 1).

2- Améliorations de la documentation

La mise au point d'une documentation exhaustive et efficace est une des tâches les plus difficiles, notamment pour l'EDP qui compile plusieurs sources sur une longue période. Les erreurs, manques ou insuffisances soulignées par les utilisateurs permettent ainsi une amélioration continue de la documentation.

a- Suivi des modifications du contenu des variables dans les sources

Un défaut possible de la documentation de l'EDP consiste en la non prise en compte de modification dans le contenu des variables. Par exemple, à partir de 2015, le questionnaire individuel des EAR a été modifié notamment sur la question de la vie en couple ainsi que sur le plus haut niveau de diplôme atteint. Ces changements de modalités ont bien été intégrés à la documentation. Cette refonte de l'enquête de recensement a aussi été accompagnée de changements de traitement de variables sur la famille (intégration des couples de même sexe dans le traitement dits de l'analyse-ménage-famille du recensement), qui n'avaient pas été mentionnés dans la documentation de l'EDP et ont conduit des utilisateurs à poser des questions au producteur, qui a complété la documentation. Deux ajouts ont été faits :
- les variables type de ménage détaillé (TYPMD) et type de famille détaillé (TYPFD) prennent en compte à partir de 2015 les couples de même sexe dans leurs modalités concernant les couples (étude sur la mobilité sociale, Insee Île-de-France)
- pour faire le lien entre les niveaux de diplôme sur longue période (depuis le recensement de 1968) des programmes pour convertir les variables de diplôme dans une nomenclature harmonisée (SAPHIR) étaient fournis dans la documentation de l'EDP. Le changement de questionnaire de 2015 n'avait cependant pas été pris en compte dans le programme fourni. Il l'est depuis la BE2016 suite à une question d'un utilisateur.

b- Amélioration de la documentation des données anciennes

Lorsque les utilisateurs mobilisent l'EDP en panel sur longue période, ils sont amenés à demander des précisions sur des variables anciennes. Si la documentation n'a pas été suffisamment complète au moment de la récupération des données, les informations ne sont pas ou plus facilement disponibles. Par exemple, la variable catégorie socio-professionnelle pour les DADS avant 1982 (CS2_ANC) n'était pas documentée car non-présente dans la documentation fournie par la source. Un utilisateur souhaitant mobiliser la variable a repéré le manque et une recherche dans les archives a permis de compléter la documentation pour la BE2016 (étude sur la mobilité sociale et le niveau de vie, France Stratégie).

3- Corrections des erreurs dans des données de l'EDP

Malgré les efforts déployés par les producteurs pour contrôler le contenu des variables dans l'EDP, des erreurs peuvent toujours passer inaperçues, d'autant que les producteurs n'utilisent pas en pratique l'intégralité des données, mais seulement une partie pour leurs propres études. L'utilisation des données conduit à mieux les connaître et facilite le repérage d'erreurs.

a- Pondération en panel

Depuis la BE2014, une variable de pondération pour utiliser les EAR en panel dans l'EDP (POIDS_PANEL_5) est disponible. Cependant, une erreur dans le programme de calcul entraînait une légère surpondération pour les départements et régions d'outre-mer qui a été détectée par un utilisateur (étude sur la mobilité résidentielle, Université de Bordeaux et INED) : dans un premier temps, une solution à mettre en oeuvre dans la BE2016 par les utilisateurs eux-mêmes a été proposée ; dans un second temps la variable de pondération corrigée sera livrée avec la BE2017.

b- Identifiant famille dans l'EAR 2004

Un des intérêts de l'EDP est de disposer d'informations sur les individus habitant le même logement que les individus EDP. Par exemple, dans les EAR l'identifiant famille (ID_FAM_DIFF) permet de connaître les caractéristiques des parents, du conjoint ou des enfants de l'individu EDP. Dans ses travaux préparatoires, un utilisateur a pu détecter l'absence d'un grand nombre d'identifiants famille dans l'EAR 2004 ce qui a été corrigé pour la BE2016 (étude sur les unions libres, Insee DG)

c- Code ilot1999

Les tests sur la table de passage îlot 1900-îlot 1999 qui a été mentionnée dans la partie 1c ont permis de détecter une erreur dans la variable îlot 1999, due à une mauvaise gestion des espaces dans l'identifiant, probablement lors de la rénovation de l'EDP aux débuts des années 2010 (étude sur la mobilité des immigrés, Paris 1).

E.3 ANNEXE 3. RESUME PUBLIC MIS A JOUR

Des données statistiques massives pour observer une société mobile

http://www.agence-nationale-recherche.fr/Projet-ANR-16-CE41-0007
(DS0806) 2016
Projet Big_Stat

Les nouveaux comportements conjugaux induisent une augmentation des mobilités individuelles familiales, ce qui rend plus difficile une description simple des situations familiales et résidentielles qui tienne compte de leur complexité. Dans le même temps, les données statistiques massives issues des fichiers administratifs exhaustifs deviennent accessibles en France pour la recherche. Le projet a pour ambition de renouveler la connaissance sur des situations familiales particulières, difficiles à observer, en tirant bénéfice de sources statistiques diverses, incluant des fichiers issus des données administratives massives, mais aussi d'évaluer scientifiquement les forces et les faiblesses des différentes sources démographiques qui ont été récemment mises à disposition de la communauté scientifique, ou le seront prochainement, par l'Institut national de la statistique et des études économiques (Insee).

Les données nécessaires pour les analyses démographiques complexes, comme l'Échantillon démographique permanent qui regroupe des données issues des recensements et de l'État civil, et a été récemment enrichi avec des données sociales et fiscales, sont maintenant disponibles. Ces données n'ont pas encore été beaucoup utilisées pour des études démographiques. Notre ambition est, dans un premier temps, d'évaluer la qualité des fichiers massifs mis à disposition et de les documenter pour les utilisateurs, en collaboration avec les personnes en charge de ces données à l'Insee, à partir d'un diagnostic partagé sur les fichiers. Nos efforts se concentreront d'abord sur les estimations de population au recensement, à partir d'une estimation inédite des doubles comptes et des omissions, ainsi que sur les représentations des structures familiales qui en découlent. Deux situations familiales particulières seront ensuite analysées à partir d'analyses diverses fondées sur des sources très hétérogènes : données administratives, recensement, enquêtes en population générale, entretiens non directifs. Tout d'abord, les relations conjugales des jeunes adultes ne sont pas toujours clairement définies et stabilisées. Les différentes définitions de la vie en couple seront mises en regard des conditions sociales et professionnelles des jeunes adultes. Ensuite, nous examinerons la situation familiale des enfants de parents séparés, particulièrement au risque de doubles comptes dans les enquêtes et le recensement. Les données administratives permettent de décrire la situation familiale des enfants et des jeunes adultes telle qu'elle est déclarée à l'État et d'analyser les conditions de vie en termes économiques, y compris pour les enfants partageant leur temps entre deux logements parentaux.

Le projet se place dans des perspectives nationale et internationale. Des contacts existent déjà avec des institutions étrangères qui utilisent et documentent des données administratives massives. Nous tirerons bénéfice de leur expérience et serons en mesure de disposer de notre propre expertise nationale, fondée sur une collaboration étroite entre nos instituts. En rendant accessibles sur un site Internet les informations sur l'accès et la qualité des données massives, et en proposant des solutions concrètes aux difficultés identifiées, le projet rendra un grand service à la communauté des chercheurs, tout en contribuant à l'amélioration de ces données. La publication d'articles méthodologiques dans des revues internationales de premier rang garantira la diffusion de nos résultats. Les avancées du projet, des informations sur des projets similaires à l'étranger et sur les modalités d'accès aux sources de données sont décrites sur le site du projet, www.big-stat.site.ined.fr, et celui des utilisateurs de l'échantillon démographique permanent, www.utiledp.site.ined.fr.

Les participants au projet travaillent à l'Institut national d'études démographiques (Ined), à l'Insee et dans les universités de Paris 1 Panthéon Sorbonne, Paris Descartes, Nanterre, Bordeaux et Strasbourg. Le projet permettra de développer et de faciliter l'utilisation par les chercheurs en sciences humaines et sociales des données massives bientôt disponibles.

Partenaires
INED Institut National d'Etudes Démographiques
Aide de l'ANR 291 585 euros
Début et durée du projet scientifique mars 2017 - 48 mois
Programme ANR : (DS0806) 2016
Référence projet : ANR-16-CE41-0007
Coordinateur du projet :
Monsieur Laurent Toulemon (Institut National d'Etudes Démographiques)

Big Statistical Data for a Mobile Society

New family and demographic behaviour are leading to greater individual and social mobility, making it more difficult to define and observe actual family and housing situations. Simultaneously, big statistical data, i.e. data from administrative files covering the whole population, are now becoming available to the research community. The project aims to extend our knowledge of complex and hard-to-measure situations, using several data sources including big data, and to assess the strengths and weaknesses of several data sources that will be disseminated in 2016 by the French National Institute of Statistics and Economic Studies (INSEE).

Data necessary for complex demographic studies, such as the French Demographic Panel based on censuses and civil registration, tax data and family allowance data, are now becoming widely available. So far, they have been rarely used for research purposes in demography. Therefore, we propose, in a first and crucial stage, to assess and document for general use the big data sources recently made available for research, in collaboration with INSEE. This collaboration is key to the constitution of reliable and well documented data sources. The knowledge from experts from several backgrounds and institutions will be essential to fully validate and test these data sources for various uses. In this step, we will check the consistency of population estimates based on censuses, surveys and administrative data, in terms of omissions and double counts, and the impact of discrepancies on the estimation of family situations and behaviours. Two research questions, which are normally difficult to evaluate with standard surveys, will then be addressed, making use of diverse methods and sources: administrative data, censuses, population surveys and qualitative data from semi-structured in-depth interviews. First, intimate relationships at young adult ages are known for their volatility, and are therefore hard to study with standard survey data. The new data sources will make it possible to look at forms of partnership and union stability in relation to income, education, occupation and labour market integration. This will vastly increase our knowledge of the dynamics of early adulthood and will further our understanding of new forms of partnership. Second, we will look more closely at the situation of children whose parents are separated, and who are a major source of double counting in surveys and censuses. New data sources will provide a more accurate picture of the family situation of children, including those in complex living arrangements, in relation to their standards of living

and poverty risk. Administrative data are very useful for studying transitions, while retrospective surveys are often complicated by recall bias and panel studies are weakened by attrition.

This project will be placed in a national and international perspective. It will provide an opportunity to combine the strengths of national institutions, while creating links with institutions abroad involved in the analysis of big administrative and census data. We will benefit from their experience and interact with big data networks to improve the quality and efficiency of our assessments and studies.

By making information, documentation and code for data use available on a website, this project will have a significant impact on the scientific community. It will also contribute to the enhancement of data quality and access. The publication of methodological and applied articles in internationally reputed journals will promote the dissemination of the project's progress and findings. More information on project outputs, on similar projects in France and in other countries, and on data access can be found on the project website, www.big-stat.site.ined.fr/en, as well as on the French Demographic Panel users website, www.utiledp.site.ined.fr/en.

Members of the project come from the French Institute for Demographic Studies (INED), INSEE, and the universities of Paris 1 Panthéon-Sorbonne, Paris Descartes, Nanterre, Bordeaux and Strasbourg. A major output of the project will be to encourage and facilitate the use of big statistical data amongst scholars working in the humanities and social sciences.

Partners
INED Institut National d'Etudes Démographiques
ANR grant: 291 585 euros
Beginning and duration: mars 2017 - 48 mois
ANR Programme: (DS0806) 2016
Project ID: ANR-16-CE41-0007
Project coordinator:
Monsieur Laurent Toulemon (Institut National d'Etudes Démographiques)
The project coordinator is the author of this abstract and is therefore responsible for the content of the summary. The ANR disclaims all responsibility in connection with its content.