Bernard AUBRY
Ancien chargé de mission INSEE-Alsace
baubry.stg@orange.fr

21 juin 2011

Comparaison de deux bases de microdonnées pour la France : IPUMS et SAPHIR

Rapport rédigé à la demande de Jean-Paul SARDON, directeur de l'Observatoire Européen (ODE) pour le compte de l'Agence Nationale pour la Recherche dans le cadre du projet CENSUS : «Evaluer et valoriser la base européenne de micro-données de recensement »

Ce rapport[1] compare deux bases de microdonnées issues des recensements de la population. La première base, IPUMS (Integrated Public Microdata Use Series[2]) est le fruit d'une collaboration entre le « *Minnesota Population Center*» et les instituts nationaux de statistique. Elle intègre actuellement des données concernant 62 pays, dont la France. La seconde base, SAPHIR (Système d'Analyse de la Population par l'Historique des Recensements) a été réalisée à l'initiative de la Direction Régionale de l'Insee-Alsace. Cette dernière base ne concerne que les seuls recensements français. Dans les deux cas, les objectifs, les principes de constitution ainsi que les publics visés sont très différents.

Au moment où les démographes européens s'engagent dans une opération similaire, il est utile de réunir les expériences et de mettre à la disposition des chercheurs une base de données d'une qualité optimale, sachant que les difficultés d'harmonisation des données, déjà sérieuses pour un pays donné, le sont bien plus encore quand il y a confrontation de systèmes statistiques nationaux ayant chacun sa propre histoire.

---

[1] *Le rapport reprend, en la développant, la communication présentée lors du séminaire « Use of Census Data in Europe in a Comparative Perspective » - Barcelone 9-11 juin 2011 organisé par le Centre d'Estudis Demografics – dans le cadre du projet Census (en collaboration avec l'Institut d'Etudes Démographiques de l'Université de Bordeaux (IEDUB), l'Institut National d'Etudes Démographiques (INED) et l'Observatoire Démographique Européen (ODE). L'auteur du rapport est le concepteur du fichier SAPHIR. Ce rapport tente de décrire les fichiers en prenant en compte notamment les remarques faites par les chercheurs qui ont travaillé simultanément sur les deux fichiers. Qu'il me soit permis de remercier tout particulièrement Khalid ELJIM pour m'avoir fait profiter de l'expérience qu'il a acquise dans le cadre de son travail de thèse de doctorat qui portait sur les immigrés originaires du Maghreb (Université de Bordeaux IV).*

[2] *IPUMS désigne l'organisme, sachant qu'en toute rigueur, on devrait parler de IPUMS $_{international,}$ pour le distinguer de IPUMS $_{USA.}$ Dans la suite du texte et du fait qu'il ne peut y avoir d'ambiguïté, on notera IPUMS pour désigner la partie internationale du projet, la seule qui nous concerne.*

PLAN

I – IPUMS – SAPHIR : deux fichiers de microdonnées

A - IPUMS – Un fichier de microdonnées censitaires à l'ambition mondiale

IPUMS-international[3] est un projet visant à collecter, harmoniser et diffuser des échantillons de microdonnées de recensements anonymisées de tous les pays du monde, pour qu'elles soient utilisées par les chercheurs. L'objectif d'IPUMS est de permettre d'effectuer des comparaisons entre les pays (ce qui n'existe pas par ailleurs) à partir de micro données, grâce à des variables harmonisées (« integrated » - qui permettent la comparaison entre les pays), tout en disposant d'informations explicites sur les métadonnées et d'une documentation des recensements complète. Par ailleurs IPUMS met également à disposition des chercheurs les données non harmonisées (variables brutes). Les microdonnées de recensements disponibles sur IPUMS concernent tous les continents, avec la mise à disposition des échantillons de microdonnées de recensements de 62 pays : 185 échantillons, 400 millions d'enregistrements (juin 2011) remontant parfois jusqu'au début des années soixante.
Pour en savoir davantage sur le contenu et la diffusion des données, on pourra consulter le lien suivant : https://international.ipums.org/international/

L'accès aux données est limité aux chercheurs. L'inscription au site passe par la rédaction d'un texte de motivation. La prise en main est rapide pour qui est familier des téléchargements des bases de données volumineuses. Le fichier est en format ASCII, mais pour le lire, l'utilisateur a le choix entre trois programmes : SAS, SPSS et STATA qui lui sont proposés automatiquement.

[3] *Rappelons que IPUMS désigne l'organisme, sachant que IPUMS$_{international}$ (IPUMS.I) et IPUMS.USA sont les noms donnés aux bases de micro-données.*
Avec un financement issu essentiellement du National Institut of Health (NIH) des Etats-Unis, l'harmonisation et la documentation complète des données sont organisées par Minnesota Population Center, en parallèle avec des activités similaires portant sur d'autres parties du monde avec le financement de la National Science Foundation.

B - SAPHIR : un fichier historique pour les recensements français

SAPHIR est un fichier interne à l'INSEE, mis au point à la Direction régionale de l'INSEE-Alsace. Son origine[4] remonte à la fin 1989 et les premières études nationales ont été publiées dans le courant de l'année 1992. Dans sa version la plus récente, il contient la quasi-totalité des informations collectées lors des six premiers recensements informatisés (1962-68-75-82-90-99), c'est-à-dire y compris les informations géographiques les plus fines (n° de quartier ou de l'Iris[5], de l'immeuble et de logement). Parmi toutes ces variables, un certain nombre ont été harmonisées et les résultats ont été largement testés. D'autres n'ont pas raison de l'être (notamment les variables collectées lors d'un seul recensement) ou ne peuvent l'être de façon systématique (l'activité économique dans une nomenclature détaillée, ou les données d'immeuble et de logement). Cependant des investissements spécifiques réalisés en collaboration avec des chercheurs ont permis d'utiliser le fichier pour des travaux s'appuyant sur les données localisées les plus fines (c'est le cas notamment pour les études sur le voisinage ou sur les grands ensembles de Strasbourg[6]).

L'investissement s'est fait sur le long terme, selon des principes tels que la simplicité, la souplesse, la transparence, le repérage des anomalies, mais aussi l'objectif de pouvoir répondre à l'attente des utilisateurs (les « clients » de l'INSEE) : le monde des élus, des aménageurs, des enseignants et des chercheurs. L'intérêt pour la connaissance des dynamiques territoriales était affiché de sorte qu'une harmonisation sur toute la période des codes de communes a été effectuée (0,4% de la population n'a cependant pas été réintégrée en 1962 et 1968, ce pourcentage étant moindre encore ultérieurement).

L'ambition du projet était de créer non seulement un fichier de données, mais un « système », d'où le nom qui a été donné à l'investissement (SAPHIR : *Système d'Analyse de la Population par l'Historique des Recensements*). La création d'un fichier famille harmonisé appliquée à la connaissance de la concentration des populations issues de l'immigration a permis la rédaction de plusieurs études sur la question largement reprise dans le débat public[7].

L'acronyme SAPHIR a été abandonné par l'INSEE. Toutefois le fichier a été mis à jour par l'intégration des résultats du RP2006. Il peut dont être utilisé par les agents de l'INSEE. Une extraction, amputée du RP62 et de certaines variables utiles à l'analyse sociologique, est disponible en ligne sur le site de l'INSEE : *« Données harmonisées des recensements de la population de 1968 à 2006* ».

Présentation synoptique des thèmes

Le tableau 1 se propose de faire apparaître les principaux domaines pour lesquels un suivi historique de données issues des recensements est possible pour la période 1962-1999. L'accent est mis sur les catégories de Saphir (caractères gras), contrairement aux fiches thématiques (annexe 1) qui présentent une liste exhaustive des variables contenues dans IPMUS. La liste des variables n'est pas exhaustive.

---

[4] On trouvera plus loin davantage d'informations sur la méthode de travail qui a conduit à la création de SAPHIR

[5] IRIS, maille de base pour la diffusion des données infra-communales

[6] Collaboration avec Michèle TRIBALAT (INED) sur le voisinage, et avec Bénédicte GERARD sur les grands ensembles. A souligner également les investissements originaux proposés par Christophe BOURGOUIGNAN (IEDUB, Bordeaux IV, rencontres AURBA-IEDUB 2007) autour des projections de population : « De la prospective démographique à l'appréhension des besoins en logement ».

[7] Le premier texte rédigé sur cette question a été publié d'abord en italien par la revue Aspenia, puis par la revue Commentaire : « Les jeunes d'origine étrangère » – Bernard AUBRY et Michèle TRIBALAT – n° 26, été 2009, pages 431 à 437.

Tableau 1 – tableau synoptique des thèmes

thème	SAPHIR	IPUMS
Logement – *Dwelling*		
Année d'achèvement	**AA**	BLTYR
Ascenseur	**ASC**	ELEVATR
Salle de bains	**BD**	BATH
Chauffage -combustible	**CC – CB**	FUELH
Nombre d'étages de l'immeuble	**ET**	STORIES
Nombre de pièces	**NP**	ROOMS
Statut d'occupation	**SO**	OWNRSHPD
Nombre de voitures	**V**	AUTOS
Autres	**HLM - NL –TL – WC**	KITCHEN-SEWAGE-ELECTR PHONE
Ménage-Famille *- Household – family*		
Identification du ménage	**NUM**	SERIAL
Ménage	**TMEN – RGM**	PERSONS-PERNUM
Famille	**TFAM – NF**	FAMSIZE-FAMUNIT
Individu-structure		
Sexe	**S**	SEX
Année de naissance	**ANN**	BIRTHYR
Formation	**DIP**	EDUCFR
nationalité	**N**	CITIZEN-NATIVTY-NATION
Autres	**-**	SCHOOL-EDAGE-EDATTAN
Individu-liens - *links*		
Avec la personne de réf. du ménage	**LCM**	RELATE
Avec la personne de réf. de la famille	**LCF**	PARRULE-SPRULE
Etat matrimonial	**M**	MARST
Autres	**COHA**	MOMLOC...STEPPOP…
Activité-emploi – *Work*		
Type d'activité	**TA**	EEMPSTAT
statut	**ST**	EMPSECT-EEMPSTAT
Catégorie sociale	**CS**	OCCISCO - ECLASSWKD
Activité économique	**AE**	INDGEN
Autres	**-**	LOOKJOB-HRSFULL-EMPLNO-EMPSTAT - TRNWRK
Géographie - migrations		
Résidence	**RR - DCR**	REGNFR
Lieu de naissance	**PN - RN -DN**	BPLFR
Lieu de résidence antérieure	**PRA – RRA - DCRA**	MGCTRY4-MIGFR
Lieu de travail	**PLT - RLT - DCLT**	PWFKFR
Autres	**-**	URBAN-MGRATEC-MGHOUFR

C - IPUMS et SAPHIR - Comparaison des deux fichiers

Les deux investissements sont nés en partie d'objectifs communs, mais on relève un certain nombre de différences dont la principale est liée aux contraintes tenant au caractère international de l'IPUMS. On trouvera plus loin (tableau 6) une présentation synthétique des caractéristiques comparées des deux fichiers.

Vouloir rendre homogènes des fichiers d'un même pays est évidemment une tâche moins ambitieuse que de chercher à fusionner des fichiers issus de pays différents, puisque chaque pays a ses propres méthodes, même si, sous la pression des institutions internationales, les méthodes tendent à s'harmoniser. De son côté, SAPHIR a été rendu possible par la mobilisation d'agents qui, parce qu'ils avaient une longue expérience, une « culture » des pratiques de leur institution, étaient en mesure d'appréhender spontanément les domaines traités, leurs difficultés, leurs pièges ; certains d'entre eux avaient encore en mémoire les expériences des premiers recensements inclus dans la série.

Les deux fichiers sont composés de micro-données (données individuelles, mais anonymes). L'identification des ménages et des familles permet de reconstituer de nouveaux fichiers en aval. L'avantage est que les utilisateurs sont libérés des tâches fastidieuses de recherche de données et que tous les croisements de variables sont a priori possibles.

C 1 - fichier IPUMS

La principale critique que l'on peut faire sans doute au fichier IPUMS est de n'avoir pas fait l'objet d'une expertise suffisante. On ne trouverait pas d'incohérences comme celles qui concernent les séries régionales ou les séries de nationalités. En effet, le tableau 2 montre comment la population de la région PACA serait passée de 3,7 à 2,5 millions d'habitants en 7 ans, de 1968 à 1975, ce qui n'est pas vraisemblable. Le cas de la Corse est différent puisque, s'agissant de cette région, les données originales sont erronées ; c'est d'ailleurs pourquoi il est suggéré aux utilisateurs de SAPHIR de ne pas prendre en compte la Corse s'ils veulent travailler sur des séries cohérentes avant 1982.

Tableau 2 – Fichier IPUMS, population totale de la Métropole de l'Ile-de-France et de trois régions méridionales

POPULATION (milliers)						Région	Variations intercensitaires			
1962	1968	1975	1982	1990	1999		68-75	75-82	82-90	90-99
46418	**49756**	**52589**	**52634**	**56660**	**58695**	**Métropole**	**2834**	**45**	**4026**	**2035**
8454	9261	9879	10065	10663	10990	Ile -de-France	617	186	598	328
1538	1697	1789	1930	2117	2309	Languedoc-Rouss.	92	141	188	191
2831	3307	**3677**	**2539**	**4259**	4526	Provence-Alpes-CA	**370**	**-1138**	**1720**	**267**
277	**292**	**219**		**250**	262	Corse	-73			12

Il reste qu'une fois soulignées ces anomalies, qui devraient pouvoir être facilement corrigées, on peut proposer un bref résumé des avantages et des inconvénients des deux fichiers.

Tableau 3 – Avantages et inconvénients respectifs des fichiers IPUMS et SAPHIR

IPUMS	**Avantages**	• Facilité d'accès aux données • Comparaisons entre pays. • Documentation en ligne – accès aux questionnaires
	Inconvénients	• Très grand nombre de variables proposées, pour tenir compte de situations différentes selon les cultures. Le surnombre de variables nuit à la limpidité (ex : les liens familiaux), • Ruptures de séries, absence de données a priori injustifiées (ex : l'année de naissance en 1999) • Harmonisations perfectibles (ex : année de construction de l'immeuble) • Unité géographique inférieure : la région, d'où impossibilité d'effectuer des totalisations sur des zones quelconques, fussent-elles plus peuplées qu'une petite région comme la Corse, ce qui est un handicap pour nombre de travaux de recherche.
SAPHIR	**Avantages**	• Simplicité, souplesse, caractère évolutif • Souci de faire apparaître les anomalies • Utilisations ciblées (familles, immigrés, aménagement du territoire) • Niveau géographique communal, voire infra communal, permettant tous les regroupements territoriaux. - Possibilité d'études sur le voisinage.
	Inconvénients	• Accès limité, sauf exception, aux chargés d'étude de l'Insee (pour le fichier détaillé) • Suivi insuffisant par l'institution qui n'a pas engagé les investissements nécessaires pour valoriser l'acquis (redressements élaborés, outils de diffusion spécifiques …)

On devrait pouvoir consulter en ligne un fichier (Excel) comparant les comptages obtenus par sommation de toutes les modalités de toutes les variables du fichier IPUMS et du fichier SAPHIR.

En prenant pour exemple, la date, ou la période, d'achèvement de l'immeuble (BLTYR pour IPUMS et AA pour SAPHIR), et en se limitant la sommation pour l'Alsace, le tableau 4 met en évidence les différences telles qu'elles apparaissent en interrogeant simultanément les fichiers IPUMS et SAPHIR[8].

[8] *On aurait pu souhaiter que le fichier IPUMS et le fichier SAPHIR soient identiques pour leur partie commune. Cela aurait été d'autant plus facile à réaliser qu'au début des années 2000, un représentant d'IPUMS, Michael RENDALL,*

Tableau 4 – Comparaison IPUMS *vs* SAPHIR - Année d'achèvement de l'Immeuble

Année/période d'achèvement	Population des logements selon l'année de collecte (pour l'Alsace, en milliers)				
	1968	**1975**	**1982**	**1990**	**1999**
I P U M S					
-				1589	1697
0	2				
1870	293	239	181		
1914	268	222	184		
1939	259	261	238		
1948	34				
1949					
1951					
1953	75				
1957	103				
1961	128	271	232		
1962					
1964	85				
1967		189	158		
1968	120				
1974			278		
1975		292			
1982			243		
9998	3				

S A P H I R					
< 1915	560	462	369	307	285
1915-48	297	263	244	236	228
1949-67	509	456	393	331	322
1968-74		290	273	255	228
1975-81			245	234	189
1982-89				215	209
1990-99					229
nd	2			5	3

(Les parties grisées permettent de faire apparaître les ensembles communs aux deux fichiers)

On constate qu'il n'y a pas de divergences importantes entre les deux fichiers, mais on se propose de montrer à travers ce tableau comment l'harmonisation a été pratiquée dans un cas et dans l'autre.

Saphir présente des périodes de construction, IPUMS donne des années, ce qui oblige l'utilisateur à reconstituer des classes spécifiques à ces besoins. Il y a donc pour l'utilisateur un meilleur confort de lecture. La plus grande précision apparente d'IPUMS n'apporte guère d'information utile. En revanche, s'agissant de SAPHIR, d'avoir regroupé toutes les années avant 1915 conduit à une légère perte d'information (ce choix a été fait du fait que les deux périodes les plus anciennes n'étaient pas disponibles aux derniers recensements).

était venu en mission à Strasbourg pour s'informer des possibilités offertes par SAPHIR. Par ailleurs on aurait pu demander à la DR d'INSEE-Alsace de vérifier la conformité à SAPHIR avant la mise en ligne des séries par l'IPUMS.

De fait le fichier IPUMS ne donne aucune information sur les recensements 1990 et 1999. Il y aurait lieu d'approfondir la question de cette absence et d'examiner les avantages et les inconvénients à ne pas retenir cette information, pourtant existante. Enfin, parce que les chiffres concernés sont mineurs, il aurait été opportun d'intégrer par affectation automatique la série dite fictive (n.d.), sans intérêt pour l'analyse (ceci est valable pour les deux fichiers).

La nomenclature des pays, comme illustration des difficultés d'harmonisation.

Les recensements français font intervenir une nomenclature des pays dans trois cas : la nationalité, le pays de naissance et le pays de résidence. D'un recensement à l'autre les nomenclatures ont changé. On a tenté dans SAPHIR de créer une nomenclature unique comprenant 75 postes. De son côté IPUMS propose une nomenclature en 27 postes (dont 19 pays) pour la seule variable nationalité (à l'exception de l'année 1999 où aucune donnée n'apparaît).

Les différences entre les deux fichiers de micro-données sont parfois très importantes. Les sommations pour la métropole apparaissent sur les deux tableaux suivants.

Tableau 5a – Effectifs par nationalité IPUMS 1962-1990

(Métropole, en milliers)

nationalité		1962	1968	1975	1982	1990	1999
.							58695
13010	Algérie	334	471	711	769	728	
13030	Libye					33	
13040	Maroc	29	88	254	401	649	
13060	Tunisie	25	60	136	175	264	
19990	Afrique*	19	33	81	156	312	
29900	Amérique*					96	
33020	Cambodge		1	4	37		
33050	Laos		1	1	33		
33110	Vietnam	7	10	11	34		
33999	Asie Sud-Est*					186	
34160	Turquie	16	17	60	128	207	
39900	Asie*	12	16	28	59	144	
41050	Pologne	177	131	96	64	211	
41080	Russie/URSS					23	
41999	Europe de l'est*					44	
43070	Italie	645	586	463	317	683	
43090	Portugal	50	303	751	758	804	
43120	Espagne	431	618	503	306	517	
43140	Yougoslavie					84	
44020	Belgique	78	67	55	48	109	
44030	France	44268	47092	49165	49083	51270	
44040	Allemagne	46	46	41	42	111	
44060	Luxembourg	6	4	3	3		
44090	Suisse	33	32	26	21		
49990	Europe*	129	148	159	151	183	
59999	Océanie*					3	
70000	Autres pays	113	32	44	51		

**autres pays du continent*

Tableau 5 b - Effectifs par nationalité, résidence antérieure et naissance- SAPHIR (1968-1999)

(Métropole, en milliers)

PAYS	NATIONALITE					PAYS DE RESIDENCE ANTERIEURE					PAYS DE NAISSANCE				
	nat68	nat75	nat82	nat90	nat99	pra68	pra75	pra82	pra90	pra99	pn68	pn75	pn82	pn90	pn99
Eu. Ouest*	0					7					5				
000	44559	46633	48110	50226	51797	46120	49652	51763	54215	55936	43542	45778	47266	49422	51225
001*	1294	1371	1408	1755	2310	51	111	157	138	149	83	153	266	322	342
Autriche	3	4	3	3	4	2	3	2	3	3	14	14	12	12	12
Belgique	62	52	48	54	64	18	29	33	42	43	145	130	123	122	119
Suisse	28	26	21	19	27	15	20	19	23	39	72	69	30	64	74
Allemagne	41	39	41	49	73	66	68	65	76	96	194	194	203	200	206
Danemark	1	1	2	3	4	1	1	2	2	3	2	3	4	5	6
Espagne	585	478	323	208	156	230	78	26	20	26	669	594	494	409	334
Irlande	1	1	2	3	5	0	1	1	3	2	1	1	2	3	5
G.Bretagne	16	23	32	47	72	12	22	29	40	63	22	31	41	59	81
Grèce	9	9	7	5	5	2	4	4	5	4	13	15	13	12	11
Italie	563	458	330	253	196	69	37	26	22	27	781	686	596	507	394
Luxemb.	3	3	3	3	3	2	4	4	4	5	12	11	10	10	8
Monaco		1	1	0	1		2	3	3	4		13	14	15	14
Norvège	1	1	1	2	2	1	1	1	2	2	1	3	2	3	2
Pays-Bas	11	10	11	17	24	4	6	7	10	15	12	13	16	21	27
Portugal	253	741	753	646	549	189	409	75	54	65	241	637	635	604	571
Suède	2	3	4	5	7	1	4	3	4	5	2	4	5	6	8
Finlande	1	1	1	1	3	1	1	1	1	2	1	2	2	2	3
Eur. Est*	11	8	3	2	6	3	1	1	2	7	15	3	3	2	12
Bulgarie	1	1	1	1	2	1	1	1	1	3	3	4	3	4	6
Tché-Slov.	5	3	2	2	2	1	1	1	1	3	17	14	14	12	11
Hongrie	8	5	3	2	3	1	2	1	1	2	16	15	13	12	10
Pologne	127	89	61	44	31	8	9	9	16	13	217	188	162	131	103
Roumanie	5	4	3	5	9	1	2	2	4	11	15	13	14	15	24
Russie/UR	19	11	6	4	11	2	3	2	3	14	40	33	26	20	25
Yougosl	44	66	62	52	50	23	29	6	6	17	50	71	69	69	77
Algérie	399	609	725	568	435	1056	164	128	74	114	1287	1340	1357	1297	1189
Maroc	66	207	406	536	484	127	161	142	115	91	241	378	517	614	689
Tunisie	53	125	178	200	142	60	75	44	29	21	245	322	342	347	327
Bénin	1	2	3	5	4	2	2	4	3	3	2	4	7	10	11
Burk-Faso	0	1	1	1	2	2	2	1	3	2	2	3	3	3	5
Centre-Afr.	0	1	1	4	3	3	3	2	2	3	2	4	3	7	8
Congo	1		8	11	35	8		8	9	15	5		5	19	42
Côte d'Iv.	1	6	11	15	20	9	16	24	31	23	5	13	20	32	44
Cameroun	2	7	12	15	19	6	9	11	16	14	6	12	17	23	35
Comores		3	1	3	5		5	3	5	7		8	14	10	16
Gabon	0	2	3	2	4	4	6	10	11	10	3	4	7	6	10
Guinée	1	1	2	6	9	2	2	1	3	2	3	5	7	10	7
Madagasc.	1	3	5	8	9	24	23	16	12	14	24	38	49	61	68
Mali	1	5	14	29	24	2	3	6	9	7	3	6	13	23	28
Mauritanie	1	3	3	4	6	3	5	3	3	3	2	4	3	4	8
Niger	0	0	1	1	0	3	3	2	3	2	1	2	2	3	3
Sénégal	3	9	24	37	31	23	20	20	23	21	19	27	43	62	73
Tchad	0	1	1	1	1	5	4	4	1	1	2	3	4	4	4
Togo	1	3	4	6	6	2	3	3	5	5	2	4	5	9	13

Afrique*	2	7	5	18	28	7	14	20	25	25	7	16	18	36	56
Egypte	2	2	4	5	8	2	2	3	4	4	17	16	18	19	21
Maurice			12	11	14			7	8	5			17	24	29
Nigéria			1	1	1			2	4	2			2	2	2
Afri. Sud			5	20	16			7	18	10			8	22	24
Cambodge	0	4	33	48	24	3	6	28	6	2	3	9	41	59	55
Laos	0	1	32	32	17	2	3	33	4	1	2	4	36	44	40
Vietnam	9	10	31	30	20	6	7	37	16	7	51	56	93	109	111
Asie*-Océ*	1	4	11	25	46	2	6	17	42	39	4	11	20	48	80
Australie	1	1	1	2	2	1	2	2	3	4	1	2	3	4	4
Inde	0	1	2	4	6	1	3	6	7	9	3	5	11	20	27
Irak			1	2	3			1	3	2			2	2	4
Iran	2	3	10	15	11	1	3	11	8	3	2	4	12	18	19
Israël	2	4	3	3	2	3	4	7	6	3	3	4	7	8	8
Japon	1	4	6	10	13	2	3	5	8	11	2	4	6	11	14
Pakistan			3	9	11			2	5	4			3	10	12
Chine	2	2	5	14	28	1	1	2	8	14	4	5	8	18	32
Liban	2	2	9	18	10	2	3	12	17	8	6	8	17	34	33
Syrie	1	2	4	7	3	1	2	3	5	3	5	6	9	13	12
Turquie	6	36	119	188	206	3	30	55	53	39	43	64	121	162	176
Canada	2	4	4	6	8	5	14	11	11	15	5	10	12	14	17
Etats-Unis	15	21	17	22	24	17	24	16	30	39	20	27	23	31	38
Amér.lat*	3	5	9	16	23	3	7	10	15	12	5	10	16	26	41
Brésil	2	3	4	6	7	4	4	5	8	9	4	6	7	13	18
Colombie			1	3	5			1	3	4			2	7	13
Chili	0	2	6	9	3	1	3	5	6	2	1	3	8	13	11
Mexique	1	1	1	2	2	1	2	2	4	4	2	2	3	4	5
Argentine	1	2	4	3	2	2	3	5	3	2	7	7	10	9	9
Vénézuela	1	1	2	1	1	1	1	2	3	2	1	2	2	2	3

**reste du continent - ** absence d'information en 1975, due à une erreur d'affectation (corrigée dans un fichier ultérieur)*
Code 000 : il s'agit des Français de naissance ou des personnes nées ou résidant antérieurement en métropole - code 001 : par convention pour la nationalité, il s'agit des personnes devenues françaises, dans les autres cas, il s'agit des personnes nées ou résidant antérieurement dans les DOM-TOM

L'examen des tableaux appelle plusieurs remarques :

a – **La frilosité**

On connaît la frilosité qui caractérise la France lorsqu'il s'agit de la question des échanges de population avec l'étranger : l'IPUMS propose une nomenclature des pays en 318 postes, cependant que la partie française du fichier n'en contient que 19 (avec en sus 8 lignes correspondant à des reliquats par continent). Par ailleurs, il manque en particulier des données sur le pays de naissance des personnes, qui auraient permis d'effectuer davantage d'études sur l'immigration. Il semble que les règles de diffusion en la matière se soient récemment très nettement assouplies et l'on peut espérer qu'à terme, les séries proposées seront plus riches et plus cohérentes.

La rétention d'information pour la France est d'autant plus regrettable que les séries détaillées issues des recensements sont particulièrement cohérentes[9]. C'est la confirmation de la bonne facture des recensements. Il est donc bien dommage de priver d'une connaissance de qualité la communauté des chercheurs, et ce d'autant plus que l'absence de repères conduit parfois à la circulation de chiffres fantaisistes dans le cas d'un domaine où le débat public est difficile.

b **– certaines incohérences**

On note des variations d'effectifs extraordinaires : le nombre d'Italiens serait ainsi passé de 317 000 à 683 000 de 1982 à 1990. Certains pays peu représentés (La Lybie, le Laos…) n'apparaissent qu'épisodiquement.

La documentation en ligne fournie par IPUMS fait allusion à la nomenclature SAPHIR. Celle-ci aurait été utilisée pour les RP 68 à 82. De fait les écarts entre les deux fichiers sont relativement moindres pour ces trois collectes.

c – la difficulté d'harmonisation

Les nomenclatures des pays illustrent bien la difficulté du travail d'harmonisation temporelle. L'inconstance des frontières et des pouvoirs politiques (décolonisation) amènent à faire des conventions : ex la Yougoslavie, l'ex URSS, la Chine, etc. Intégrer un nouveau pays, le rayer de la série, c'est a priori rendre impossible l'établissement de séries cohérentes, sauf à établir des données fictives pour composer des reliquats (par continent) cohérents.

[9] *C'était l'objet d'une communication au colloque de l'AIDELF (Association Internationale des Démographes de Langue Française - (Bernard AUBRY – les immigrés en France 1962-1999 - Budapest – 2004) où précisément a été montrée la bonne cohérence des collectes successives (bilans par génération).*

C 3 - Les caractéristiques comparées : tableau de synthèse

Afin de disposer d'une vision synthétique des deux fichiers, on présente ci-après un tableau donnant les principales caractéristiques des bases IPUMS et SAPHIR.

Tableau 6 – Caractéristiques comparées IPUMS – SAPHIR

	IPUMS	SAPHIR
Objectifs	• *Base de données individuelles*	• *Bases de données* • *système d'analyse*
Géographie	• *62 pays*	• *France métropolitaine*
Années d'observation	• *1962-2006*	• *1962-2006*
Principes	• *Fournir une base aux chercheurs*	• *Accessibilité aux données* • *Transparence des résultats* • *Mise en évidence des anomalies* • *Améliorer l'offre de diffusion de l'Insee*
Echantillon	• *1/20ème ou 1/25ème (RP90)*	• *1/ 4 ou 1/5ème (RP75)*
Variables	• *Variables brutes* • *Variables harmonisées*	• *Variables brutes* *variables harmonisées (sauf pour l'activité économique et la profession)*
Accès aux données	• *Aisé et élaborée* • *Documentation en ligne*	• *Très aisé pour les agents connaissant SAS)*
Diffusion	• *limitée aux chercheurs*	• *Produit interne à l'Insee, les informations disponibles en ligne ne permettent pas de mettre à profit les spécificités du produit.*
Applications en aval	• *Limitées (impossibilité de sélectionner les immigrés par ex.)*	• *Fichier famille et fichier ménage* *Fusions avec d'autres fichiers (ex : enq. sur les forces de travail*

II - Réflexions sur la constitution d'un fichier européen

I - La base européenne de micro-données de recensement intégrées (IECM)

Le Centre d'Estudis Demografics (Barcelone) a obtenu le soutien du sixième programme de travail de l'Union européenne pour l'amélioration, l'harmonisation et la diffusion des données européennes intégrées et des métadonnées, ainsi que pour coordonner les tâches basées en Europe. Ce programme prévoit la constitution d'une base européenne de micro-données (IECM) construite à partir des fichiers fournis par IPUMS (le sous-fichier Europe correspond à peu près au quart de l'ensemble de la base mondiale).

C'est dans ce contexte qu'est présentée ici l'expérience acquise avec la composition du fichier SAPHIR. Pour assurer au fichier européen une bonne fiabilité, un certain nombre de réflexions semblent s'imposer.

I - Le premier objectif est de créer la **confiance**. Il faut donc que les utilisateurs puissent quasi-immédiatement disposer de données sans avoir à s'interroger sur la qualité des informations. En effet, dès que le doute s'instaure, dès que des erreurs ou des anomalies graves sont repérées, c'est souvent, et parfois à tort, que *l'ensemble du travail* est remis en cause. Bien sûr, la qualité parfaite est un idéal que l'on ne saurait atteindre.

Dans l'optique d'une utilisation du fichier historique par les chercheurs, on peut envisager deux applications tout à fait différentes, sans exclure évidemment la constitution de sous-fichiers intermédiaires spécifiques.

La première c'est celle d'une **base de données brutes**, parfaitement documentée. Dans ce cas le chercheur est seulement libéré des problèmes liés à la quête des données. Il doit pouvoir accéder aux éléments facilitant son travail : questionnaire de la collecte, définitions, etc. La base IPUMS répond à cette attente à travers la mise en ligne des données brutes et de la documentation afférente.

Pour « rentabiliser » le travail que les utilisateurs pourraient effectuer sur ces données, pour éviter que les mêmes investissements longs et fastidieux ne soient reproduits à chaque fois, il serait souhaitable que les chercheurs soient invités à valoriser leur travail et par là-même qu'ils trouvent un intérêt (de notoriété) à pouvoir porter leurs résultats de recherche dans un recueil comme le rend possible aujourd'hui l'informatique (un « Wikipedia » ou une encyclopédie en quelque sorte). Donc une perspective de ***capitalisation des investissements*** qui devrait pouvoir profiter à tous.

La seconde est celle d'une **base de séries historiques de référence**. On imagine facilement certaines applications. C'est, par exemple, l'épidémiologiste qui pour des calculs de prévalence, a besoin de rapporter ses propres résultats à des dénominateurs de référence, la population par sexe et âge, voire par origine (pour la drépanocytose par exemple). C'est l'économiste qui souhaite pouvoir se référer à des indicateurs cohérents portant sur la population active (au lieu de résidence) et sur l'emploi (au lieu de travail).

Entre les deux extrêmes, des bases spécifiques pourraient être créées (pour certains domaines et certains groupes de pays)

II – Avancer par étapes et par thèmes, responsabiliser les acteurs

La constitution de séries historiques cohérentes ne se fait pas de façon aisée. Les difficultés sont déjà grandes au niveau d'un même pays, elles le sont plus encore entre pays voisins. Quand il s'agit de pays lointains, l'idée même d'harmonisation est illusoire. Une variable élémentaire aussi simple que l'âge pose même en soi parfois de nombreux problèmes de comparaison.

Dans une optique limitée à l'Europe, on peut sans doute aller assez loin dans l'harmonisation des données. On devrait découper le champ des données censitaires en plusieurs domaines relativement indépendants. Parmi les catégories suggérées : les caractéristiques du logement et de l'équipement, l'éducation et l'apprentissage, la qualification et l'emploi, les liens familiaux, les migrations intérieures et les migrations extérieures.

Chaque domaine serait étudié par un groupe de chercheurs spécialisés qui désignerait en son sein un responsable. Le concours actif des instituts nationaux de statistique (INS) aiderait les chercheurs de façon significative. En retour les INS apprendraient à mieux connaître le travail de leurs partenaires et les utilisations qui sont faites par les chercheurs des données qu'ils produisent. La collaboration se ferait de deux façons. D'une part à travers la fourniture au groupe de recherche de micro-données sur les seules variables nécessaires à l'investigation du domaine[10] et d'autre part par la désignation de correspondants permanents internes auxquels le responsable pourrait s'adresser en cas de nécessité.

Dans chaque domaine la progression se ferait par étape. Après consensus, on diffuserait chaque fichier d'un domaine donné en le caractérisant par un indicateur millésimé. Ainsi s'agissant de la formation, DIP.1 serait une première version d'un fichier historique. Il serait suivi plus tard par une autre DIP.2, et ainsi de suite. Les versions successives intégreraient des redressements de plus en plus élaborés. En cas d'aiguillage, plusieurs versions pourraient le cas échéant cohabiter. Une version faiblement redressée et une version grandement reconstruite. Un exemple : on sait que la qualité de la collecte pour certaines catégories, les jeunes, les étrangers, est médiocre (doubles comptes, mais plus souvent omissions). Sur la base de certaines hypothèses, avec toutefois des risques d'incohérences créées par volonté d'amélioration, on pourrait procéder à des redressements qui permettraient de construire une « démographie historique » cohérente.

III – Mettre en évidence les anomalies les plus graves

En premier lieu, il convient de ne focaliser l'attention que sur les anomalies les plus graves, sachant que toute opération aussi importante qu'un recensement de la population est entaché de nombreuses incertitudes d'origine très différentes. Les variations intercensitaires suspectes, celles qui sont susceptibles de conduire les utilisateurs à des interprétations erronées, seraient soulignées. Comme cela arrive parfois, on évitera l'accumulation de listes d'erreurs anodines[11].

---

[10] *Pour améliorer la qualité des redressements, on pourrait souhaiter par ailleurs la fourniture d'informations tirées d'autres sources telles que l'enquête européenne sur les forces de travail qui pour certains domaines permettent d'affiner les contenus des modalités et donnent ainsi des clés pour des redressements.*

[11] *On ne peut limiter le travail d'harmonisation à la seule prise en compte des consignes de codification. Un exemple sur l'activité des femmes vivant dans les zones rurales. Les séries des recensements des années soixante et soixante-dix montrent de fortes irrégularités à la fois spatiales et temporelles. C'est que certains cadres locaux ont interprété de façon personnelle les consignes reçues du niveau central. Tel ou tel avait « décidé » à tel ou tel*

IV – D'autres suggestions

Bien entendu il convient de rédiger une documentation appropriée, mêlant les informations de base (un minimum d'informations devrait être imprimées sous forme de brochures attrayantes) et des documents en ligne incluant des textes plus élaborés, des tableaux synoptiques et des renvois (hypertextes). Cette documentation doit être elle-même évolutive, prenant en compte les réactions des utilisateurs : nécessité de répondre aux remarques et d'intégrer les apports dans la base documentaire. Ces opérations relèvent d'une pratique maintenant classique de l'outil informatique.

On pourra s'orienter vers d'autres directions, comme :

- la réalisation de fichiers nationaux optimisant l'information disponible (pour la France, amélioration du fichier existant, enrichissement par des données provenant des autres enquêtes auprès des ménages : sur le logement, les forces de travail, etc.
- la réalisation de fichiers harmonisés pour quelques pays statistiquement proches.
- la présentation de tables de contrôles commentées, croisant plusieurs variables.
- la présentation de bilans intercensitaires. Ce peut être simplement des bilans d'effectifs par cohorte, pour connaître en première approximation la qualité relative de recensements successifs (calcul d'une émigration apparente). On peut imaginer aussi des bilans plus élaborés, comme l'INSEE en a proposé (*bilans des ressources humaines*) sur l'équilibre, par génération entre population totale, population active résidente et emploi (lieu de travail).

  Enfin, dans le prolongement de ce qui a été dit plus haut sur la capitalisation, on pourrait souhaiter la création d'une base documentaire qui ne serait pas seulement l'empilement des travaux et études réalisées (thèses de doctorat notamment, desquelles seraient extraites les quelques « perles » qui viendraient enrichir le collier).

---

recensement que les femmes présentes dans une ferme étaient actives par définition, même si elles n'avaient déclaré aucun établissement de travail.

ANNEXES

I – TABLEAUX DES VARIABLES IPUMS

Pour chacun des 6 domaines déjà décrits dans le tableau 1 (dans l'optique SAPHIR) on présente maintenant l'ensemble des variables renseignées pour la France dans le fichier IPUMS harmonisé (à l'exception des 9 variables portant sur le handicap qui n'étaient présentes qu'au RP62. On remarque que les quelque 85 variables recensées ne représentent qu'une faible partie de la longue liste des variables prévues dans le fichier IPUMS. Cette liste est d'ailleurs intéressante à examiner en soi puisqu'elle montre bien la variété des domaines d'investigation dans les différents pays du monde.

Pour chaque domaine et pour chaque variable, on présente les informations suivantes :

- le nom de code de la variable dans IPUMS, son label et une traduction proposée.

- le nombre de modalités théoriques (d'après la nomenclature) et le nombre des modalités observées en effectuant des sommations sur la partie Alsace du fichier IPUMS. A travers la différence entre les deux nombres, on mesure parfois l'inadéquation de la nomenclature IPUMS eu égard à la nomenclature des fichiers originaux. Noter que s'il y a 4 modalités, l'information se réduit en fait à deux catégories, puisque l'une concerne les hors champ, l'autre les non déclarés.

- Une indication de la présence/absence à chaque RP de l'information dans le fichier IPUMS.

- Le code SAPHIR correspondant lorsqu'il existe une équivalence suffisamment proche[12].

- Une sommation pour l'Alsace, portant sur la seule population des ménages ordinaires, a été effectuée systématiquement pour chacune des modalités de chaque variable dans chacun des deux fichiers (IPUMS et SAPHIR). Ces résultats seront mis en ligne (format Excel). Il aurait été souhaitable, bien entendu, de pouvoir faire la même chose pour l'ensemble de la France, en effectuant des croisements adéquats de variable (ex : âge et formation pour mettre en évidence des effets de seuils chez les jeunes).

- Chaque tableau fait l'objet d'un bref commentaire établi notamment sur la base des résultats obtenus par sommation des effectifs. Compte tenu de la complexité qui s'attache aux problèmes de comparaisons des variables, en soi, et plus encore, dans une perspective temporelle, la description des variables n'est donnée que de façon indicative, avec le libellé exact en anglais, mais avec une traduction qui devrait parfois être revue. Les tableaux sont donc à considérer comme une première base d'informations avant une poursuite du travail par des investigations approfondies.

[12] *En référence à la note du 22 avril 2004 décrivant le fichier du moment. Une version ultérieure (2005), reprend sensiblement les mêmes labels.*

1 – LOGEMENT – EQUIPEMENT

Fichier IPUMS										Codes SAPHIR
libellés	Codes	Nombre de modalités		Disponibilité IPUMS						
		Théoriques	Représentées en Alsace	62	68	75	82	90	99	
Household serial number *numéro d'ordre du ménage*	**SERIAL**	-	-	x	x	x	x	x	x	NUM
Year structure was built *Année de construction*	**BLTYR**	130	19	x	x	x	x			AA
Stories in structure *Nombre d'étages*	**STORIES**	21	12		x					ET
Number of rooms *Nombre de pièces*	**ROOMS**	32	16	x		x	x			NP
Kitchen or cooking facilities *Cuisine*	**KITCHEN**	10	4		x	x	x			
Number of bedrooms *Chambres*	**BEDRMS**	23	11		x					
Bathing facilities *Bain*	**BATH**	6	4		x	x	x	x	x	
Tout à l'égout Sewage	**SEWAGE**	6	5		x	x	x			
Electricity	**ELECTRC**	4	4		x					
Elevator in structure *Ascenseur*	**ELEVATR**	4	4		x	x	x			ASC
Fuel for heating *Combustible*	**FUELH**	18	9		x	x	x	x	x	CB
Central heating *chauffage central*	**HEAT**	7	5		x	x	x	x		CC
Automobile available *Nombre d'automobiles*	**AUTOS**	10	6		x	x	x	x	x	V
Téléphone availability	**PHONE**	4	5		x	x	x	x		
Ownership of dwelling *Statut d'occupation*	**OWNRSHP**	4	4		x	x	x	x	x	SO
Ownership of dwelling D *« « détaillé*	**OWNRSHPD**	45	11		x	x	x	x	x	

*Saphir propose quelques autres variables : **HLM** (logement social ou non) – **NL** (nombre de logements dans l'immeuble – **TL** (type de logement)*

Peu de séries couvrent toute la période 1962-99. Certaines ne portent d'ailleurs que sur une seule collecte. On note l'incomplétude de certaines séries alors même que l'information existe (BLTYR). Certaines incohérentes temporelles apparaissent, dues sans doute à des changements de définition dont il faudrait mesurer les conséquences (présence ou non d'une cuisine selon un critère de superficie).

2 – MENAGES – FAMILLES

Fichier IPUMS										Codes SAPHIR
libellés	Codes	Nombre de modalités		Disponibilité IPUMS						
		Théoriques.	*Représentées en Alsace*	62	68	75	82	90	99	
Group quarters status *Type de ménage*	**GQ**	4	_	x	x	x	x	x	x	CP
Household classification *Classification*	**HHTYPE**	13	8	x	x	x	x	x	x	COHA
Number of persons records *Nombre de personnes*	**PERSONS**		18	x	x	x	x	x	x	TMEN
Number of families *Nombre de familles*	**NFAMS**	10	9	x	x	x	x	x	x	NFAM
Nb of own family members *Taille de la famille*	**FAMSIZE**	99	18	x	x	x	x	x	x	TFAM
Family unit membership *Numéro d'ordre de la famille*	**FAMUNIT**	99	14	x	x	x	x	x	x	NF
Nb of married couples *Nombre de couples*	**NCOUPLS**	10	6	x	x	x	x	x	x	
Nb of mothers *Nombre de mères*	**NMOTHRS**	10	6	x	x	x	x	x	x	
Nb of fathers in HH *Nombre de pères*	**NFATHRS**	10	6	x	x	x	x	x	x	
Nb of own childrens in HH *Nombre d'enfants du ménage*	**NCHILD**	10	10	x	x	x	x	x	x	
Nb of own childrens under 5 y. *Nombre d'enfants de moins de 5 ans*	**NCHLT5**	10	7	x	x	x	x	x	x	
Age of youngest own child **HH** *Age de l'enfant le plus jeune*	**YNGCH**	52	52	x	x	x	x	x	x	
Age of eldest own child *âge de l'enfant le plus âgé*	**ELDCH**	52	52	x	x	x	x	x	x	

Les séries sont toujours complètes et semblent cohérentes. Les séries sont nombreuses car souvent reconstruites (et donc peu informatives en soi).

La définition de la famille est celle d'IPUMS $_{USA,}$ plus large que la définition française. C'est pourquoi le nombre de familles, qui n'est jamais supérieur à 3 en France, peut atteindre 9. Le site d'IPUMS décrit sur des exemples les différentes situations rencontrées (*Family interrelationships*).

3 – INDIVIDUS

Fichier IPUMS										Codes SAPHIR
libellés	Codes	Nombre de modalités		Disponibilité IPUMS						
		Théoriques.	Représentées en Alsace	62	68	75	82	90	99	
Person number *Rang dans le ménage*	**PERNUM**	-	-	x	x	x	x	x	x	RGM
Sexe	**SEX**	3	2	x	x	x	x	x	x	S
Year of birth *Année de naissance*	**BIRTHYR**	128	128	x	x	x	x	x		ANN
Age	**AGE**	102	101	x	x	x	x	x	x	AGE
Age2 grouped into intervals *Groupes d'âge*	**AGE2**	22	17	x	x	x	x	x	x	
citizenship *Nationalité*	**CITIZEN**	7	3	x	x	x	x	x	x	N
Nativity status *Statut à la naissance*	**NATIVTY**	4	2	x	x	x	x	x	x	
School attendance *Fréquentation d'une école*	**SCHOOL**	6	4		x		x		x	
Educational attainment – Int *niveau d'étude le plus élevé - Int*	**EDATTAN**	6	5	x	x	x	x	x	x	
Educational attainment – Int D *Niveau d'étude - Int D*	**EDATTAND**	16	7	x	x	x	x	x	x	
Educational attainment – FR	**EDUCFR**	13	13	x	x	x	x	x	x	DIP
Educational attainment – EU	**EEDATTA**	10	6	x	x	x	x	x	x	
Age when complete education *Age de fin d'étude*	**EDAGE**	25	25		x	x	x			

codification : Int internationale – EU européenne – FR française - D codification détaillée

L'année de naissance n'apparaît pas en 1999 alors que l'information existe. Le code NATIVTY (2 modalités, nés français ou non) est trop réducteur puisque des analyses un tant soit peu approfondies sur l'immigration par origine nécessiteraient de connaître la nationalité détaillée à la naissance. On ne connaît que la nationalité du moment (NATION, voir $ 6).

Le niveau d'études le plus élevé est une variable délicate car probablement souvent mal comprise et par ailleurs soumise à des changements permanents dans le temps du fait des réformes successives du système d'éducation. IPUMS propose des séries spécifiques à la France (EDUCFR en 13 postes mais de nombreuses ruptures) et à l'Europe (EEDATTA, 6 postes apparemment bien cohérentes), cette dernière étant semble-t-il assez proche de la série SAPHIR (en 9 postes, série établie après avis d'expert, mais nécessairement perfectible).

4 – INDIVIDUS – LIENS

Fichier IPUMS										Codes SAPHIR
libellés	Codes	Nombre de modalités		Disponibilité IPUMS						
		Théoriques.	Représentées en Alsace	62	68	75	82	90	99	
Relationship to HH head *Lien avec le chef de ménage*	**RELATE**	7	5	x	x	x	x	x	x	LCM
Relationship to HH head *Lien avec le chef de ménage - D*	**RELATED**	74	11	x	x	x	x	x	x	
Relationship to HH head –Europe *Lien avec le chef de ménage - Eu*	**ERELATE**	14	12	x	x	x	x	x	x	
Nb of unrelated persons *Nb de personnes isolées*	**UNREL**	10	10	x	x	x	x	x	x	
Marital status E *Etat matrimonial – Int*	**MARST**	6	4	x	x	x	x	x	x	
Marital status E *Etat matrimonial détaillé –Int*	**MARSTD**	28	4	x	x	x	x	x	x	
Marital status - Europe *Statut matrimonial –E*	**EMARST**	7	4	x	x	x	x	x	x	M
Head location in HH *Rang du chef de ménage*	**HEADLOC**	3 car.	11	x	x	x	x	x	x	
Spouse's location in HH *Rang de l'épouse*	**SPLOC**	3 car.	11	x	x	x	x	x	x	
Mother's location in HH *Rang de la mère*	**MOMLOC**	3 car.	11	x	x	x	x	x	x	
Fathers's location in HH *Rang du père*	**POPLOC**	3 car.	11	x	x	x	x	x	x	
Rule for linking parent *Dépendance au père*	**PARRULE**	13	7	x	x	x	x	x	x	
Rule for linking spouse *Dépendance à l'épouse*	**SPRULE**	7	5	x	x	x	x	x	x	
Probable stepmother *Lien biologique de l'enfant avec la mère*	**STEPMOM**	7	2	x	x	x	x	x	x	
Probable stepfather *Lien biologique de l'enfant avec le père*	**STEPPOP**	4	5	x	x	x	x	x	x	

Variables SAPHIR : **LCM** (liens avec la personne de référence du ménage – **LCF** (id. de la famille) – **COHA** : indicateur spécifique de cohabitation en 10 postes

Le fichier IPUMS étant destiné a priori à permettre des comparaisons entre des pays ayant des comportements matrimoniaux très différents (ainsi 74 modalités sont prévues a priori), les variables et les modalités prévues ne s'adaptent pas bien à la France. En effet, on ne connaît dans les recensements que l'état matrimonial ainsi que l'appartenance ou non à un couple, sans référence à l'état matrimonial. On dispose aussi d'un lien au sein du ménage (resp. au sein de la famille) avec la personne de référence (dans le passé il s'agissait du « chef » de ménage, resp. de la famille).

Par conséquent les affectations qui ont été effectuées pour harmoniser les concepts ne donnent pas de résultats satisfaisants et, du reste, peuvent même apparaître sans objet. HEADLOC, POPLOC, MOMLOC et

SPLOC qui renvoient chaque individu au rang de la personne correspondante *(ainsi sur l'enregistrement d'un enfant apparaît le numéro d'ordre de sa mère)*, ce qui facilite des rapprochements s'il s'agit de créer des groupes que l'on souhaite comparer entre eux. Les variables PARRULE et SPRULE, STEPPOP et STEPMOM qui renvoient à l'intensité réelle ou supposée des liens qui unissent deux personnes au sein du ménage ou de la famille.
En revanche MARST et EMARST (Etat matrimonial) sont cohérents avec les données françaises. Il en est de même pour RELATE et ERELATE qui correspondent aux liens avec la personne de référence du ménage.

Il reste qu'avec LCM et LCF (liens avec la personne de référence du ménage et de la famille), M (état matrimonial) et COHA (indicateur de cohabitation recalculé en 9 postes), le fichier SAPHIR réduit à 4 le nombre des variables du domaine. Bien que les séries obtenues soient cohérentes, il conviendrait cependant d'expertiser les choix de façon à prendre la mesure de la perte d'information éventuelle que cette simplification a créée. Rappelons que dans le fichier SAPHIR, le concept de famille est restreint. Les familles ont été reconstituées sur toute la période dans le sens qui avait cours dans les années quatre-vingt-dix (âge limite de 24 ans pour appartenir à une famille). Le programme de traitement permet cependant de changer aisément ce seuil si l'on faire réaliser des investigations spécifiques sur cette question.

5 – TRAVAIL – EMPLOI

Fichier IPUMS										Codes SAPHIR
libellés	codes	Nombre de modalités		Disponibilité IPUMS						
		Théoriques.	Représentées en Alsace	62	68	75	82	90	99	
Period seeking work *Recherche d'emploi*	**LOOKJOB**	11	10	x	x	x		x	x	
Full-time or part-type work *Temps plein/partiel*	**HRSFULL**	4	4					x	x	
Means of transport. school or work *Mode de transport*	**TRNWRK**	29	15		x			x	x	
Class worker (general version) *Statut*	**CLASSWK**	6	4	x	x	x	x	x	x	
Class of worker (general version) *« « détaillé*	**CLASSWKD**	57	9	x	x	x	x	x	x	
Class of worker – (Europe) *Classe de travailleurs*	**ECLASWK**	8	5	x	x	x	x	x	x	
Employment status *Statut – Int*	**EMPSTAT**	5	4	x	x	x	x	x	x	
Employment status (detailled) *Statut détaillé – Int*	**EMPSTATD**	52	10	x	x	x	x	x	x	
Employment status (Europe) *Statut – Eu*	**EEMPSTA**	9	7	x	x	x	x	x	x	
Sector of employment *Secteur*	**EMPSECT**	12	6					x	x	
Number of employees *Nombre d'employés*	**EMPLNO**	10	10		x	x	x	x	x	
Occupation - unrecoded *« « « détaillé*	**OCC**	4 car.	146	x	x	x	x	x	x	
Occupation - ISCO	**OSSISCO**	14	11	x	x	x	x	x	x	
Industry - unrecoded *Activité*	**IND**	5 car.	339	x	x	x	x	x	x	
Industry – general recode *Activité regroupée*	**INDGEN**	20	18	x	x	x	x	x	x	

ISCO : International Classification of Occupation
*Codes SAPHIR : **TA** : type d'activité (9 postes) – **ST** : statut (10 postes) - **CS** : catégorie socioprofessionnelle (30 poste.) Variables non harmonisées : **PR** : profession et **AE** (Activité économique)*

Il s'agit d'un domaine à la marge de la démographie, un domaine très évolutif en raison des transformations touchant à la fois aux activités et aux qualifications. Comment suivre dans le temps, quel sens même cela peut-il avoir de présenter des séries sur les emplois dans l'informatique et les nouvelles technologies ? La question est délicate, mais la disponibilité d'une information détaillée (croisement possible de deux variables élémentaires ayant chacune plusieurs centaines de modalités) fournit une information de base extrêmement riche. Les regroupements sont délicats si on veut leur donner une valeur générale, mais des analyses spécifiques permettront éventuellement à des chercheurs de suivre des séries

homogènes dès lors qu'ils les définiront eux-mêmes. Ainsi par exemple, si l'on veut isoler les catégories dites *stratégiques*, on pourra par exemple sélectionner un ensemble de modalités spécifiques du moment : le charbon et l'acier en début de période, puis le pétrole, puis l'informatique, etc.)
En se référant à la statistique française et à SAPHIR qui en découle, il y a lieu de distinguer trois notions essentielles, à savoir :

- le statut (ST) qui distingue la position de l'individu dans une nomenclature regroupée en 9 postes.
- La catégorie socioprofessionnelle (CS), une catégorie spécifiquement française qui a fait l'objet d'une transformation importante à partir de 1982, mais qui a été rendue quasiment homogène dans Saphir.
- l'activité économique (AE), modifiée plusieurs fois (dont en 1999).

De son côté IPUMS propose trois groupes et des variantes..

- groupe CLASSWK (y c détaillé et Europe) : les séries sont cohérentes, notamment la série ECLASWK
- groupe EMPSTA : la série européenne n'est pas tout à fait cohérente (irrégularités fortes)
- groupe OCC : Les séries OCC pour l'activité économique est une série d'activité économique en une dizaine de postes qui semble très cohérente.

Par ailleurs IPUMS propose des séries qui, bien qu'elles ne soient pas complètes dans le temps, n'en ont pas moins un intérêt évident : HRSFULL (temps plein ou partiel), LOOKJOB (durée de la recherche d'emploi), TRNWRK (mode de transport) et EMPLNO (nombre de salariés).

6 – GEOGRAPHIE - MIGRATIONS

Fichier IPUMS										Codes SAPHIR
libellés	Codes	Nombre de modalités		Disponibilité IPUMS						
		Théoriques.	Représentées en Alsace	62	68	75	82	90	99	
Continent or region of country *Région continentale*	**REGIONW**	18	_	x	x	x	x	x	x	s.o.
Region - France *Région de résidence*	**REGNFR**	22	_	x	x	x	x	x	x	RR
Region Europe Nuts1 *Région nuts 1*	**ENUTS1**	8	-	x	x	x	x	x	x	
Region Europe Nuts 2 *Région nuts 2*	**ENUTS2**	22	-	x	x	x	x	x	x	RR
Urban-rural status *Caractère urbain/rural*	**URBAN**	4	2	x	x	x	x	x	x	
Region of birth - France *Région de naissance – Fr*	**BPLFR**	32	32	x	x	x	x	x	x	RN
Region of birth – Europe nuts1 *Région de naissance*	**EBPLNT1**	12	11	x	x	x	x	x	x	
Region of birth – Europe nuts2 *Région de naissance*	**EBPLNT2**	28	17	x	x	x	x	x	x	
Country of citizenship *Nationalité*	**NATION**	318	27	x	x	x	x	x	x	
Same house last census *Chang. de résid.antérieure – Fr*	**MGHOUFR**	4	4				x			
Country of residence last census *Pays de Résidence antérieure*	**MGCTRY4**	22	8			x			x	PRA
Region de residence at last census *Région de résid. antérieure– Fr*	**MIGFR**	23	23	x	x	x	x	x	x	RRA
Migration status, last census *Statut migratoire au RP précédent*	**MGRATEC**	4	4	x	x	x	x	x	x	
Region of work - France *Région de travail – Fr*	**PWRKFR**	24	23	x	x	x	x	x	x	RLT

*SAPHIR : **DCR** – **DCRA** – **DCLT** : commune de résidence, de résidence antérieure, de travail*
***DN** : département de travail*

Ce domaine peut paraître relativement complexe alors qu'il peut, pour l'essentiel, se résumer en 4 groupes correspondant respectivement au lieu de résidence (au moment du recensement) au lieu de naissance (LN), au lieu de résidence antérieure (LRA) et au lieu de travail (LT). A cela s'ajoute le caractère rural/urbain (URBAN) de la commune de résidence. Par ailleurs on a intégré dans ce domaine le code NATION (nationalité au moment de la collecte).

Le niveau géographique le plus fin d'IPUMS est la région (au sens de la Nuts2[13]. IPUMS distingue parfois les DOM-TOM, seuls ou regroupés, mais d'une façon telle que parfois les séries sont incohérentes (EBPLNT1).

Lieu de résidence (LR). IPUMS propose outre REGIONW (région au sens d'un groupe de pays, sans intérêt en l'occurence), la Nuts2 et son regroupement en Nuts1 pour la « grande région», une entité sans pouvoir en France (contrairement au Land ou à la Communidad).

Lieu de naissance (LN) (EBPLNT1 – EBPLT2 pour la Nuts, BPLFR pour la France). Toutefois l'utilisateur devra être attentif puisque les DOM-TOM sont classés à l'étranger avant 1999.

Lieu de résidence antérieure (LRA) (MGCTRY4 et MIGFR pour la France) : mêmes remarques

Lieu de travail (LT) (PWRKFR) : intéressant pour les régions ayant des travailleurs frontaliers.

Le code NATION donne en principe une répartition très détaillée des nationalités (théoriquement en 318 modalités), mais le fichier n'en retient que 27. De son côté Saphir propose trois séries cohérentes de 75 modalités, pour le lieu de naissance, le lieu résidence antérieure et la nationalité (lieu de travail se réduit à quelques pays frontaliers).

- **IPUMS : des analyses territoriales impossibles**

Tout en travaillant sur des territoires géographiques importants en population, nombre d'utilisateurs souhaitent disposer d'informations sur des territoires quelconques (non composées d'une seule ou d'un ensemble de régions). C'est impossible dans l'état actuel du fichier IPUMS. En revanche du fait que la localisation est très fine dans SAPHIR, tout regroupement est a priori possible. Mais la diffusion des résultats n'est pas autorisée, sauf accord explicite de l'INSEE. C'est grâce à des conventions spécifiques que plusieurs chercheurs ont pu ainsi travailler sur des territoires de petite taille.
Au **lieu de résidence,** le niveau le plus fin est la commune, mais on peut accéder aussi au numéro de l'Iris (depuis 1990), de l'immeuble et du logement. Certes la reconstitution de territoires urbains cohérents dans le temps est difficile mais, s'agissant des grands ensembles et au prix d'investissements longs, on peut recomposer des ensembles identiques sur plusieurs périodes successives (un travail en ce sens a été réalisé avec l'Université de Strasbourg sur les grands ensembles – Bénédicte GERARD, voir bibliographie). Au **lieu de résidence antérieure** et de **travail** on dispose de la commune, mais pour le **lieu de naissance**, on ne dispose que du département. L'harmonisation communale sur longue période a été possible dans la majorité des cas, mais pour environ 0,4% de la population un investissement supplémentaire aurait été nécessaire pour assurer une meilleure cohérence spatiale.

[13] *nuts : nomenclature européenne des unités territoriales statistiques*

II – En savoir plus sur SAPHIR

Longtemps les recensements de la population ont été, en France, exploités de façon indépendante les uns les autres. En conséquence, les nombreuses publications qui faisaient suite à la collecte ne concernaient qu'une seule date. Tout travail d'analyse d'un domaine ou d'un territoire cherchant à prendre en compte la durée impliquait par conséquent la compulsion de plusieurs documents, avec obligation de porter attention aux changements de méthodes opérés d'une opération à l'autre.

C'était le cas notamment dans une région comme l'Alsace où a été développé le fichier SAPHIR. Celle-ci présentait un certain nombre de spécificités qu'il était intéressant de regarder de plus près. Un déficit de femmes dans certaines zones industrielles rurales, un mouvement frontalier en pleine croissance qui perturbait les marchés locaux du travail et interpellait les élus. Tour cela s'inscrivait dans un contexte national soucieux de mieux de comprendre les transformations profondes qui traversaient la société : mutations économiques et courants migratoires intenses : échanges Paris-province et nord-sud. La DATAR (Délégation à l'Aménagement du Territoire) fut d'ailleurs l'une des premières institutions à s'intéresser aux possibilités d'investigation offertes par SAPHIR.

Première étape, la proposition en octobre 1989 de créer une base de données migratoires[14] a été bien accueillie par la hiérarchie nationale. Carte blanche nous a été donnée pour nous engager dans un investissement à long terme : constitution d'un premier fichier national au niveau départemental portant sur les premiers recensements informatisés (1962, 68, 75, 82, 90, puis 1999), ensuite extension par la prise en compte de la commune, ce qui autorise a priori tout regroupement géographique à la demande. En intégrant les dernières collectes, c'est pratiquement un demi-siècle de la démographie du pays qui peut être regroupée en une même base cohérente.

Les principes

Pour donner une chance de pérennité au travail, pour éviter ce qu'il arrive parfois quand on lance un projet dont les contours ne sont pas très clairs, on a procédé de façon empirique. Cependant dès le départ, plusieurs principes ont été mis en exergue, à savoir la nécessité de valoriser l'information existante, en privilégiant la simplicité, la souplesse, la transparence et en donnant au projet un caractère évolutif.

Simplicité - souplesse. En premier lieu on a cherché à faciliter l'accès aux données des recensements antérieurs. Il existait alors nombre de fichiers archivés, créés lors de chaque opération mais conservés de façon hétéroclite, alors même que la collecte des données, réalisée selon des procédures identiques sur l'ensemble du territoire, offrait une masse d'informations *cohérentes.* La première phase du travail a conduit à faire l'inventaire des informations disponibles et à élaborer des tableaux synoptiques des variables à qui on a cherché à donner des symboles facilement mémorisables. Il fallait notamment distinguer les variables utiles – qui apportent l'information originale - des variables redondantes ou des variables reconstruites.

Du fait que ce travail ne correspondait pas à une commande, qu'il n'était pas inscrit dans un programme et qu'aucun délai contraignant n'était fixé, il a été possible de progresser d'une façon très souple. Des idées parfois jugées intéressantes ont pu être abandonnées. On pouvait toujours reculer pour repartir sur de

[14] *« Pour un fichier historique des recensements - Mieux connaître les brassages de la population » -– note interne à l'INSEE - octobre 1989*

nouvelles bases, tout en préservant l'acquis. Cette façon de procéder n'a pas que des avantages, mais elle évite parfois de s'engager trop longtemps dans des impasses, comme on le constate parfois lorsque les opérations sont préalablement corsetées par des règles contractuelles.

La transparence, le repérage des anomalies

En dehors de son utilité intrinsèque, un fichier historique a ceci d'intéressant qu'il permet de mettre immédiatement en évidence les anomalies. Le manque de transparence dans les procédures de traitement des données est souvent reproché par les utilisateurs de la statistique publique. Les chiffres successivement publiés par l'INSEE à l'issue d'un recensement sont le résultat de plusieurs opérations qui chacune donnent des résultats différents : d'abord de simples comptages, puis exploitation exhaustive (dite légère) sur quelques variables, enfin exploitation par sondage (dite lourde) sur l'ensemble des variables[15]. Que dire aux décideurs quand les chiffres pour un même territoire, une même variable à une même date sont contradictoires et conduisent à des interprétations différentes? Or l'un des objectifs du travail était de pouvoir fournir aux élus locaux des tableaux présentant la dynamique sur long terme de leur territoire, avec la possibilité d'effectuer des comparaisons avec d'autres zones comparables. Par ailleurs, en un temps où l'évaluation des politiques publiques est devenue un impératif toujours plus marqué, un fichier historique des recensements judicieusement exploités permet de mettre en regard les choix et les résultats (par exemple, les effets du TGV sur le territoire).

Le souci de pédagogie était clairement affiché de façon à ce que le monde de l'éducation, voire le grand public trouvent dans la base, à travers des présentations adaptées de résultats, un intérêt certain dans la mesure où les informations des recensements sont basiques et donc essentielles à la connaissance de la société.

Mais du fait des restrictions en matière de diffusion des données (Commission nationale Informatique et Libertés – CNIL), mais aussi parce que les techniques de diffusion étaient encore balbutiantes (la mise en ligne n'était alors pas envisageable), l'outil devait rester un produit interne à l'INSEE, un outil d'études devant servir à la fois à des fins d'étude et de diffusion par les canaux habituels du moment.

Souci de créer un « système »

Une fois les premières opérations terminées, l'idée est venue de chercher à donner à ce travail une dimension supplémentaire. Il s'agissait d'aller au-delà d'un simple fichier pour un faire un « *système* » d'analyse. D'où l'acronyme SAPHIR (Système d'Analyse de la Population par l'Historique des recensements) qui a été donné à l'investissement quand il s'est trouvé bien engagé.

Plusieurs pistes ont été suggérées. Parmi ces pistes on peut souligner le souci de corriger les erreurs les plus graves, de créer des cadres d'analyse, mais aussi d'intégrer d'autres sources d'information, notamment les données des enquêtes sur les forces de travail.

Il faut reconnaître que peu d'investissements ont été développés dans ce sens à l'initiative de l'institution. Signalons, cependant, la constitution de bilans intercensitaires des ressources humaines, mais il y avait bien

[15] *Un fichier historique établi sur une même base (le sondage au quart ou au cinquième) présente au moins l'avantage d'une certaine cohérence, si les chiffres ne sont pas forcément les plus justes.*

d’autres investigations possibles. Ainsi, celle qui aurait consisté à enrichir la base par la prise en compte des informations insuffisamment mises en valeur des données issues des exploitations exhaustives.

Toutefois, la facilité d’accès à toutes les données issues de l’exploitation des sondages a permis d’investir quelques champs nouveaux. Pour n’en citer que deux : la création d’un fichier familles avec application au thème de l’immigration et la création d’un fichier sur le voisinage, par la prise en compte de la proximité, en probabilité, des logements entre eux. Sans compter les nombreuses thèses de doctorat qui ont été nourries peu ou prou par le contenu de la base.

English abstract

Use of Census Data in Europe in a Comparative Perspective[16]

Barcelona June 9-11 2011

The paper compares two microdata bases compiled from population censuses. The first base, IPUMS (Integrated Public Microdata Series International) is the result of collaboration between the "Minnesota Population Center" and the national statistical institutes. It currently incorporates 159 censuses conducted in 55 countries, including France. The second base, SAPHIR, (Système d'Analyse de la Population par l'Historique des Recensements - Population Analysis System using Historical Census) was produced through an initiative of the Regional Directorate of the Insee-Alsace. This latter is concerned only with French censuses. In each case, the objectives, the principles of the constitution and the public targeted are very different.

At a time when European demographers are undertaking a similar operation, it is helpful to combine experiences and provide researchers with a data base of optimum quality, knowing that the difficulties of harmonising the data, already considerable for a given country, become even more acute when there is confrontment of national statistical systems each with its own history.

Saphir: a historical French census data file

For a long time now the population censuses in France have been used in ways that are completely independent of each other. Consequently, the numerous publications subsequent to the data collection only concern a single date. As a result, any work analysing a field or a territory trying to take into account a period of time required several documents, with a need to pay attention to the changes in methods used from one operation to the other.

This was the case particularly in a region like Alsace where the SAPHIR data file has been developed. A number of specific characteristics were presented that were worth a closer look. A deficit of women in some rural industrial areas, and a growing [cross] border movement which disrupted the local work markets raised questions for the elected representatives. All of this took place in a national context concerned to gain a better understanding of the profound transformations which were cutting across society: economic changes and intense migratory currents: Paris-province and north-south. Moreover, the DATAR (Delegation for Territorial Planning) was one of the first institutions to show an interest in the investigation possibilities offered by SAPHIR.

The first stage concerning the proposal in October 1989 to create a migratory data base[17] *was well-received by the national hierarchy. A free-hand was given to us to invest in the work: constitution of the first national data file at department level involving the first computerised censuses (1962, 68, 75, 82, 90, then 1999), then extended to take account of the commune, which enabled, on the face of it, any geographic grouping on request. By integrating the latest collections, almost half a century of the country`s demography can be amalgamated in the same coherent data base.*

16 *Texte remis aux participants du colloque de Barcelone*

17 *To make known the movements of the population - For a historical data file of the censuses – October 1989*

The Principles

To give the work a chance of permanence, to avoid what sometimes happens when the parameters of a project are unclear, we went ahead in an empirical way. However, from the outset, several principles were highlighted, namely, the need to take full advantage of the existing information, by favouring simplicity, flexibility, and transparency, whilst giving the project an evolving character.

Simplicity - flexibility. In the first instance, we sought to facilitate access to data from previous censuses. At the time there was a number of archived data files in existence, created during each operation but kept heterogeneously, even though the data collected using identical procedures over the entire territory, offered a mass of coherent information. The first phase of the work resulted in an inventory of the information available and the production of synoptic tables of the variables with easily memorisable symbols. It was necessary to distinguish the useful variables – which provide the original information – from the redundant variables or reconstructed variables.

As this work did not relate to an order, was not part of a programme, and as there was no tight deadline, it was possible to move forward in a very flexible way. Some ideas considered to be attractive at times could be abandoned. It was always possible to step back and start again on new bases, whilst retaining the acquisitions. This way of proceeding does not only have advantages, but it sometimes avoids spending too much time on insurmountable impasses, as is sometimes the case when operations are already constrained by contractual rules.

Transparency, the locating of anomalies -

Apart from its intrinsic usefulness, a historical data file is advantageous because it immediately highlights anomalies. The lack of transparency in the data processing procedures is often criticised by the users of public statistics. The figures successively published by INSEE [National Economic Studies and Statistics Institute] following a census are the result of several operations each giving different results: firstly, simple counting, then exhaustive `exploitation` (called light) of some variables, finally, `exploitation` through sampling (called heavy) of all of the variables[18]. What do we say to the decision makers when the figures for the same territory and the same variable at the same date are contradictory leading to different interpretations? Now, one of the objectives of the work was to be able to provide the local elected representatives with tables showing the long-term dynamic of their territory, with the possibility of making comparisons with other comparable areas. Moreover, at a time when an evaluation of public policies has clearly become an essential requirement, an historical census data file wisely processed would enable the choices made by the elected representatives and their consequences to be compared (as the effects of the TGV on the territory).

The concern of pedagogy was clearly displayed so that the world of education and even the general public showed definite interest in the base, through adaptive presentations of results, insofar as the information from censuses is basic and therefore essential for a knowledge of society.

[18] *A historical data file established on the same base (sampling of a quarter or a fifth) presents at the least, the advantage of a certain coherence, if the figures are not necessarily the most sound.*

However, due to restrictions regarding the dissemination of data (French Data Protection Authority, the CNIL, Commissison Nationale Informatique et Libertés), but also because the dissemination techniques were still hesitant (putting on-line could still not be envisaged), the tool had to remain a product used only internally by INSEE, a study tool which had to simultaneously serve the purposes of study and dissemination through the usual channels of the time.

The creation of a "system"

Once the first operations were complete, the idea occurred to give this work an extra dimension. It involved going beyond a simple data file to create a "system" of analysis, hence the acronym SAPHIR (Système d'Analyse de la Population par l'Historique des Recensements), which was given to the investment when it became well-established.

Several routes were suggested. Underlining these routes lay a concern to correct the most serious errors, to create frameworks of analysis, but also to integrate other sources of information, particularly data from labour force surveys.

It has to be acknowledged that there was little development of investment in this sense on the initiative of the institution. Let us report, however, the constitution of human resource assessments between census periods, but there were many other possible investigations. These could have included an investigation which would have been used to enrich the base by taking into account information insufficiently used from data resulting from the exhaustive `exploitations`.

However, the facility for accessing all the data from the `exploitation` of the sampling has enabled some new fields to be created. To cite just two of them: the creation of a families data file with application to the theme of immigration, and the creation of a data file on vicinity, taking into account proximity of housing in terms of probability. This is without counting the numerous doctoral theses information for which has been more or less supplied from the content of the base.

Bibliographie

La création de SAPHIR et son exploitation ont conduit à la rédaction de nombreuses notes techniques et de documents divers. Il est prévu de mettre en ligne les différentes publications connues.

Ce rapport s'appuie notamment sur deux notes techniques. L'une datée d'août 1993 (*SAPHIR – Un système d'Analyse de la Population par l'Historique des Recensements*) a paru dans la série « *Rectangle* » de la Direction de la Diffusion et de l'Action Régionale (réf H9305). L'autre est datée du 22 avril 2004 (note interne à l'INSEE intitulée *SAPHIR3 – un fichier historique des ménages*) décrit les variables utilisées dans une perspective plus large (la note fournit notamment une liste exhaustive des variables disponibles dans les fichiers archivés, variables originales, mais aussi variables reconstruites).

On signale seulement ici quelques ouvrages, publications ou thèses parmi les plus importants dont le contenu s'appuie largement sur des exploitations plus ou moins originales du fichier SAPHIR. Nombre de ces travaux ont été élaborés dans le cadre de deux associations, l'AIDELF (Association Internationale des Démographes de Langue Française) et de la CUDEP (Conférence Universitaire de Démographie et d'Etude des Populations).

A - Quelques ouvrages ou articles ayant utilisé SAPHIR :

- Bernard BARBIER (Académie de Marseille, 1995) : *un quartier, l'Ile Saint-Louis*, Cette brève étude montre un exemple d'utilisation de SAPHIR au niveau infra-communal (de tels travaux infra-communaux sont possibles en fusionnant les fichiers de sondage et exhaustifs).
- *Atlas des villes nouvelles d'Ile-de-France* mai 1995 - Convention avec le ministère de l'Equipement Direction de l'Architecture et de l'Urbanisme), à l'initiative de Jean-Claude RENAUD
- Emmanuel AMOUGOU - *Etudiants d'Afrique noire, une jeunesse sacrifiée* ? - l'Harmattan 1999
- Jean BASTIE - *Nouvelle histoire de Paris* - Bibliothèque historique de la ville de Paris 2000
- Atlas de France - Territoire et aménagement - Reclus – n°14 La Documentation française 2001 (sur les migrations Paris-province)
- Bernard AUBRY - *les immigrés en France 1962-1999* - Colloque AIDELF (Association internationale des Démographes de Langue Française) - Budapest - septembre 2004.
- Bénédicte GERARD – *Connaissance de l'échelon infra-urbain à partir des recensements. L'exemple des grands ensembles d'habitation strasbourgeois (1968-1999)* - Cahiers de démographie locale - Néothèque - 2008
- Bernard AUBRY -– *le voisinage, proposition d'indicateurs* – Cahiers de démographie locale - Néothèque -2008
- Bernard AUBRY et Michèle TRIBALAT – *Les jeunes d'origine étrangère* – Commentaire – juin 2009
- TRIBALAT Michèle – *Les yeux grands fermés – L'immigration en France*- Denoël 2010

B - Thèses (IEDUB, Institut d'Etudes Démographiques de l'Université de Bordeaux)

- Christophe BERGOUIGNAN - Thèse de doctorat (20 janvier 1999) - *Les sources administratives, un outil pour le développement local*
- Christophe BERGOUIGNAN - Thèse d'habilitation à diriger des recherches (HDR) de doctorat (10 décembre 2004) - *Les confrontations en analyse démographique*
- Mélanie CAILLOT - Thèse de doctorat de (28 novembre 2008) *Analyse démographique de l'élargissement de l'accès à l'enseignement supérieur*
- Ceren INAN –Thèse de doctorat (4 décembre 2009) – *Dynamique démographique de la population active occupée en France*
- Khalid ELJIM - Thèse de doctorat (24 novembre 2009) - *Maghreb-France : Quelle émigration pour l'avenir ? Bilan et perspectives.*