Pondérations de l'enquête ChIPRe

Géraldine Charrance, janvier 2022

L'enquête ChIPRe est la première expérience en France du recours à la méthode d'échantillonnage « Network sampling with memory » (NSM). Cette méthode constitue une nouvelle variante de sondage par chaînage (boule de neige, Respondant Driven Sampling (RDS)…). Ces méthodes consistent à enquêter au sein de réseaux en sélectionnant au départ quelques individus appelés « graines ». Seuls ces individus sont désignés par le sondeur. Par la suite, ce sont les enquêtés eux-mêmes qui recrutent/désignent leurs pairs qui seront sollicités à leur tour pour participer à l'enquête. Les méthodes par chaînage permettent en théorie d'atteindre des pans de la population non directement accessibles à des enquêteurs, mais souffrent d'un biais de sélection très fort.

Dans le but de pallier ce défaut, une équipe de l'université de Caroline du Nord a développé la méthode NSM qui présente des avantages théoriques notamment sur la précision des estimations, au prix d'une complexité de mise en œuvre plus importante. La particularité de NSM est de recréer, au fur et à mesure du terrain, une base de sondage de la population cible composée des personnes citées par les répondants et de tirer aléatoirement les futurs enquêtés dans cette base. Contrairement à la méthode RDS, elle ne cherche pas à enquêter tous les contacts cités mais vise à intégrer une dimension aléatoire dans une méthode de sondage empirique. L'algorithme comprend une première phase exploratoire dite « Search », à la recherche des divers pans du réseau, puis une seconde phase de tirages aléatoires dans le réseau en prenant en compte la structure découverte en première étape. Selon ses concepteurs, la méthode NSM devrait permettre d'obtenir des estimations d'une précision équivalente à celle d'un sondage aléatoire simple.

Après deux expériences à l'étranger (en Tanzanie et en Caroline du Nord), la méthode a été utilisée pour la première fois en France entre septembre 2020 et juin 2021 dans le cadre d'une enquête menée auprès des immigrés chinois en Ile-de-France (ChIPRe). A l'issue du terrain, 501 questionnaires et quelques 1700 citations ont été collectés.

Ce document a pour objectif de détailler le calcul des poids attribués à chacun des 501 répondants à l'enquête. Différents scénarios ont été envisagés, qui combinent ou non plusieurs étapes : le calcul des poids de sondage, la correction de la non-réponse totale par modélisation et GRH, le redressement de l'échantillon par calage sur marges et la troncature des poids. Le scénario finalement retenu repose sur un calage sur marges des poids de sondage calculés selon la méthode NSM, qui sont ensuite tronqués au seuil de 1%.

Calcul des poids de sondage

Comme indiqué dans l'introduction, la méthode NSM repose sur une sélection aléatoire des individus à enquêter (hors graines) dans la base de sondage constituée au fil du terrain. Comme pour toutes méthodes probabilistes, on peut donc calculer pour chaque personne échantillonnée une probabilité de sondage et un poids (découlant de la probabilité de sondage).

Les poids de sondage sont calculés pour les 844 individus échantillonnés non graines.

D'après l'article de Ted Mouw and Ashton M. Verdery (2012), la formule pour obtenir le poids de sondage est la suivante :

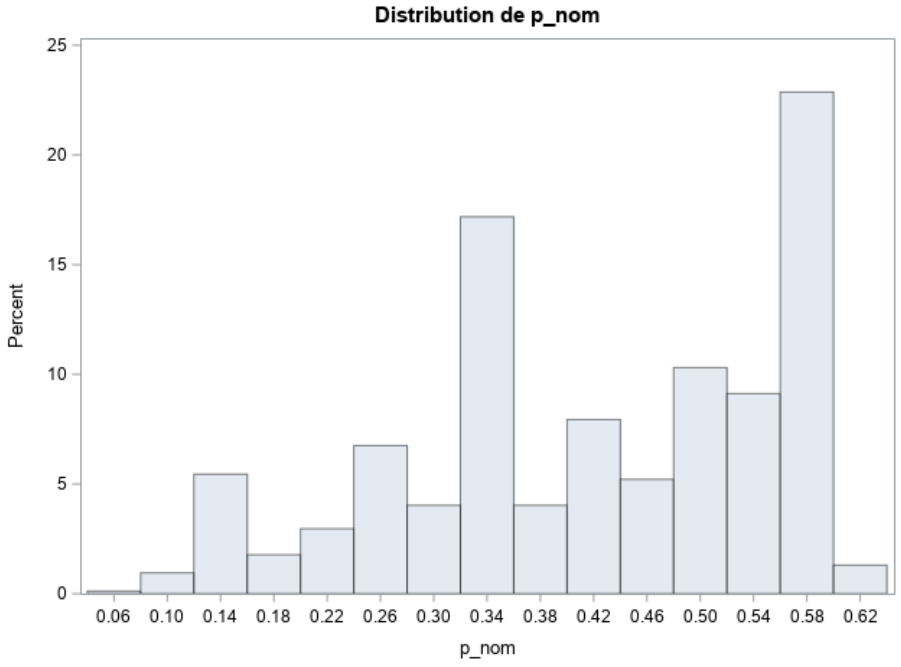
W_{Sond} = 1/ (p_nom*Csr_final)

Il faut donc mobiliser:

- La probabilité d'être nommé (P_nom) au moment du tirage pour chaque RID échantillonné (dans la base synthese_tirage pour chaque tirage)

Principales statistiques de P_nom

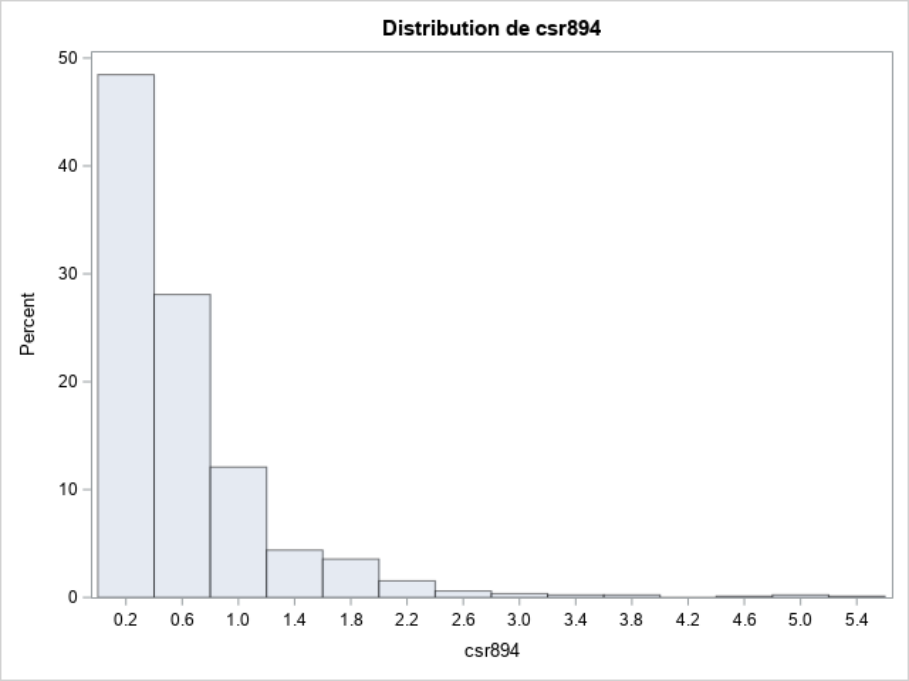
	Valeurs pour P_nom
Moyenne	0,4198
Min	0,0553
1%	0,1188
50%	0,4347
99%	0,6130
Max	0,6130



- Le cumul des taux de sondage à la fin du processus d'échantillonnage pour chaque RID échantillonné (correspond à la variable CSR894 disponibles dans la base CSR du dernier tirage)

Principales statistiques du dernier CSR

	Valeurs pour CSR894
Moyenne	0,5988
Min	0,0333
1%	0,0337
50%	0,4166
99%	3,1321
Max	5,3986



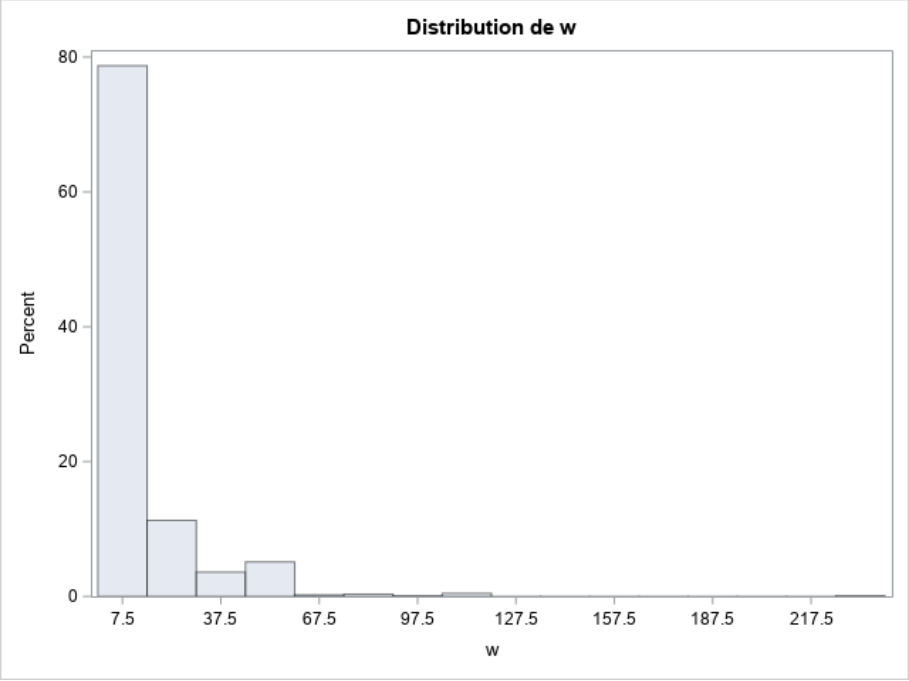
Les graines, quant à elles, ont un poids fixé à 1.

A l'issue de ce calcul, nous obtenons un système de poids de sondage pour les 844 individus échantillonnés et les 16 graines.

Principales statistiques de poids

	Sur les 860 individus graines ou échantillonnés	Sur les 501 répondants
N	860	501
Somme	10253,35	6437,97
Coefficient de variation	143,69	149,08
Moyenne	11,92	12,85
Min	0,57	0,57
1%	0,88	0.94
50%	5,93	6.37
99%	75,86	75.86
Max	235,20	235,20
Max/Min	412,63	412,63
99%/1%	86,45	80,69

Sur les 860 graines ou échantillonnés



Correction de la non-réponse totale

Dans la base constituée grâce aux rosters collectés (ayant servi de base de sondage), nous disposons d'un certain nombre d'informations sur les individus qui peuvent être mobilisées dans le cadre d'une correction de la non-réponse totale.

Dans la base des rosters, les informations sociodémographiques d'un individu sont renseignées (et donc estimées) par le.s citant.s. Nous avons donc été confrontés à trois difficultés :

- L'information peut être erronée, étant donné qu'elle est délivrée par un tiers et non par la personne elle-même
- L'information peut être manquante (le citant refuse ou ne sait pas donner telle ou telle information sur la personne citée)
- L'information peut être multiple et discordante si la personne est citée par plusieurs répondants (selon un citant, la personne a entre 18 et 24 ans et selon un autre, entre 25 et 29 ans).

Afin de construire des variables sans valeur manquante et adaptées à la modélisation, un certain nombre de traitements ont été réalisés sur les données du roster (plus de détails en annexe) :

- Réduction du nombre de modalités pour la classe d'âge (10 → 5), la province d'origine (33 → 12) et la durée de résidence (8 → 4)
- En cas de citations multiples : la valeur la plus fréquente est retenue. En l'absence de valeur plus fréquente, la modalité «0 – Informations discordantes » est affectée.
- Les refus/nsp sont une modalité à part entière pour chaque variable. En cas de citations multiples, les autres modalités priment toujours sur les modalités « refus/nsp».

La participation a ensuite été modélisée sur la base des 825 personnes échantillonnées et non identifiées comme hors cible (les graines étant choisies par le sondeur et leur appartenance à l'échantillon étant conditionnée par l'acceptation de participer, leur taux de réponse est donc de 100%, elles sont exclues donc de la modélisation). Sur ce champ, le taux de réponse brut est de 58,8% (=485/825).

La modélisation a été réalisée avec les variables explicatives suivantes (cf annexe pour plus d'informations):

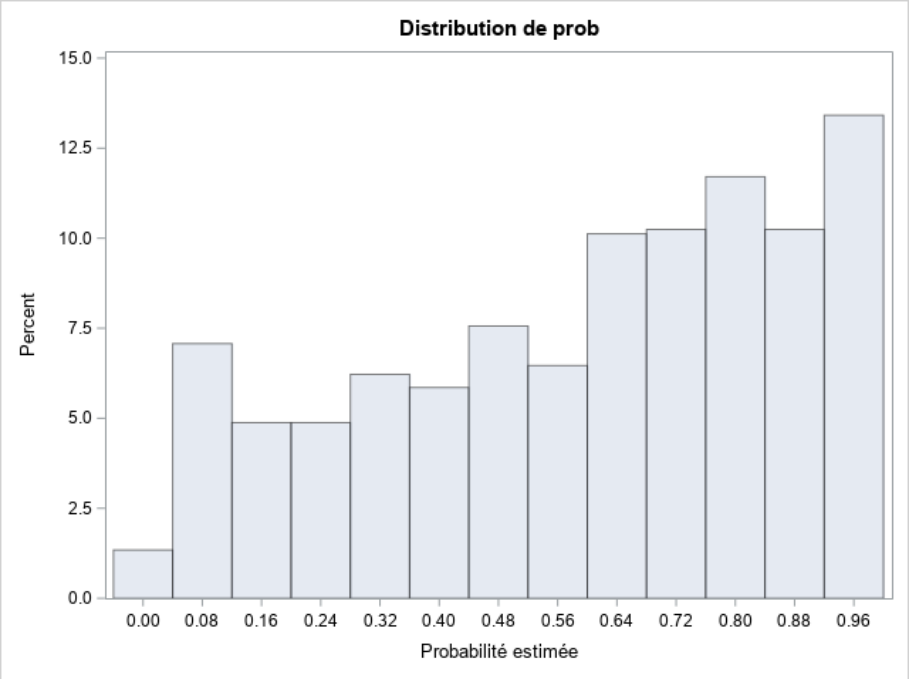
- Sexe
- Classe d'âge
- Province d'origine
- Durée de résidence en France
- Avoir été cité par une graine (oui/non)
- Avoir été cité plusieurs fois (oui/non)
- La présence d'un numéro de téléphone (oui/non)
- La présence d'un compte wechat (oui/non)
- Un des citants est un ami proche (oui/non)
- Un des citants est un membre de la famille (oui/non)
- Un des citants n'a pas su ou voulu définir la relation (oui/non)
- Enquêteur en charge de l'exploitation

Seules les variables suivantes ont un effet propre et significatif sur la participation :

- La durée de résidence en France : les personnes pour lesquelles la durée de résidence n'a pu être définie en raison d'incohérence entre les citants ont une probabilité moindre de participer (que les personnes en France depuis 3 à 10 ans).
- Le fait d'avoir été cité par plusieurs enquêtés est associé positivement à la participation
- Le fait d'avoir été cité par une graine est associé positivement à la participation
- La présence d'un numéro de téléphone, tout comme celle d'un identifiant wechat, semble jouer positivement sur la participation
- L'enquêteur chargé de l'exploitation semble corrélé à la probabilité de participer à l'enquête

Principales statistiques : Probabilité de réponse

	Sur les 820 individus échantillonnés et non hors cible
N	820
Coefficient de variation	47.80
Moyenne	0.59
Min	0.0000017
Max	0.9936682



A partir de la probabilité prédite de participer, des groupes de réponse homogènes ont été constitués.

Probabilités prédites, effectifs de répondants et taux de réponse (brut et pondéré) par GRH

Probabilité prédite de participer	Effectifs de répondants	Effectif total	Taux de réponse brut	Taux de réponse pondéré
[0-35%[	28	197	14,2	15,8
[35-50%[	45	98	45,9	60,3
[50%-65%[	74	122	60,7	59,9
[65%-80%[	126	157	80,3	77,8
[80%-90%[	94	115	81,7	88,0
[90%-100%]	118	131	90,1	91,3
Total	485	820	59,2	64,3

On obtient ensuite le poids de sondage corrigé de la non-réponse totale en divisant le poids de sondage par le taux de réponse pondéré observé dans le groupe auquel il appartient.

Le poids des graines reste, quant à lui, inchangé, leur probabilité de réponse étant de 100%.

Principales statistiques : Poids de sondage corrigé de la non-réponse totale

N	501
Somme	9998,24
Coefficient de variation	181,75
Moyenne	19,96
Min	0,62
1%	1
50%	8,65
99%	133,67
Max	480,07
Max/Min	774,31
99%/1%	133,67

Redressement / calage sur marges

Choix de la source

Pour obtenir des marges et procéder au redressement, deux possibilités s'offraient à nous :

- Utiliser le réseau dévoilé (mêmes difficultés que celles citées précédemment)
- Utiliser une/des sources externes à l'enquête (recensement 2018, base AGDREF)

En comparant la structure du réseau dévoilé à celle de la population chinoise dans le recensement, nous avons été forcés de constater que le réseau ChIPRe souffrait a priori de biais importants : surreprésentation des jeunes, des personnes arrivées récemment en France (corrélé à l'âge) et des personnes diplômées. Il ne semblait donc pas pertinent d'utiliser la structure du réseau comme marge

de calage. Ce biais constaté sur le réseau s'explique en partie par le profil des graines choisies (beaucoup dans le milieu étudiant, notamment).

Profils des graines

	Effectifs	Pourcentages
Total	16	100%
Sexe		
Hommes	5	31,3%
Femmes	11	68,7%
Classe d'âge		
18-29 ans	5	31,3%
30-39 ans	4	25,0%
40-49 ans	1	6,2%
50-59 ans	4	25,0%
60 ans et plus	2	12,5%
Département de résidence		
75 – paris	3	18,8%
77 – Seine et Marne	1	6,2%
78 – Yvelines	2	12,5%
91 – Essonne	0	0,0%
92 – Hauts de Seine	3	18,8%
93 – Seine Saint Denis	6	37,5%
94 – Val-de-Marne	1	6,2%
95 – Val d'Oise	0	0,0%

Nous avons finalement opté pour l'utilisation des données du recensement de 2018 pour produire nos marges de calage. Nous suspectons une sous-estimation de la population immigrée chinoise en Ile-de-France dans le recensement[1]. Malgré cela, le recensement nous parait être, à ce jour, la source d'informations la plus fiable sur cette population. Par ailleurs, nous ne pouvons vérifier si la sous-représentation suspectée touche uniformément les différents pans de la population ou non.

Une seconde source avait été envisagée (base issue de l'application de gestion des dossiers des ressortissants étrangers en France, AGDREF) mais a finalement été écartée en raison de la difficulté de produire des chiffres à l'échelle individu (un individu pouvant apparaître plusieurs fois s'il a réalisé plusieurs demandes) et ne couvrant que la population ayant demandé un titre de séjour, contrairement au recensement. Elle a néanmoins été utilisée pour des comparaisons sur le sous-champ des individus ayant déclaré avoir rempli un dossier dans l'enquête ChIPRe.

Définition des marges

Une fois la source externe déterminée, il nous faut choisir des variables qui serviront de marges. Pour cela, les variables doivent être disponibles et comparables dans les deux sources (données de l'enquête (questionnaire) et source externe).

[1] Attané, Isabelle. « L'immigration chinoise en France », *Population*, vol. 77, no. 2, 2022, pp. 229-262.

De plus, compte tenu de la forte variation des poids de sondage, nous avons fait le choix de sélectionner un nombre très limité de variables afin de ne pas trop modifier la structure des poids.

Nous avons donc choisi de réaliser le calage sur les variables suivantes :

	Effectifs	Pourcentages
Total	68 704	100%
Sexe		
Hommes	28 534	41,5%
Femmes	40 170	58,5%
Classe d'âge		
18-29 ans	14 572	21,2%
30-39 ans	18 601	27,1%
40-49 ans	15 646	22,8%
50-59 ans	11 478	16,7%
60 ans et plus	8 407	12,2%
Département de résidence		
75 – paris	20439	29,7%
77 – Seine et Marne	2335	3,4%
78 – Yvelines	3164	4,6%
91 – Essonne	2401	3,5%
92 – Hauts de Seine	9452	13,8%
93 – Seine Saint Denis	17538	25,5%
94 – Val-de-Marne	11376	16,6%
95 – Val d'Oise	1999	2,9%

Troncatures

A l'issue du calage sur marges, lorsque les poids sont trop dispersés, une troncature des poids extrêmes est souvent pratiquée. Dans notre cas, deux niveaux de troncatures ont été utilisées et comparées : 1% et 5%.

En pratique, pour la troncature à 1%, cela consiste à ramener les poids inférieurs au 1er centile (1%) à la valeur du 1er centile, et les poids supérieurs au dernier centile (99%) à la valeur du dernier centile.

Le procédé est le même pour la troncature à 5%.

Une fois la troncature réalisée, on ramène la somme des poids à la taille de la population (par une simple règle de trois).

Résultats des différents scenarios testés

Plusieurs scénarios ont été testés afin de disposer d'éléments étayés pour choisir le scenario le plus satisfaisant (structure de l'échantillon final proche de la cible et statistiques de poids finaux acceptables). Dans les scenarios 3 et 4, le poids de sondage n'a pas été pris en compte[2], et dans les scenarios 2 et 4, l'étape spécifique de correction de la non-réponse par groupes de réponse homogène

[2] La prise en compte des poids de sondage dans l'exploitation d'une enquête probabiliste est très fortement recommandée, la pondération permettant de corriger les biais liés au processus d'échantillonnage.

n'a pas été faite. A contrario, dans tous les scenarios, un calage sur les trois variables sexe, classe d'âge et département de résidence a été réalisé.

Récapitulatif des scenarios testés

	Poids de sondage	Correction de la non-réponse par GRH	Calage sur marges
Scénario 1	Oui	Oui	Sexe Classe d'âge Département de résidence
Scénario 2	Oui	Non	
Scénario 3	Non	Oui	
Scénario 4	Non	Non	

Comme indiqué précédemment, pour les scenarios pour lesquels les statistiques de poids étaient peu satisfaisantes (notamment un rapport poids maximum/poids minimum important), deux troncatures ont été réalisées :

- Une troncature des 1% max et 1% min (les poids de 2% de l'échantillon ont été modifiés)
- Une troncature des 5% max et 5% min (les poids de 10% de l'échantillon ont été modifiés)

C'est donc au final 10 pondérations qui ont été comparées sur différents aspects : structure de l'échantillon sur les trois variables de calage et d'autres variables du recensement (non contrôlées dans le calage), statistiques de poids, structure de l'échantillon par rapport au réseau ChIPRe.

Récapitulatif des pondérations produites

Poids_S1	Scenario 1 (sans troncature)
Poids_S1_T1	Scenario 1 + troncature des 1% max et 1% min
Poids_S1_T5	Scenario 1 + troncature des 5% max et 5% min
Poids_S2	Scenario 2 (sans troncature)
Poids_S2_T1	Scenario 2 + troncature des 1% max et 1% min
Poids_S2_T5	Scenario 2 + troncature des 5% max et 5% min
Poids_S3	Scenario 3 (sans troncature)
Poids_S3_T1	Scenario 3 + troncature des 1% max et 1% min
Poids_S3_T5	Scenario 3 + troncature des 5% max et 5% min
Poids_S4	Scenario 4 (sans troncature)

1. Statistiques de poids

Comme indiqué précédemment, certains scenarios conduisent à des statistiques de poids peu satisfaisantes, avec notamment un rapport max/min très élevé, qui nous ont amenés à faire des troncatures pour réduire ce ratio. Ces troncatures ont été réalisées sur les scenarios 1 à 3, le scenario 4 (sans troncature) conduisant à des résultats satisfaisants.

Les scenarios 3 et 4 conduisent à des poids moins dispersés que les scenarios 1 et 2. La forte dispersion des poids finaux est donc principalement imputable à la prise en compte des poids de sondage, cependant, il ne semble pas envisageable de faire l'impasse sur cette 1ère étape.

De même, la correction de la non-réponse totale par GRH opérée dans le scenario 1 conduit à détériorer les poids (comparaison avec le scenario 2). Il convient donc de regarder ce qu'apporte cette correction sur les différents critères étudiés ci-après afin de choisir la pondération à retenir.

Principales statistiques des poids

	S1	S1_T1	S1_T5	S2	S2_T1	S2_T5	S3	S3_T1	S3_T5	S4
Somme	68704	68704	68704	68704	68704	68704	68704	68704	68704	68704
Moyenne	137,13	137,13	137,13	137,13	137,13	137,13	137,13	137,13	137,13	137,13
Coef. de variation	173,63	164,42	129,08	163,76	155,78	116,95	270,88	203,73	93,59	68,93
Minimum	2,40	3,94	10,90	3,33	5,98	14,57	20,61	23,46	45,23	71,00
1er ctl	3,86	3,94	10,90	5,86	5,98	14,57	21,11	23,46	45,23	71,54
5ème ctl	9,22	9,42	10,90	12,21	12,45	14,57	28,53	31,70	45,23	72,74
95ème ctl	565,11	577,59	667,91	502,12	512,09	599,09	344,30	382,65	545,92	375,66
99ème ctl	1277,29	1305,50	667,91	1203,09	1226,97	599,09	1671,29	1857,45	545,92	436,34
Maximum	1879,70	1305,50	667,91	1645,65	1226,97	599,09	4617,78	1857,45	545,92	466,09
Max/min	781,97	331,31	61,30	493,50	205,32	41,13	224,01	79,19	12,07	6,56

2. Structure de l'échantillon sur les variables de calage

La structure de l'échantillon en mobilisant les poids sans troncature (Poids_s1, poids_s2, poids_s3 et poids_s4) est identique à celle de la cible (recensement), par définition. C'est donc l'effet de la troncature qui est étudié ici en comparant son impact dans les différents scenarios sur l'évolution de la structure.

Ecarts à la cible pour les scenarios avec troncature

	S1_T1	S1_T5	S2_T1	S2_T5	S3_T1	S3_T5
Sexe						
Hommes	-0,48	-0,4	-0,25	-1,22	-5,17	-4,01
Femmes	0,48	0,4	0,25	1,22	5,17	4,01
Classe d'âge						
18-29 ans	-0,04	1,89	0,43	3,53	2,36	2,45
30-39 ans	0,43	3,47	0,54	4,61	2,25	4,3
40-49 ans	-0,39	-1,27	-0,62	-2,72	-2,23	0,7
50-59 ans	-0,27	-2,38	-0,07	-2,59	0,51	-3,35
60 ans et plus	0,27	-1,71	-0,28	-2,82	-2,89	-4,11
Département de résidence						
75	-0,02	-1,33	0,37	1,09	-1,45	1,19
77	0,07	-0,43	-0,23	-0,91	0,38	-1,07
78	-0,03	-0,53	-0,31	-0,77	0,51	-0,42
91	0,08	0,07	0,07	0,03	0,39	1,26
92	0,3	2,57	0,28	2,51	1,53	2,62
93	-0,85	-3,8	-0,57	-4,41	-2,76	-9,04
94	0,36	2,91	0,33	1,88	1,07	3,75
95	0,06	0,53	0,06	0,56	0,32	1,7

Un code couleur a été utilisé pour rendre les écarts plus visibles: plus le vert est intense, plus l'écart est important avec une surreprésentation dans la catégorie concernée dans l'échantillon. De la même façon, plus le rouge est prononcé, plus l'écart est important dans le sens d'une sous-représentation de la catégorie dans l'échantillon. De façon similaire dans les trois scenarios, la troncature des 5% max et 5% min conduit à une plus grande distorsion de l'échantillon. Si l'on compare les trois poids avec troncature des 1% max et 1% min, on constate qu'avec le scenario 3, la structure est davantage modifiée qu'avec les scenarios 1 et 2 (pour lesquels les écarts à la cible restent faibles). A ce stade, les poids S1_T1 et S2_T2 semblent être à privilégier.

3. Structure de l'échantillon sur les variables issues de sources externes non mobilisées dans le calage

Nous avions repéré dans les données du recensement de 2018 quelques autres variables nous permettant de comparer notre échantillon ChIPRe à la population cible. Cependant, ces variables n'ont pas été retenues comme marges car nous doutions de la comparabilité de celles-ci dans les deux sources (questions posées différemment, notamment).

Ces variables sont le statut d'occupation du logement, le plus haut diplôme obtenu & l'année d'arrivée en France.

Ecarts à la cible (recensement) en points de pourcentage

	Scenario 1			Scenario 2			Scenario 3			Scen. 4
	S1	S1_T1	S1_T5	S2	S2_T1	S2_T5	S3	S3_T1	S3_T5	S4
Statut d'occupation du logement										
Propriétaire	-20,9	-20,4	-18,0	-19,9	-19,7	-18,5	-11,3	-11,3	-10,8	-11,2
Locataire ou sous-locataire	18,2	17,6	15,5	17,0	17,0	16,6	10,9	10,4	8,8	9,0
Logé gratuitement	4,1	4,2	3,7	4,0	3,8	2,7	2,0	2,4	3,1	3,3
Autre situation	-1,4	-1,4	-1,2	-1,1	-1,0	-0,8	-1,6	-1,5	-1,1	-1,1
Plus haut diplôme obtenu										
Aucun diplôme	-30,5	-30,5	-32,1	-31,7	-31,6	-32,8	-36,2	-35,9	-34,8	-33,0
Brevet, CEP, CAP, BEP, etc	7,5	7,0	3,9	8,3	7,1	3,4	15,6	8,8	3,5	6,2
Bac	6,4	6,2	6,8	5,1	5,4	5,8	5,2	6,2	5,7	6,7
Bac +2 à bac+5 (modalités 4 et 5)	12,3	12,9	17,4	13,1	14,2	18,9	9,7	15,0	22,1	17,4
Doctorat	4,4	4,4	4,1	5,1	4,9	4,7	5,7	5,9	3,4	2,8
Année d'arrivée en France (arrivée non renseignée ou postérieure à 2018 non pris en compte dans le TAP)										
Avant 1939	-12,3	-12,3	-12,3	-12,3	-12,3	-12,3	-12,3	-12,3	-12,3	-12,3
1939-1988	-4,7	-4,5	-4,3	-4,5	-4,4	-4,8	0,2	-2,9	-4,7	-2,7
1989-1998	-1,5	-2,2	-5,5	-0,4	-1,4	-5,4	-8,1	-7,0	-7,8	-4,6
1999-2008	-0,7	0,1	2,7	-1,0	-0,4	0,6	12,1	10,8	9,5	6,4
2009-2013	7,1	7,5	8,4	7,2	7,2	9,0	3,5	4,7	6,6	5,0
2014-2018	12,0	11,5	11,0	11,1	11,3	12,9	4,6	6,7	8,7	8,2

Pour le statut d'occupation, quel que soit le scenario et la troncature retenus, les propriétaires sont largement sous-représentés dans l'échantillon ChIPRe et les locataires largement surreprésentés. Les scenarios 3 et 4 offrent cependant des résultats moins contrastés.

Pour le niveau de diplôme, les écarts sont très forts sur les personnes non diplômées, très largement sous-représentés dans ChIPRe. Cette fois ci, c'est le scénario 1 qui offre les résultats les moins éloignés du recensement).

Pour l'année d'arrivée en France, l'échantillon ChIPRe souffre d'une sous-représentation des arrivées avant 1999, et d'une surreprésentation des arrivées récentes.

Quel que soit le poids retenu, les écarts entre l'échantillon ChIPRe et le recensement restent très marqués et ne peuvent pas vraiment nous aider à trancher sur un poids plutôt qu'un autre.

Comme mentionné précédemment, nous avons également mobilisé les données issues de l'application de gestion des dossiers des ressortissants étrangers en France (AGDREF) pour comparer notre échantillon ChIPRe avec la population cible. La base AGDREF ne couvrant pas en intégralité le champ de l'enquête ChIPRe, nous avons défini un sous-échantillon parmi les répondants ChIPRe.

Seules ont été conservées, pour la comparaison, les personnes étant entrées en France pour la 1ere fois entre 1999 et 2018 et ayant un visa de touriste, un visa étudiant, un visa professionnel ou encore un passeport ou une carte d'identité d'un autre pays européen. 340 répondants sur les 501 de l'échantillon vérifient ces conditions.

Ecarts à la cible (AGDREF) en points de pourcentage

	Scenario 1			Scenario 2			Scenario 3			S4
	S1	S1_T1	S1_T5	S2	S2_T1	S2_T5	S3	S3_T1	S3_T5	S4
Sexe										
Hommes	-0,3	-0,4	-1,0	-1,4	-1,4	-2,0	-8,5	-9,0	-6,4	-4,6
Femmes	0,3	0,4	1,0	1,4	1,4	2,0	8,5	9,0	6,4	4,6
Classe d'âge										
18-29 ans	10,7	10,3	10,4	11,2	11,2	11,9	9,1	9,8	7,4	6,7
30-39 ans	-1,5	-0,8	1,8	0,7	0,7	1,9	0,8	0,6	2,1	-0,1
40-49 ans	-4,2	-3,9	-4,5	-3,0	-3,0	-5,3	-7,5	-7,2	-3,9	-3,4
50-59 ans	-1,0	-1,8	-3,9	-4,9	-4,9	-4,5	2,2	1,4	-1,5	-0,2
60 ans ou +	-3,9	-3,9	-3,8	-4,0	-4,0	-3,9	-4,6	-4,6	-4,1	-3,0
Situation matrimoniale										
Célibataire	-14,7	-14,7	-14,7	-14,7	-14,7	-15,1	-18,2	-18,4	-20,1	-21,8
Marié	9,3	9,1	9,2	8,6	8,6	8,6	2,4	3,3	10,8	13,1
Divorcé	3,7	3,8	3,5	3,3	3,3	3,6	4,2	4,3	3,9	3,7
Veuf	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	0,1	0,1	0,3	0,6
Autre situation (Pacs)	2,2	2,3	2,4	3,2	3,2	3,3	11,6	10,7	5,0	4,3
Département (Gestionnaire pour AGDREF, de résidence pour ChIPRe										
75	-14,5	-14,5	-15,9	-14,5	-14,2	-13,4	-14,5	-16,0	-13,3	-14,5
77	1,4	1,5	1,0	1,4	1,2	0,5	1,4	1,8	0,3	1,4

78	0,2	0,2	-0,3	0,2	-0,1	-0,6	0,2	0,7	-0,2	0,2
91	-0,8	-0,7	-0,7	-0,8	-0,7	-0,7	-0,8	-0,4	0,5	-0,8
92	0,9	1,2	3,5	0,9	1,2	3,4	0,9	2,4	3,5	0,9
93	10,3	9,4	6,5	10,3	9,7	5,8	10,3	7,5	1,2	10,3
94	2,9	3,3	5,8	2,9	3,3	4,8	2,9	4,0	6,7	2,9
95	-0,4	-0,3	0,1	-0,4	-0,3	0,2	-0,4	-0,1	1,3	-0,4

Sur le sexe, les écarts sont assez restreints avec les scenarios 1 et 2, plus importants avec une surreprésentation des femmes dans ChIPRe pour les scenarios 3 et 4.

Sur la classe d'âge, nous observons pour toutes les pondérations, une surreprésentation des moins de 30 ans, bien qu'un calage sur la distribution par âge du recensement 2018 ait été fait. Cette surreprésentation résiduelle des moins de 30 ans parmi les demandeurs de titres entre 1999 et 2018 est très certainement due à notre excédent d'étudiants parmi les répondants ChIPRe.

Pour la situation matrimoniale, nous avons un déficit de célibataires et un excédent de personnes mariées dans notre échantillon, ce qui semble contradictoire avec l'effet d'âge. Cependant, dans la base AGDREF, il s'agit du statut à l'entrée alors que dans ChIPRe, c'est le statut actuel qui est renseigné. Ceci peut expliquer les différences observées.

Concernant le département, il n'est pas vraiment comparable dans les deux sources. Dans AGDREF, il s'agit du département gestionnaire de la demande de titres de séjour (à l'entrée en France) alors que dans ChIPRe, il s'agit du département actuel de résidence. Il est donc difficile de conclure sur les biais éventuels.

4. Structure de l'échantillon par rapport au réseau

Dans cette partie, l'objectif est de voir quel est l'impact de la correction de la non-réponse totale par GRH sur la structure de l'échantillon. Dans les scenarios 2 et 4, une correction de la non-réponse par modélisation et GRH n'a pas été faite, contrairement aux scenarios 1 et 3. L'objectif ici est donc de voir si les scenarios 1 et 3, intégrant une phase de CNRT par GRH, offrent une structure d'échantillon plus proche du réseau ChIPRe ou si le calage annule cet effet.

Ecarts au réseau en points de pourcentage

	Scenario 1			Scenario 2			Scenario 3			S4
	S1	S1_T1	S1_T5	S2	S2_T1	S2_T5	S3	S3_T1	S3_T5	S4
Sexe										
Homme	3,8%	3,3%	3,4%	3,7%	3,5%	2,5%	4,0%	-1,1%	0,0%	3,7%
Femme	-3,7%	-3,2%	-3,2%	-3,6%	-3,3%	-2,3%	-3,9%	1,3%	0,1%	-3,6%
REFUS/NSP	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%	-0,1%
Classe d'âge										
18-30 ans	-15,1%	-15,1%	-12,7%	-14,8%	-14,4%	-11,0%	-16,3%	-13,8%	-13,0%	-15,5%
31-40 ans	-8,3%	-8,4%	-7,1%	-10,6%	-10,1%	-7,4%	-6,1%	-3,7%	-1,8%	-8,5%
41-50 ans	9,0%	8,5%	7,3%	9,0%	8,3%	6,4%	7,9%	5,2%	6,3%	7,2%
51-60 ans	8,0%	8,2%	6,1%	9,0%	8,6%	5,2%	12,0%	9,2%	5,0%	8,2%
Plus de 60 ans	3,7%	3,9%	3,2%	3,7%	3,8%	2,7%	1,5%	1,9%	1,5%	3,3%

	Scenario 1			Scenario 2			Scenario 3			S4
	S1	S1_T1	S1_T5	S2	S2_T1	S2_T5	S3	S3_T1	S3_T5	S4
REFUS/NSP	2,8%	3,0%	3,2%	3,6%	3,5%	3,6%	-1,8%	-1,3%	0,9%	3,9%
INFOS DISCORDANTES	-0,1%	-0,1%	0,0%	0,2%	0,2%	0,5%	2,8%	2,6%	1,1%	1,3%
Province d'origine en Chine										
Province=1	-0,4%	-0,2%	0,6%	-2,0%	-1,9%	-1,0%	-1,2%	-0,3%	1,0%	-0,8%
Province=2	-1,5%	-1,4%	-1,4%	-0,1%	-0,3%	-0,4%	0,0%	1,1%	-0,3%	0,1%
Province=3	0,0%	0,1%	-0,1%	0,9%	1,0%	0,5%	-2,9%	-2,7%	-1,6%	-0,3%
Province=4	0,4%	0,1%	-0,8%	-1,1%	-1,3%	-1,4%	-2,3%	-1,5%	-0,3%	-1,6%
Province=5	6,6%	6,0%	5,5%	5,2%	5,0%	3,9%	12,7%	7,1%	6,6%	7,7%
Province=6	3,5%	3,3%	2,7%	3,6%	3,9%	3,2%	-0,8%	0,3%	0,4%	0,9%
Province=7	-0,7%	-0,7%	-0,3%	-1,5%	-1,4%	-1,1%	-2,0%	-1,8%	-1,2%	-1,8%
Province=8	-3,9%	-3,8%	-2,8%	-3,0%	-2,9%	-1,9%	-3,1%	-3,1%	-3,4%	-3,2%
Province=9	-2,3%	-2,3%	-1,9%	-1,9%	-1,9%	-1,4%	0,0%	0,6%	-1,1%	-2,0%
Province=10	-0,5%	-0,5%	-0,5%	-0,5%	-0,5%	-0,5%	-0,4%	-0,3%	-0,2%	-0,2%
Province=11	-0,5%	-0,4%	-1,0%	-0,5%	-0,4%	-1,2%	-2,0%	-1,9%	-1,4%	-1,4%
Province=12	0,1%	0,3%	0,0%	0,9%	0,7%	0,4%	6,7%	6,6%	2,1%	1,1%
Province=Refus/nsp	0,0%	0,2%	0,8%	0,6%	0,6%	1,4%	-4,1%	-3,3%	0,2%	2,1%
Province=infos discordantes	-0,7%	-0,7%	-0,6%	-0,6%	-0,6%	-0,5%	-0,7%	-0,7%	-0,6%	-0,5%
Durée de résidence en France										
Durée_Fr=0-2 ans	-1,2%	-1,5%	-2,2%	-1,9%	-1,7%	-1,5%	-4,1%	-3,4%	-2,6%	-1,8%
Durée_Fr=3-10 ans	-12,0%	-11,9%	-10,2%	-12,6%	-12,3%	-9,8%	-14,1%	-11,9%	-10,4%	-13,5%
Durée_Fr=11-20 ans	5,5%	4,9%	3,9%	3,5%	2,9%	2,8%	7,3%	9,4%	7,3%	2,1%
Durée_Fr=21+	3,7%	4,0%	3,6%	5,6%	5,9%	3,6%	9,7%	7,6%	4,5%	7,4%
Durée_Fr=Refus/nsp	3,8%	4,2%	4,6%	5,8%	5,6%	5,1%	-5,3%	-4,3%	0,1%	5,8%
Durée_Fr=infos discordantes	0,3%	0,3%	0,3%	-0,4%	-0,4%	-0,3%	6,5%	2,6%	1,0%	0,1%

Conclusion

Au vu des résultats présentés dans cette dernière partie, nous avons choisi de retenir la pondération POIDS_S2_T1 pour l'exploitation de l'enquête. Cependant, cette pondération pose toujours des problèmes et une attention particulière doit être prise lors de son utilisation : le rapport min/max reste important (ce qui peut dégrader la précision des estimateurs), et des écarts par rapport à la structure du recensement de la population (sur la part de propriétaires par exemple) persistent, ce qui semble montrer un écart de structure.

Annexes

Recodage des variables disponibles dans le roster pour modéliser la non-réponse totale

	Modalités de la variable dans le roster	Modalités de la variables dans la modélisation de la participation
Sexe	1 Homme 2 Femme 88 Refus 99 NSP	1 Homme 2 Femme 3 Refus/NSP
Classe d'âge	1 18-25 ans 2 26-30 ans 3 31-35 ans 4 36-40 ans 5 41-45 ans 6 46-50 ans 7 51-55 ans 8 56-60 ans 9 61-65 ans 10 Plus de 65 ans 88 Refus 99 NSP	0 Infos discordantes selon citants 1 18-30 ans 2 31-40 ans 3 41-50 ans 4 51-60 ans 5 Plus de 60 ans 6 Refus/NSP
Province d'origine en Chine	1 Anhui 2 Beijing 3 Chongqing 4 Fujian 5 Gansu 6 Guangdong 7 Guangxi 8 Guizhou 9 Hainan 10 Hebei 11 Heilongjiang 12 Henan 13 Hubei 14 Hunan 15 Hong Kong 16 Jiangsu 17 Jiangxi 18 Jilin 19 Liaoning 20 Macao 21 Nei Menggu 22 Ningxia 23 Qinghai 24 Shandong 25 Shanghai 26 Shanxi 27 Shaanxi 28 Sichuan 29 Tianjin	0 Infos discordantes selon citants 1 Heilongjiang (11) / Jilin (18) / Liaoning (19) 2 Beijing (2)/ Shanghai (25) / Tianjin (29) 3 Hebei (10) / Shanxi (26) 4 Jiangsu (16) / Shandong (24) 5 Zhejiang (33) 6 Fujian (4) / Guangdong (6) / Hainan (9) / Hong Kong (15) / Macao (20) 7 Gansu (5) / Nei Menggu (21) / Ningxia (22) / Shaanxi (27) 8 Anhui (1) / Henan (12) / Hubei (13) 9 Chongqing (3) / Sichuan (28) 10 Qinghai (23) / Tibet (30) / Xinjiang (32) 11 Guangxi (7) / Guizhou (8) / Yunnan (31) 12 Hunan (14) / Jiangxi (17) 13 Refus/NSP

	30 Tibet 31 Yunnan 32 Xinjiang 33 Zhejiang 88 Refus 99 NSP	
Durée de résidence en France	1 Moins d'un an 2 De 1 à 2 ans 3 De 3 à 5 ans 4 De 6 à 10 ans 5 De 11 à 15 ans 6 De 16 à 20 ans 7 De 21 à 25 ans 8 Plus de 25 ans 88 Refus 99 NSP	0 Infos discordantes selon citants 1 Moins de 3 ans 2 De 3 à 10 ans 3 De 11 à 20 ans 4 Plus de 20 ans 5 Refus/NSP

Variables construites à partir de la base des rsoters et mobilisées dans la modélisation de la participation

Variable	Modalités
Multicitation	0 Une seule citation 1 Plusieurs citations
Cite_par_graine	0 Non 1 Oui
Tel_dispo	0 Aucun numéro renseigné 1 Numéro de téléphone renseigné
WC_dispo	0 Aucun compte wechat renseigné 1 Compte WeChat renseigné
Cite_par_ami_proche	0 Aucun des citants ne considère le cité comme un ami proche 1 Un des citants considère le cité comme un ami proche
Cite_par_famille	0 Aucun des citants n'est un membre de la famille du cité 1 Un des citants est un membre de la famille du cité
Rel_non_def	0 Aucun des citants n'a su définir la relation qui le lie au cité 1 Un des citants n'a pas su définir la relation qui le lie au cité
Enq_exploit	11 Enq 11 12 Enq 12 13 Enq 13 14 Enq 14 15 Enq 15 16 Enq 16 17 Enq 17 18 Enq 18 19 Enq 19 20 Enq 20

Bibliographie

[1] Mouw T, Verdery AM. « Network Sampling with Memory: A proposal for more efficient sampling from social networks». Sociol Methodol. 2012 Aug;42(1):206-256.
[2] Goel S, Salganik M.J, «Assessing respondent-driven sampling», PNAS April 13, 2010 107 (15) 6743-6747
[3] Merli M G, Verdery A., Mouw T, Li J., «Sampling Migrants from their Social Networks: The Demography and Social Organization of Chinese Migrants in Dar es Salaam, Tanzania», Migr Stud. 2016 Jul;4(2):182-214.
[4] Mouw T, Chavez S, Edelblute H, Verdery A., «Binational Social Networks and Assimilation: A Test of the Importance of Transnationalism», Social Problems Vol. 61, No. 3 (August 2014), pp. 329-359
[5] Merli G, Mouw T, Stolte A, Le Barbenchon C, Florey-Eischen F, Using Multiple Modes of Data Collection to Recruit Migrant Samples With Network Sampling With Memory: The Chinese Immigrants in Raleigh-Durham (ChIRDU) Study, 2019 Annual Meeting, Austin, TX April 10-13