Compter les absents : estimer indirectement une diaspora insulaire

Clément Digoin, Eva Lelièvre, Thomas Merly-Alpa, Celio Sierra-Paycha

DOCUMENTS DE TRAVAIL 298

Compter les absents : estimer indirectement une diaspora insulaire

Clément Digoin, Éva Lelièvre, Thomas Merly-Alpa et Célio Sierra-Paycha

Septembre 2024

ined
INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES

https://doi.org/10.48756/ined-dt-298.0924

COMPTER LES ABSENTS : ESTIMER INDIRECTEMENT UNE DIASPORA INSULAIRE

Clément Digoin[1], Éva Lelièvre[2], Thomas Merly-Alpa[23] & Célio Sierra-Paycha[24]

[1] *ENSAI, France, clementdigoin@gmail.com*
[2] *Ined, France, eva.lelievre@ined.fr*
[3] *Insee, France, thomas.merly-alpa@insee.fr*
[4] *Cridup, Université Panthéon Sorbonne, France, celio.sierra-paycha@univ-paris1.fr*

Résumé. L'objectif de cette étude est d'évaluer le degré de fiabilité de l'estimation de la diaspora insulaire à partir de l'information obtenue sur le lieu de résidence des absents apparentés (parents, frères et sœurs et enfants) des répondants. Cette étude est conduite à partir des données de la première enquête Famille de la Polynésie française (*Feti'i e fenua*, ISPF-Ined, 2020). Ces estimations vont dépendre du calcul des poids de chaque déclaration : un même individu peut être cité par plusieurs membres de sa famille interrogés dans l'enquête *Feti'i e fenua*. En utilisant et en adaptant la Méthode Généralisée de Partage des Poids (MGPP) proposée par Deville et Lavallée (2006) une série d'estimations de la diaspora polynésienne dans l'Hexagone est initialement conduite. Puis disposant des données du recensement 2019 de l'Hexagone qui permettent de cerner la population née en Polynésie vivant en France métropolitaine, il est alors possible de confronter les estimations obtenues via la MGPP et de discuter de la nature des biais observés. Un calage simultané (Caron et Sautory, 2004) calculé sur les répondants en Polynésie et les résidents polynésiens en France métropolitaine est alors proposé pour améliorer les poids de l'enquête.

Mots-clés. Estimation, méthode généralisée du partage des poids, calage simultané.

Abstract. The objective of this paper is to assess the degree of reliability of estimating the diaspora of an insular diaspora using information on the place of residence of the related absentees (parents, siblings, and children) of respondents. This study is conducted using data from the first French Polynesia Family Survey (*Feti'i e fenua*, ISPF-Ined, 2020). These estimates depend on the calculation of the weights of each statement: the same individual can be cited by several members of the family interviewed in the *Feti'i e fenua* survey. Using and adapting the Generalized Weight Sharing Method (GWSM) proposed by Deville and Lavallée (2006), we initially conduct a series of estimates of the Polynesian diaspora in France. Then, relying on the 2019 French census data (mainland France) which allow to identify the Polynesian population living in Metropolitan France, it is then possible to compare the estimates obtained via the GWSM and discuss the nature of the observed biases. A simultaneous calibration (Caron and Sautory, 2004) calculated on the respondents in Polynesia and the Polynesian residents in Metropolitan France is finally adopted to improve the weighting of the survey.

Keywords. Estimation, Generalized Weight Sharing Method, Simultaneous Calibration.

Introduction

Les petits pays insulaires sont caractérisés par deux spécificités : leur population est peu nombreuse et leur diaspora, au sens de natifs résidant en dehors de l'archipel ou de l'île d'origine, relativement importante (Poirine, 1994 ; Bertram, 1999 ; Rallu et coll., 2010). En raison de l'ampleur de l'émigration, les politiques publiques et la recherche gagneraient à disposer de données quantitatives fiables sur les dimensions et la structure de ces diasporas. Toutefois, la petite taille des populations d'origine et *a fortiori* des populations migrantes ne garantit pas une bonne visibilité statistique dans les pays d'immigration (Sierra-Paycha, 2022). Si l'information sur les pays de naissance est bien collectée dans la plupart des recensements ou registres dans les pays d'accueil (Walmsley, Ahmed, et Parsons, 2007 ; Özden, Parsons, Schiff et Walmsley, 2011), la façon dont est construite cette donnée peut conduire à une invisibilisation statistique des diasporas insulaires. En effet, certains petits Etats insulaires sont en situation d'autonomie politique relative vis-à-vis de plus grands Etats comme par exemple la Polynésie française avec la France, Palau avec les Etats-Unis ou les îles Cook avec la Nouvelle-Zélande. Cela a pour effet de fondre les effectifs de personnes originaires de ces pays insulaires dans des agrégats nationaux de rang supérieur : natifs de France, des Etats-Unis ou de Nouvelle-Zélande pour reprendre l'exemple cité plus haut. Ainsi, dans la base bilatérale des migrations rassemblant pour chaque pays d'origine la distribution des ressortissants dans tous les pays du monde à partir des statistiques dans les pays de résidence (Parsons et coll., 2007), ne figurent ni la Polynésie française, ni les îles Cook. Quant à Palau, la base indique moins de 500 natifs résidant à l'extérieur pour l'année 2020 ce qui paraît bien peu, même pour une population nationale d'à peine quelque dizaines de milliers d'habitants. En raison de la taille très faible des effectifs des diasporas insulaires, pour des raisons d'anonymisation ou de facilités de codification, les instituts statistiques des pays d'accueil des diasporas ont également

tendance à agréger dans un échelon national supérieur les catégories liminales avec la modalité principale. Cette invisibilisation qui s'explique par les faibles effectifs et par l'association politique à un Etat plus puissant, est d'autant plus problématique dans les contextes insulaires car l'émigration y fait très souvent partie intégrante d'un modèle de développement. Cette place prépondérante de l'émigration a permis de qualifier ce modèle par l'acronyme MIRAB qui signifie *Migration, Remittances, Aids and Bureaucracy* (Bertram et Watters 1984, 1985, 1986 ; Bertram 1986, 1993, 2004).

Cet article est une expérimentation visant à tester la possibilité d'étudier les émigrants, les absents, à partir de données d'enquête sur les membres de la famille présent dans le pays insulaire d'origine. Faire un sondage indirect pour compter les absents permet en effet d'estimer l'émigration d'une région dont l'économie dépend des transferts, des possibles retours. Cette question importante de l'estimation de la diaspora demande néanmoins à être validée. L'article cherchera à évaluer dans quelle mesure une enquête qui collecte la localisation des absents apparentés auprès des enquêtés permet une estimation de l'ensemble des absents et sa marge d'erreur. Pour cette démonstration, nous étudierons le cas des natifs de Polynésie française résidant dans l'Hexagone dont on dispose par ailleurs de données exceptionnellement fiables grâce au recensement mené par l'Insee (Institut National de la Statistique et des Études Économiques) qui collecte des informations détaillées sur les lieux de naissance des population ultramarines.

La première étape de cette étude consiste à évaluer la taille des natifs de Polynésie résidant dans l'Hexagone à partir de la première enquête Famille (*Feti'i e fenua*, ISPF-Ined, 2020) collectée sur ce territoire ultramarin en utilisant la Méthode Généralisée de Partage des Poids (MGPP) proposée par Lavallée (1995). Puis, en procédant à la comparaison des estimations obtenues avec les données censitaires de France métropolitaine (RP-2019) dénombrant les personnes

nées en Polynésie, nous détaillons les biais et adaptons la MGPP aux spécificités des données de *Feti'i e fenua*. L'analyse des pyramides des âges, issues de la MGPP nous montre que la structure des estimateurs obtenus diverge des valeurs du recensement. C'est pourquoi, nous analysons la dispersion des poids pour tenter de se recaler sur les marges du recensement 2019. Afin d'améliorer les estimations, nous procédons enfin à un calage simultané sur les poids des répondants, prenant des marges sur la population de Polynésie française et sur les Polynésiens vivant en France métropolitaine.

1 Présentation des données utilisées

1.1 L'enquête *Feti'i e fenua* (ISPF-Ined, 2020) en Polynésie française

L'enquête *Feti'i e fenua* (ISPF-Ined, 2020), littéralement « liens et territoire », a été menée en Polynésie française en 2019 et 2020 par l'Ined et par l'Institut de la Statistique de la Polynésie française (ISPF). Cette enquête s'inspire de l'enquête *Famille et logements* (2011) de l'Insee en France métropolitaine. Elle a permis de recueillir différentes informations concernant l'organisation familiale des Polynésiens, les relations entre les membres d'une même famille et leur mobilité au sein du territoire et au-delà (Fardeau et Lelièvre, 2021 ; Fardeau et coll., 2020).

Elle recueille la composition et la dispersion spatiale des familles à travers les lieux de résidences de leurs membres, y compris en dehors du territoire, ainsi que leurs occupations et les liens qu'ils entretiennent avec les lieux d'origine (lieu de naissance, île ou région de socialisation avant 6 ans, lieux de scolarité, etc.). La collecte s'est déroulée sur l'ensemble des cinq archipels dans 31 îles[1] (sur les 74 habitées) auprès d'individus âgés de 40 à 59 ans[2] tirés

[1] https://www.ispf.pf/ispf/enquete/FeF

[2] Ils constituent les « adultes pivots »[2] entre des parents vieillissants et des enfants en âge de décohabiter. Le périmètre de la famille qui fait l'objet du recueil porte donc sur 3 générations.

au hasard dans l'échantillon des ménages issus du recensement de 2017. Le plan de sondage est bâti sur une classification des îles, basée sur le nombre de ménages éligibles résidant sur l'île (un ménage est éligible s'il compte au moins un individu qui a atteint un âge compris entre 40 à 59 ans lors du recensement de 2017), et la présence d'un aéroport pour les plus petites îles. Par ailleurs, les îles comptant moins de 20 ménages éligibles ont été exclues de l'enquête. Chaque île de Polynésie française appartient ainsi à une strate (tableau 1.1)[3]. Sur 5 964 ménages échantillonnés, l'information sur les familles de 5 139 répondants a été collectée, ce qui correspond au remarquable taux de réponse de 86%. Compte tenu du plan de sondage, la pondération a été calculée en 3 étapes, que sont les poids de sondage, la correction de la non-réponse totale et le calage sur marges.

Tableau 1.1 : Strates du plan de sondage de l'enquête *Feti'i e fenua* (Ined-ISPF, 2000).

	Nombre d'îles	Nombre d'îles sélectionnées	*Taux de sondage*
Iles comptant plus de 500 ménages éligibles	7	7	100%
Iles comptant entre 200 et 500 ménages éligibles	6	6	100%
Iles comptant moins de 200 ménages éligibles et équipées d'un aéroport	16	7	43,75%
Iles comptant moins de 200 ménages éligibles et non équipées d'un aéroport	44	11	25%
Total	73	31	42,47%

Le questionnaire est individuel, une seule personne (âgée de 40 à 59 ans) est enquêtée par ménage qui fournit les informations concernant la situation de chacun sur trois générations : parents, frères et sœurs, enfants ainsi que, le cas échéant, famille de son conjoint. On a ainsi le portrait spatialisé de ces familles qui résulte et résume les migrations effectuées (univers familial présenté en figure 1.1).

[3] Pour les deux strates comptant les îles de moins de 200 ménages éligibles, nous avons opté pour un choix raisonné. Cependant, pour le calcul des poids de sondage, nous nous sommes replacés dans une situation d'un sondage aléatoire d'îles, en utilisant le nombre d'îles sélectionnées et le nombre d'îles dans la strate pour pondérer les îles sélectionnées

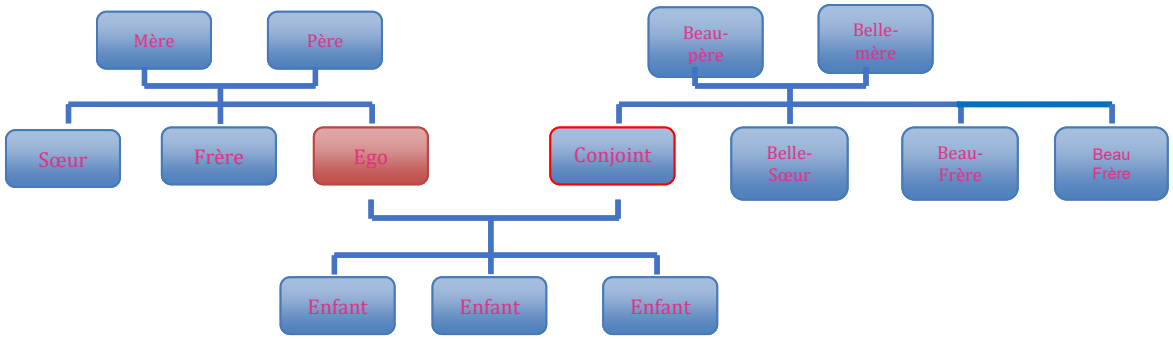


Figure 1.1 : l'univers familial décrit par les répondants de l'enquête *Feti'i e fenua* (ISPF-Ined, 2020)

1.2 Les données de recensement France métropolitaine 2019

On mobilisera dans la suite de l'analyse à des fins de comparaison les données du recensement de la population en France métropolitaine de 2019 (Insee) qui résultent des enquêtes annuelles de recensement de 2017 à 2021. Nous utiliserons ici le sous-échantillon des personnes nées en Polynésie française et résidant en France métropolitaine. On ne conserve donc que les individus qui vivent dans l'Hexagone et nés en Polynésie française. Par ce filtre, on obtient un échantillon de 5 530 individus vivant en France métropolitaine et qui ont déclaré être natifs de Polynésie française.

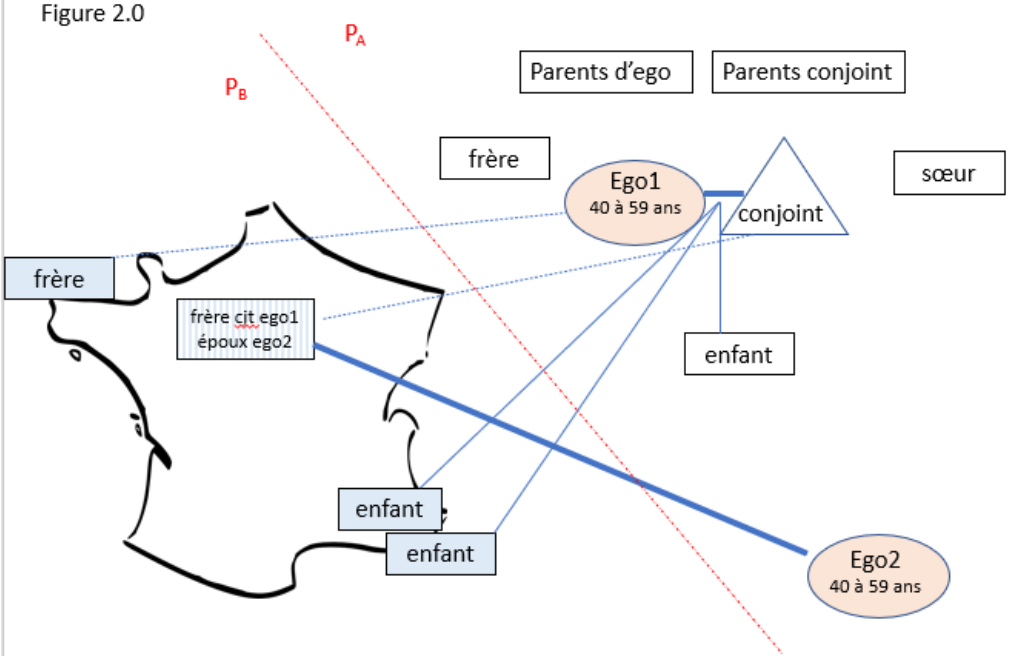
La méthode de recensement en continu employée depuis 2004 repose sur un échantillonnage rotatif de grande taille issu d'un plan de sondage complexe (Godinot, 2005) ; les estimations obtenues sont très précises, notamment en raison du fort taux de sondage (autour de 8%), mais il est nécessaire d'utiliser une pondération afin de rendre l'échantillon représentatif de la population française. La somme pondérée de ces individus conduit à une estimation de 17 278 individus nés en Polynésie française vivant en France métropolitaine. Cette estimation, ainsi que sa répartition selon différentes caractéristiques (pyramides des âges, etc.), seront utilisées comme points de référence pour tester nos méthodes.

2 Le sondage indirect et la méthode généralisée de partage des poids

Pour évaluer la diaspora polynésienne, nous sommes dans le cadre d'un sondage indirect. En effet, on dispose d'un échantillon de la population P^A : celle des personnes vivant en Polynésie française au moment de l'enquête âgées de 40 et 59 ans et l'on cherche à estimer l'effectif de la population P^B soit l'ensemble des individus ayant émigré de Polynésie française vers l'Hexagone. Notons s^A l'échantillon tiré dans la population P^A et π^A_j la probabilité d'inclusion de l'individu j de P^A dans s^A. Si les π^A_j sont connues, les probabilités d'inclusion des individus i dans l'échantillon s^B de P^B sont généralement inconnues, d'autant plus quand les relations entre les individus de la population P^A et P^B ne sont pas bijectives (un individu de P^B n'est pas uniquement relié à un individu de P^A et inversement).

Dans les données de *Feti'i e fenua*, un individu ayant entre 40 et 59 ans (population P^A) peut avoir plusieurs membres de sa famille qui a émigré (population P^B). Inversement, un individu qui a émigré (population P^B) peut être décrit par plusieurs personnes interrogées dans l'enquête (population P^A) (figure 2.0).

Figure 2.0



Ego1 interrogé par l'enquête (population P_A) peut avoir plusieurs membres de sa famille dans l'Hexagone (population P_B). Inversement, un individu qui a émigré (population P_B) peut être décrit par plusieurs personnes interrogées dans l'enquête (population P_A).

La présence de ces liens non bijectifs nous conduit à utiliser la Méthode Généralisée de Partage des Poids (MGPP) comme présentée par J.C. Deville et P. Lavallée en 2006.

Les poids obtenus par la MGPP w_i sont calculés comme

$$\boldsymbol{w_i} = \sum_{j=1}^{n^A} \frac{L_{ij}}{\pi_j^A \, x \, L_i} \quad (1)$$

Où L_{ij} est l'indicatrice de la présence d'un lien entre l'individu i de P^A et j de P^B qui vaut 1 si l'individu j de s^A est lié à l'individu i de s^B, 0 sinon et L_i le nombre total de liens pour un individu i de s^B avec des individus de P^A.

Pour implémenter la MGPP il faut définir ce que nous entendons par "lien" entre les individus de P^A et s^B. En effet, ce choix va conditionner le mode de calcul des indicatrices L_{ij} et du nombre total de liens L_i.

Or, nous ne connaissons pas l'ensemble des liens reliant les individus de s^B aux individus de P^A. En effet, nous ne disposons d'information que sur l'univers familial des répondants (Figure 2.2), et il est possible qu'un individu en métropole présent dans s^B soit relié à d'autres individus de P^A qui n'appartiennent à l'univers familial d'aucun répondant de s^A. Les seules informations dont nous disposons est le lien familial entre les individus de l'ensemble des univers familiaux des répondants de s^A, ensemble que l'on notera F^A, et de s^B ; cela nous permet de connaître les personnes de F^A pouvant citer dans le cadre de l'enquête *Feti'i e fenua* les individus de s^B. Il faut donc approximer P^A par F^A pour compter le nombre de liens pour chaque individu de s^B. Ce problème correspond à celui de la non-réponse de lien dans l'échantillonnage indirect (Xu et Lavallée, 2009) : il conduit à une surestimation des estimations (ici du nombre d'individus) à l'aide de l'échantillonnage indirect.

En résumé, nous considérons les ensembles suivants :

- P^A : L'ensemble des individus ayant entre 40 et 59 ans et vivant en Polynésie française

- P^B : L'ensemble des individus nés en Polynésie française et vivant en France métropolitaine
- s^A : Les individus enquêtés de *Feti'i e fenua*
- s^B : Les individus mentionnés par des individus de s^A nés en Polynésie française et vivant en France métropolitaine
- F^A : L'ensemble des individus ayant entre 40 et 59 ans, vivant en Polynésie française et ayant été enquêté ou mentionné dans l'enquête *Feti'i e fenua*.

Ils vérifient $s^A \subset F^A \subset P^A$.

2.1 Exemple d'une famille dans *Feti'i et fenua* et notations liées

On suppose que tous les membres de la famille de Bertrand sont nés en Polynésie française.

- s^A = Bertrand
- s^B = Paul
- F^A = Bertrand, Emilie, Carole

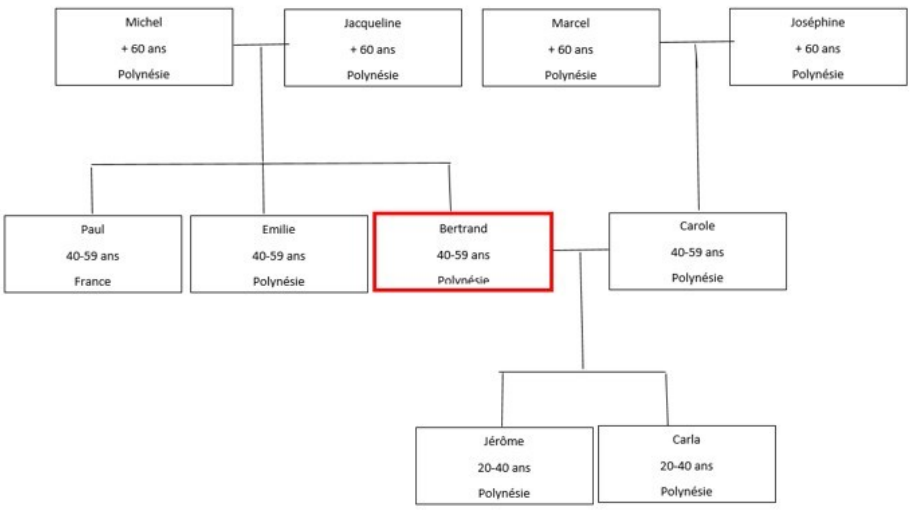


Figure 2.1 : l'univers familial de Bertrand (cas-exemple)

2.2 Méthode naïve

On va donner à chaque individu de s^B le poids de l'individu de s^A qui l'a cité.

En d'autres termes, le lien qui lie les individus de F^A et de s^B est le suivant : un individu de s^B est lié à un individu de F^A, si et seulement si l'individu i de s^B est cité par l'individu j de

F^A dans *Feti'i e fenua*. La valeur de ce lien sera de 1 si cela est vrai, 0 sinon. Le nombre total de liens L_i est pour un individu cité égal est au nombre d'individus de s^A l'ayant cité comme faisant partie de son univers familial.

L'équation (1) est toujours valide, même si au sens strict on ne partage pas le poids de l'individu de s^A, ou du moins, on ne le partage qu'en un.

Dans le cas particulier où l'individu i n'est cité que par un seul individu j, son poids peut se réécrire : $w_i = 1 / \pi^A_j$

Si cette première méthode naïve est plutôt simple, elle ne donne pas de résultats satisfaisants. En effet, on obtient après pondération 27 729 individus nés en Polynésie vivant en France métropolitaine, soit un écart de 9 951 avec les effectifs du recensement (17 278 personnes). Cette méthode conduit à surestimer grandement le nombre d'individus vivant en France métropolitaine. Cette surestimation s'explique par la définition de lien que nous avons donnée : ne prendre en compte que les individus de s^A pour compter le nombre de liens conduit à augmenter artificiellement le poids de chaque individu de s^B.

Une autre façon de le voir est de dire que cette hypothèse conduit à une forte non-réponse de liens (Xu et Lavallée, 2009), ce qui implique un biais de surestimation important ; nous pouvons également le voir comme une mauvaise standardisation de la matrice de liens lié à la méconnaissance du nombre total de liens (Medous et al., 2023). Pour améliorer nos résultats, nous améliorons la prise en compte des liens familiaux par la méthode généralisée du partage des poids.

2.3 Méthode inclusive tenant compte de l'univers familial et des fratries recomposées

Puisque la première méthode conduisait à une surestimation importante, on va raffiner

l'identification des liens entre les apparentés.

La définition du lien qui lie les individus ne change pas ; mais ici, nous allons considérer l'ensemble des individus de F^A pour calculer le nombre total de liens L_i. Pour cela, nous évaluons pour chaque individu de F^A et de s^B la présence d'un lien de la façon suivante : un individu de s^B est lié à un individu de F^A, si et seulement si dans le cas où l'individu j de F^A aurait participé à l'enquête *Feti'i e fenua*, il aurait mentionné l'individu i de s^B. La valeur de ce lien sera de 1 si cela est vrai, 0 sinon.

Dans la famille de Bertrand (Figure 2.1), trois personnes ont entre 40 et 59 ans (Bertrand, Emilie et Carole) et vivent en Polynésie Française : elles sont donc éligibles à l'enquête.

Si Emilie était interrogée, elle citerait Paul car c'est son frère.

Si Carole était interrogée, elle citerait Paul car c'est son beau-frère.

- ⇨ Donc Paul à trois liens au total ($L_i = 3$). Si Paul n'est mentionné par aucun autre individu de s^A, son poids vaut donc $w = 1/(3\pi)$ avec π la probabilité d'inclusion de Bertrand dans l'enquête

La prise en compte de ces liens permet de réduire la non-réponse de liens. Il reste cependant à gérer quelques cas spécifiques. Par exemple, dans les cas où figurent aussi dans l'univers familial de la personne interrogée (respectivement son conjoint) des demi-frères et des demi-sœurs, il est difficile de savoir avec assurance si ces demi-frères et demi-sœurs de la personne interrogée pourraient citer la personne appartenant à s^B. Une possibilité serait de les exclure : mais dans ce cas, nous réintroduirions du biais de surestimation. Dans ce cas, il est décidé d'allouer un lien de force plus faible à ces individus, conformément à l'article de J.C. Deville et P. Lavallée (2006) : entre ½ (pour les cas où l'individu dans s^B est un parent de la personne interrogée) et 2/3 (pour les cas où l'individu dans s^B fait partie de la fratrie de la personne

interrogée). Le choix de ces coefficients repose sur l'énumération des cas possibles : supposons un enquêté ayant un demi-frère faisant partie de F^A. Pour la mère de l'individu enquêté, deux situations sont possibles : elle est soit la mère également de son demi-frère, soit elle ne l'est pas. Pour la demi-sœur de l'individu enquêté, elle peut partager les deux mêmes parents avec le demi-frère, un seul des deux ou aucun, d'où un choix de pondération du lien plus fort. Ces pondérations sont prises en compte dans le calcul du lien L_{ij} et du nombre total de liens L_i.

En appliquant la MGPP, on obtient après pondération un effectif de 10 240 polynésiens vivant en France métropolitaine. Il manque donc 7 038 individus par rapport au recensement. Sauf mention contraire, on utilisera cette méthode dans la suite de cet article.

3 Comparer les pyramides des âges (Estimation / RP) et identifier les biais

En tenant compte du sexe et de l'âge de chaque individu, on peut construire les pyramides des âges des personnes nées en Polynésie française à partir des absents pondérés calculés à partir de données de *Feti'i e fenua* et des données du recensement de France métropolitaine. La Figure 3.1 montre la superposition de la pyramide des âges issue du recensement (en couleur) et de diaspora estimée par la MGPP (encadrée) en France métropolitaine. Des tranches d'âge de cinq ans sont utilisées pour pouvoir analyser en particulier la population jeune ; pour les âges les plus élevés, l'incertitude des estimations ne permettra pas une analyse très fine.

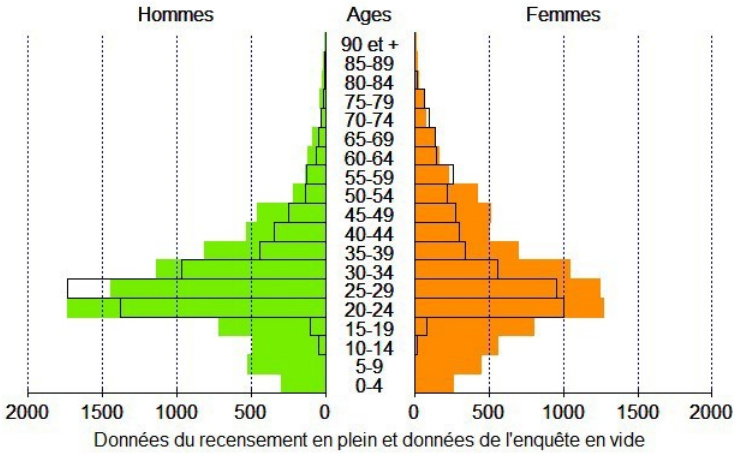


Figure 3.1 – Pyramide des âges des personnes nées en Polynésie française (diaspora estimée comparée au recensement en France métropolitaine)

Cette population est globalement jeune et se concentre autour des 20-35 ans, puis décroît plus l'âge avance. La prépondérance de cette population n'est pas étonnante car un certain nombre de jeunes Polynésiens quittent la Polynésie et rejoignent l'Hexagone dans le cadre de l'armée ou y poursuivent leurs études. La MGPP permet d'obtenir une structure proche de celle du recensement au vu de la figure 3.1, mais de nombreux écarts existent.

La différence principale est l'absence des moins de 20 ans qui ne sont quasiment pas cités dans l'enquête du fait du sondage particulier de *Feti'i e fenua* collectées auprès d'un échantillon représentatif des 40-59 ans. En effet, les jeunes enfants des enquêtés vivent encore avec eux ; réciproquement, les jeunes enfants nés en Polynésie présents en métropole sont souvent partis avec leurs parents. Il convient donc de supprimer ces tranches d'âges de la comparaison :

⇨ En supprimant les moins de 18 ans de notre étude, on obtient 10 151 individus par la méthode MGPP correspondant aux 13 958 personnes recensées dans l'Hexagone.

Les autres biais subsistants sont les suivants :

- La population masculine des 25-29 ans est surestimée (+286) par l'estimation de la diaspora ; cela peut s'expliquer par une sous-estimation des militaires ;
- Les femmes sont sous-estimées notamment chez les 25-40 ans

L'écart entre les populations étant une sous-évaluation du nombre de personnes par la MGPP,

cela nous rassure sur les risques liés à la non-réponse de liens, qui conduisent systématiquement à une surestimation. Ces écarts sont donc potentiellement liés d'une part à une incertitude probabiliste, qui pourrait être estimée à partir du plan de sondage de *Feti'i e fenua*, et d'autre part la présence d'individus dans l'Hexagone n'ayant plus de liens avec la Polynésie.

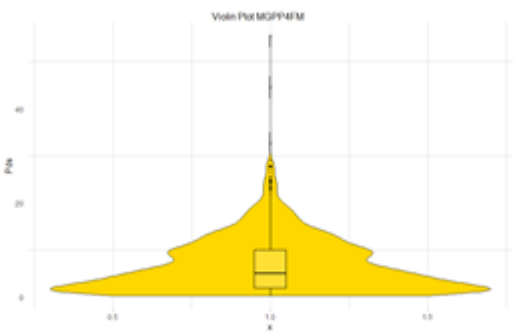
4 Calage sur marge des estimations

Dans cette partie, nous allons mobiliser le fait de disposer des marges issues du recensement de la population en France métropolitaine pour voir si de nouveaux jeux de pondérations peuvent être construits afin de réduire les écarts observés entre les effectifs recensés en France métropolitaine et l'estimation de la diaspora à partir des données de *Feti'i e fenua*. On envisage tout d'abord d'effectuer un calage sur les marges du recensement métropolitain 2019, afin d'améliorer l'écart des estimateurs issus de la MGPP aux valeurs du recensement, puis un calage simultané sur ces marges et sur les données issues du recensement polynésien.

4.1 Dispersion des poids

Avant de pouvoir procéder aux calages, on vérifie à l'aide de *violinplots* la dispersion des poids issus de la MGPP. La Figure 4.1 détaille les statistiques descriptives et le *violinplot* : la boîte noire représente la boîte de Tukey (et les points noirs les valeurs extrêmes de la distribution), tandis que la courbe jaune est la courbe de distribution des poids de la MGPP pour la France métropolitaine. L'avantage principal des *violinplot* est qu'ils permettent d'associer la courbe de distribution d'une variable et ses statistiques descriptives sur un même graphique.

Minimum	0,1706
Q1	1,9090
Médiane	5,0195
Moyenne	6,6409
Q3	9,9135



Maximum	55,4445
Maximum/Minimum	325,0568
Total	10240

Figure 4.1 : Description des poids construits par la MGPP et *violinplot*

L'allure de la courbe de la distribution des poids ne permet pas d'envisager d'effectuer un calage sur marges sans auparavant effectuer une troncature. En effet, pour la France métropolitaine, on constate que la distribution des poids est très étalée et que les rapports maximum/minimum sont supérieurs à 200. Il faut donc ramener tous les poids supérieurs à la valeur de la troncature.

Plusieurs types troncatures sont envisagées : nous présentons en Figure 4.2 les résultats d'une troncature à la VAS (Valeur Adjacente Supérieure), calculé à partir du troisième quartile auquel on rajoute une fois et demi l'écart interquartiles.

Minimum	0,1732
Q1	1,9385
Médiane	5,0972
Moyenne	6,6410
Q3	10,0668
Maximum	22,2594
Maximum/Minimum	128,5129
Troncature initiale	21,92
Total	10 240

Figure 4.2 : Descriptifs des poids MGPP ***après*** troncature

4.2 Marges métropolitaines

Une fois la troncature effectuée, on peut essayer un calage sur marges pour la France métropolitaine.

Le jeu de marges est issu du recensement de la France métropolitaine.

Le jeu de données à caler est celui des Polynésiens vivant en France métropolitaine d'après les

déclarations des enquêtés dans l'enquête *Feti'i e fenua*, obtenu après MGPP, troncature à la VAS et recalibrage.

On utilise quatre variables pour effectuer le calage : le sexe (2 modalités), l'âge (en ne tenant pas compte des moins de 20 ans et des plus de 90 ans, mal cités dans l'enquête), l'activité (2 modalités), la région de résidence (13 modalités).

Plusieurs découpages sont envisagés pour la variable d'âge : tranches quinquennales (qui conduisent à une bonne adéquation aux marges, mais une dispersion des poids importante), décennales (avec le résultat inverse), âge continu. Il est finalement décidé d'utiliser les tranches suivantes : 20-24 ans, 25-29 ans, 30-34 ans, 35-39 ans, 40-49 ans, 50-59 ans et plus de 60 ans.

4.3 Calage simultané

Après la MGPP, les résultats ne sont pas totalement satisfaisants pour la France métropolitaine. En effet, sur la population générale, il nous manque 7 038 individus, et 3 807 en ne prenant en compte que les plus de 18 ans. Les différents scenarii de calage sur marges améliorent certes les résultats, mais il est nécessaire de concilier une structure de population proche du recensement (avec nos pyramides des âges) et une distribution des poids resserrée pour que ces poids soient utilisables *a posteriori*.

Nous allons donc effectuer un calage simultané sur les poids des répondants de *Feti'i e fenua* en prenant des marges sur les habitants de Polynésie française et sur les Polynésiens vivant en France métropolitaine, via les recensements respectifs de ces deux populations. Pour reprendre les concepts théoriques de Caron et Sautory (2004) les unités primaires seraient les répondants de *Feti'i e fenua*, et les unités secondaires les personnes que ces répondants citent comme vivant en France métropolitaine. Ce calage simultané permet aux répondants de *Feti'i e fenua* de

représenter correctement la structure de la population en Polynésie française, mais aussi celle des Polynésiens vivant en France métropolitaine.

En raison du contexte de lien non bijectif entre les deux populations ayant conduit à l'introduction de la MGPP (voir partie 2), nous devons adapter légèrement les variables mobilisées pour le calage simultané. Nous reprenons ici l'approche introduite par Deville (1998) pour l'analyse de panels et le calage simultané à plusieurs dates ; en effet, les variables associées aux unités secondaires ne sont pas entièrement affectées à un individu de la population des unités primaires, comme ce serait le cas pour un lien individu-logement, par exemple, mais sont pondérées en prenant en compte la force du lien de la MGPP. Un individu de s^B pouvant être cité 4 fois par les individus de F^A sera par exemple affecté d'un coefficient de ¼ dans le calage simultané de l'individu de s^A qui l'a cité.

Le calage simultané est réalisé sur les marges suivantes :

- Pour les unités primaires (individus en Polynésie Française), on utilise les marges suivantes, issues du recensement de la population en Polynésie Française : sexe, âge, indicatrice de vie en couple sous le même toit, lieu de naissance, situation vis-à-vis de l'emploi ;
- Pour les unités secondaires, on se limite au sexe et à l'âge (en tranches définies en partie 4.2).

Nous effectuons le calage avec une méthode logit avec comme borne inférieure 0,7 et supérieure 2 (méthode choisie car permettant la meilleure dispersion des poids). Avec ces nouveaux poids pour les répondants à l'enquête, on peut calculer des nouveaux poids pour les Polynésiens vivant en France métropolitaine à partir de l'équation (1).

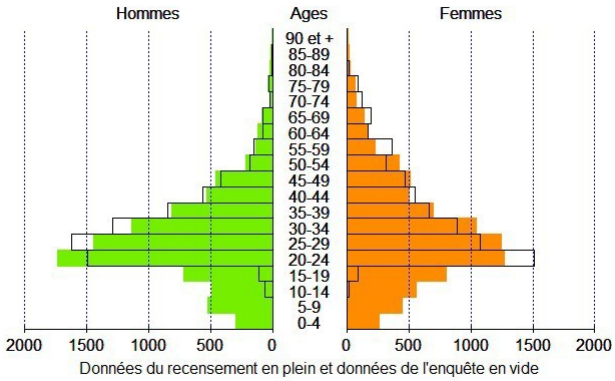


Figure 4.3 : Pyramide des âges comparant les effectifs du recensement 2019 (couleur et plein) aux effectifs d'absents de l'enquête *Feti'i e fenua* (ISPF-Ined, 2020) après calage simultané et MGPP.

Au regard de la pyramide représenté en figure 4.3, les résultats sont très bons pour les plus de 20 ans en termes de structure de population. On surestime quelques classes d'âges mais cette surestimation ne dépasse pas les 300 individus et elle compense la sous-estimation de l'autre sexe. Quelques dissonances sont à noter et des hypothèses peuvent être formulées :

1. La surestimation des 20-24 ans chez les femmes peut éventuellement s'expliquer par une surestimation des étudiantes

2. La surestimation des 25-34 ans chez les hommes peut éventuellement s'expliquer par une surestimation des militaires

3. La surestimation des femmes dans les classes d'âges supérieures à 60 ans pourrait s'expliquer de la façon suivante : les femmes vivent plus longtemps que les hommes, elles ont donc plus de chance d'être citées. Comme le calage cherche à avoir le bon nombre de +60 ans, et qu'il est plus probable de citer des femmes âgées que des hommes, le poids des femmes est augmenté au détriment des hommes.

Conclusion

En conclusion, la méthode que nous avons mise en place, MGPP et calage simultané, pour évaluer la diaspora est plutôt efficace sur l'exemple de la Polynésie française. En effet, la pondération des individus cités dans l'enquête par la MGPP permet d'approximer de façon satisfaisante le nombre de Polynésiens vivant en France métropolitaine.

L'écart constaté avec le recensement, de 7 000 individus en moins pour les absents estimés par les données de l'enquête, est certes relativement important (pour une population totale respective d'environ 17 200) mais imputable en grande partie à la construction de l'enquête. Tout d'abord, si le fait de n'avoir interrogé que les 40-59 ans, ces adultes "pivots", permet de capter une grande partie de la population émigrante, néanmoins, les absents des classes d'âges extrêmes sont difficilement atteignables notamment les plus jeunes. En ne prenant plus en compte les mineurs, notre estimation est grandement améliorée (on passe d'un écart de 7 000 à 3 700) par rapport aux données du recensement.

Le reste de l'écart est explicable en très grande partie par le phénomène que l'on étudie. En effet, la MGPP ne permet que de retrouver les personnes ayant encore des parents proches (et dans la tranche d'âge interrogée) dans le territoire d'origine. Si certains émigrants n'ont plus aucun parent proche (ou dans la tranche d'âge interrogée) dans ce territoire, la MGPP ne permet pas de les retrouver. Ce phénomène explique encore en partie l'écart avec le recensement.

Le caractère encore expérimental de cette méthode d'estimation est une invitation à poursuivre d'autres recherches méthodologiques visant à affiner la connaissance statistique des diasporas insulaires à partir de données indirectes. Pour mieux connaître leur diaspora, d'autres instituts de statistiques nationaux de petits pays insulaires pourraient reproduire cette méthode d'estimation indirecte des caractéristiques de l'émigration. En effet, au-delà de la Polynésie

française, bien d'autres territoires archipélagiques sont soumis au défi de couverture statistique posé par l'émigration et pourraient s'en inspirer, et ce, d'autant plus que la dépendance croissante des populations de petits États à l'égard des continents (de leurs universités et offres d'emploi) et les changements climatiques devraient s'intensifier dans les décennies qui viennent (Cassin, Melindi-Ghidi et Prieur, 2022).

Bibliographie

Bertram, G., (1986). “Sustainable development” in Pacific micro-economies." *World Development*, 14(7), 809-822.

Bertram, G. (1993). "Sustainability, aid, and material welfare in small South Pacific Island economies, 1900–1990." *World Development*, 21(2), 247-258.

Bertram, G. (1999). The MIRAB model twelve years on. *The Contemporary Pacific*, 1(1), 105-138.

Bertram, G. (2004). "On the convergence of small island economies with their metropolitan patrons." *World development*, 32(2), 343-364.

Bertram, G. et Watters, R. (1984). New Zealand and its small island neighbours: A review of New Zealand policy toward the Cook Islands, Niue, Tokelau, Kiribati and Tuvalu. *Institute of public policies working paper*, 1.

Bertram, G. et Watters, R. (1985). The MIRAB economy in South Pacific microstates. *Pacific Viewpoint*, 26(3), 497-519.

Bertram, G. et Watters, R. (1986). The MIRAB process: Earlier analyses in context. *Pacific Viewpoint*, 27(1), 47-59.

Caron, N. et Sautory, O. (2004). Calages simultanés pour différentes unités d’une même enquête. *Document de travail Méthodologie statistique INSEE*, 0403.

Cassin, L., Melindi-Ghidi, P. et Prieur, F. (2022). Confronting climate change: Adaptation vs. migration in Small Island Developing States. *Resource and Energy Economics*, 69, 101301.

Deville, J.-C. (1998). Les enquêtes par panel : en quoi différent-elles des autres enquêtes ? suivi de : Comment attraper une population en se servant d’une autre. *Actes des Journées de Méthodologie Statistique,* INSEE Méthodes, Paris: Insee, 84-85-86, 63-82.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : Les fondements de la méthode généralisée du partage des poids. *Techniques d’enquête*, 38.

Fardeau L., Lelièvre E. et L’équipe ATOLLS. (2021). L’enquête Feti’i e fenua (Enquête Famille, territoire et relations intergénérationnelles en Polynésie française) : Apurement et imputation des données. *Documents de travail*, 262. Article accessible à l’adresse https://www.ined.fr/fr/publications/editions/document-travail/enquete-feti-e-fenua-enquete-famille-territoire-et-relations-intergenerationnelles-en-polynesie-francaise-apurement-et-imputation-des-donnees/

Fardeau, L., Lelièvre, E., Sierra-Paycha, C. (2020). La première enquête Famille en Polynésie française : Feti’i e fenua. *Points études et bilans de la Polynésie française*,1276, 1-4.

Godinot, A. (2005). Pour comprendre le recensement de la population. *Insee Méthodes* [en ligne]. Hors-Série. Article accessible à l’adresse https://www.insee.fr/fr/information/2579979

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A., and Puech, P. (2023). Many-to-one indirect sampling with application to the french postal traffic estimation. TheAnnals of Applied Statistics, 17(1): 838–859.

Özden, Ç., Parsons, C. R., Schiff, M. et Walmsley, T. L. (2011). Where on earth is everybody? The evolution of global bilateral migration 1960–2000. *The World Bank Economic Review*, 25(1), 12-56.

Parsons, C. R., Skeldon, R., Walmsley, T. L. et Winters, L. A. (2007). Quantifying international migration: A database of bilateral migrant stocks. *World Bank Policy Research Working Paper*, (4165).

Poirine, B. (1994). Rent, emigration and unemployment in small islands: The MIRAB model and the French overseas departments and territories. *World development*, 22(12), 1997-2009.

Rallu, J.-L., Rogers G. et Reay-Jones, R. (2010). The Demography of Oceania from the 1950s to the 2000s. *Population*, 65(1), 9-115.

Sierra-Paycha, C. (2022). L'expansion du champ migratoire polynésien au XXIe siècle: le fait d'une jeunesse qualifiée. *L'Espace géographique*, 51(1), 74-94.

Walmsley, T. L., Ahmed, S. A. et Parsons, C. R. (2007). A global bilateral migration data base: skilled labor, wages and remittances. *Global Trade Analysis Project*, 1880, 1-32.

Xu, X. et Lavallée, P. (2009). Traitements de la non-réponse de lien dans l'échantillonnage indirect. *Techniques d'Enquêtes*, 35.