JMS Journées de méthodologie statistique de l'Insee 2018

Évaluation de la robustesse d'une projection de population

Jérôme FABRE, Khaled LARBI (*)

(*) Direction régionale de l'Insee Hauts-de-France

jerome.fabre@insee.fr khaled.larbi@insee.fr

Mots-clés. Projections de population, robustesse, analyse des données, Omphale.

Résumé

Le but de cet article est de proposer une définition du concept de robustesse dans le cadre d'une projection de population obtenue à l'aide d'Omphale. L'introduction de cette définition est nécessaire dans la mesure où Omphale est un outil de projection : le but n'est pas de prédire l'avenir et donc la robustesse d'une projection ne se base pas sur sa qualité prédictive. Une projection Omphale est robuste si elle retranscrit correctement les hypothèses du modèles. Des travaux exploratoires ont été menés et deux tests sont proposés : un test sur l'évolution globale de la structure de la population et un test par classe d'âge.

Abstract

This article aims to provide a definition of the robustness of a population projection from Omphale. Introducing such a definition is necessary because population projections differs from forecasts : one cannot judge the quality of a projection using the predictive accurancy. An Omphale projection is robust if it reflects accurately the input assumptions. After some exploratory works, we introduce two tests : one based on the global evolution and the second one based on age-range evolutions.

Introduction

Omphale est un outil de projections démographiques locales développé à l'Insee. Il fonctionne sur tout zonage communal ou supra communal connexe et d'au moins 50 000 habitants. Les projections locales d'Omphale sont une déclinaison locale des projections nationales de l'Insee ([3]) : un mécanisme de calage permet d'assurer la cohérence entre projections nationales et territoriales. Comme tout modèle de projections de population, Omphale se base sur une observation du présent et du passé récent ainsi que sur un certain nombre d'hypothèses permettant de faire évoluer la situation présente au fil du temps.

1 Constat sur les projections et notion de robustesse des projections Omphale

1.1 Omphale : un outil de projection de population locale selon la méthode par cohorte et composante

La situation de départ d'Omphale est la structure de la population par sexe et âge détaillé du territoire observée dans le recensement de la population 2013, structure qui peut être représentée sous la forme d'une pyramide des âges (Figure 1). Le principe d'une projection de population par la méthode dite par cohorte et composante est de déformer année après année la pyramide des âges du territoire en faisant intervenir quatre types d'évènements : des naissances, des décès, des mobilités résidentielles à l'intérieur du pays et celles avec l'étranger. Les trois premiers événements sont modélisés d'une part à partir des caractéristiques observées de ce territoire et d'autre part prolongés dans le temps sur la base d'hypothèses nationales. Le traitement des mobilités internationales est spécifique.

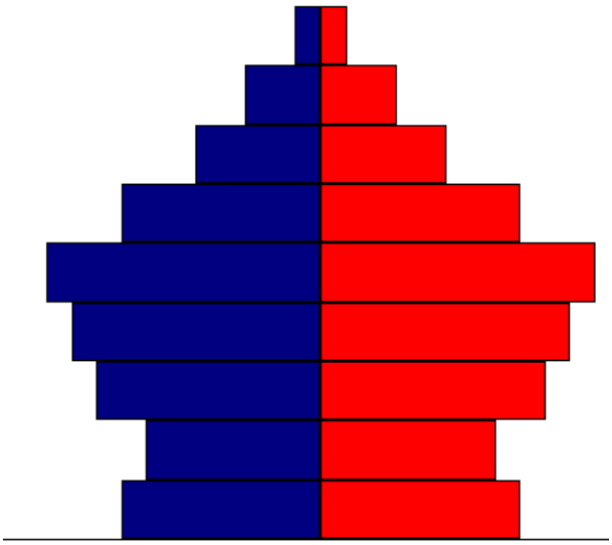
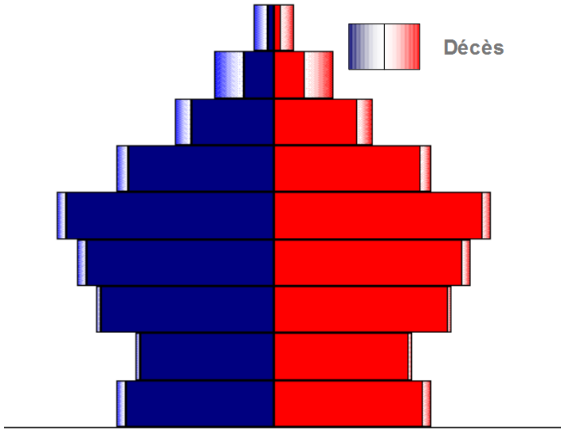


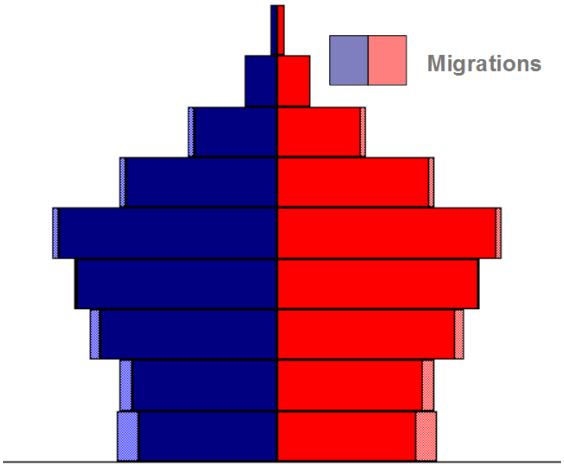
FIGURE 1 – Pyramide des âges 2013 simplifiée d'un territoire fictif

Les taux de fécondité, de mortalité par sexe et âge (uniquement par âge de la mère pour la fécondité) sont appréhendés sur le territoire à l'aide de l'état civil et du recensement de la population. Pour les migrations avec le reste de la France, seuls les taux d'émigration par sexe et âge sont estimés, la bi localisation des flux avec les zones d'échange permettant *in fine* d'appréhender conjointement les entrées et les sorties du territoire. Ces taux calculés pour 2013 au niveau local de la zone projetée sont ensuite prolongés dans le temps selon des hypothèses nationales aménagées dans plusieurs scénarios. À titre d'exemple, le scénario dit « central » vise à prolonger les dernières tendances observées : l'espérance de vie y atteindrait 90,3 ans pour les femmes en 2050 et 86,8 ans pour les hommes ; la fécondité se stabiliserait à un niveau proche de l'actuel. Par conséquent, un territoire où l'espérance de vie est faible en 2013 verrait bien cette dernière augmenter selon les tendances nationales mais conserverait en 2050 sa caractéristique de décès précoces par rapport à la moyenne française. L'hypothèse concernant les taux de migration internes à la France est également constituée par une stabilité : les mobilités internes au pays sont d'une part un jeu à somme nulle et, d'autre part, extrêmement complexes avec des polarisations différenciées selon le cycle de vie. Il n'y aurait que peu de sens à toutes les augmenter ou à toutes les diminuer dans les mêmes proportions.

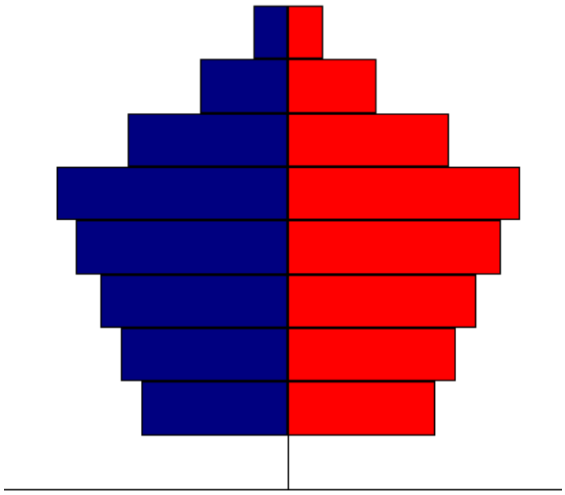
Une fois ces ratios mesurés, ils peuvent être appliqués à la population à laquelle ils se réfèrent pour estimer un nombre de nouveaux nés, de décès et de migrants sur le territoire. Ces effectifs sont ensuite déduits ou ajoutés au stock d'habitants du territoire que l'on fait également vieillir d'un an (Figure 2 a à c) pour passer à la pyramide des âges de l'année 2014. Le nombre d'enfants en bas âge est mesuré à partir d'un calcul spécifique intégrant les quotients de fécondité et permettant de tenir compte de manière indirecte de la mortalité infantile et des migrations résidentielles de nourrissons (Figure 2d). Cette opération est répétée jusqu'à l'année horizon c'est-à-dire 2050.



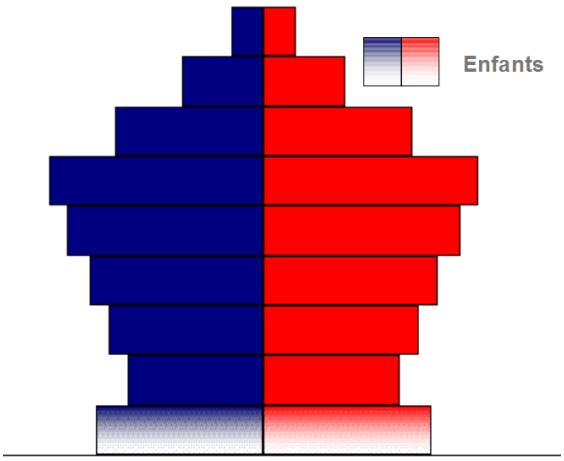
(a) Effet de l'application des quotients de décès sur la pyramide des âges 2013



(b) Effet des migrations internes et avec l'étranger sur la pyramide des âges 2013



(c) Effet du vieillissement de un an de la population pour passer de la pyramide des âges 2013 à celle de 2014



(d) Pyramide des âges 2014 suite à l'intégration des jeunes enfants

FIGURE 2 – Application d'un pas de projection

Concernant le solde avec l'étranger, l'hypothèse nationale tendancielle est fixée à + 70 000 personnes par an soit proche de la moyenne des soldes observés sur la dernière décennie. La ventilation locale de ce solde est complexe. Le recensement renseigne sur les entrées depuis l'étranger, et donc permet de les localiser, mais n'enquête que les résidents actuels : de ce fait aucune information n'est disponible sur l'origine géographique des personnes ayant quitté le territoire national.

La ventilation des sortants issus du solde national est donc faite au prorata du poids du territoire dans les entrées avec l'étranger. Autrement dit, l'hypothèse est faite d'un lien géographique fort entre entrées et sorties avec l'étranger. Or cette relation n'est pas vérifiée partout dans le pays, par exemple en Île-de-France, ce qui peut être à l'origine d'un manque de robustesse de l'outil.

1.2 Qu'est ce qu'une projection de population robuste ?

Le but d'une projection de population n'est pas de prédire l'avenir d'un territoire. Certes les hypothèses introduites dans le modèle ont été fixées avec un objectif de crédibilité. Pour autant, elles reposent sur l'idée d'un prolongement des tendances structurelles de la démographie française. Or, les tendances démographiques connaissent régulièrement des soubresauts, comme peut en attester par exemple le pic des naissances de l'an 2000. À un niveau local, les ruptures démographiques peuvent être encore plus soudaines notamment en ce qui concerne les mouvements migratoires de proximité. Il arrive que l'attractivité résidentielle d'un territoire fluctue très rapidement en fonction de différents facteurs comme la disponibilité du foncier, son prix, les politiques d'aménagement du territoire ... Une projection de population n'est pas en capacité de prévoir ces ruptures ; ce n'est non plus pas son objectif.

Dans ce cadre, la robustesse d'une projection ne peut en aucun cas être évaluée à l'aune de sa qualité prédictive. En d'autres termes, l'exercice *ex-post* de comparaison entre la population projetée et la réalisation dans le recensement de la population n'est pas l'outil adéquat pour juger de la qualité de projection. C'est d'autant plus le cas quand on ne dispose pas d'un cycle complet d'enquêtes annuelles de recensement pour évaluer une vraie tendance d'évolution sur 5 ans. Dans le cas d'Omphale qui débute en 2013, il faudrait en théorie attendre le recensement de 2018 pour observer une tendance suffisamment robuste pour la comparer à celle d'une projection.

Dans le cas d'un territoire qui a connu une rupture démographique forte dans les premières années de projection, il n'était pas du ressort d'Omphale de l'anticiper. L'exemple de l'arrondissement de Calais (Figure 3) fournit un cas d'école à ce titre. Les migrants résidant en habitations mobiles y ont été recensés en 2016. L'enquête annuelle de recensement 2016 intervient pour la première fois dans un millésime complet du recensement en 2014 d'où une forte hausse de la population entre 2013 et 2014, hausse non anticipée par Omphale. Or, cette forte croissance de la population tient à deux facteurs indépendants à la logique d'une projection :

- des dynamiques migratoires internationales extrêmement volatiles avec des effets de concentrations sur certains lieux ;
- le calendrier de collecte du recensement de la population qui enquête ces populations tous les cinq ans : dans la réalité l'effet des migrants à Calais a sans doute été plus diffus mais la méthode de collecte dans le recensement conduit à concentrer cet effet sur l'année 2014.

La robustesse des projections Omphale ne peut donc être jugée au regard de sa capacité à anticiper l'avenir. En ce sens, on peut s'interroger sur leur intérêt pour les acteurs locaux s'il est assumé que les projections Omphale n'ont que peu de chances de se réaliser. Dans le cadre du scénario tendanciel, Omphale vise à présenter un futur possible pour un territoire compte tenu des hypothèses sous-jacentes. Le résultat est donc conditionné à ces hypothèses, elles-mêmes non probabilistes, mais pour le scénario central déterminées de façon à prolonger les dernières évolutions de court ou moyen terme. Les projections centrales d'Omphale visent donc à déterminer la taille et la structure de la population dans le cadre d'une poursuite des tendances récentes, et notamment d'un maintien des quotients d'émigrations pourtant par nature volatiles. La non adéquation fréquente entre projection et réalisation n'est donc pas un indicateur de manque de robustesse d'Omphale mais représente plutôt un marqueur de ruptures fortes en matière démo-

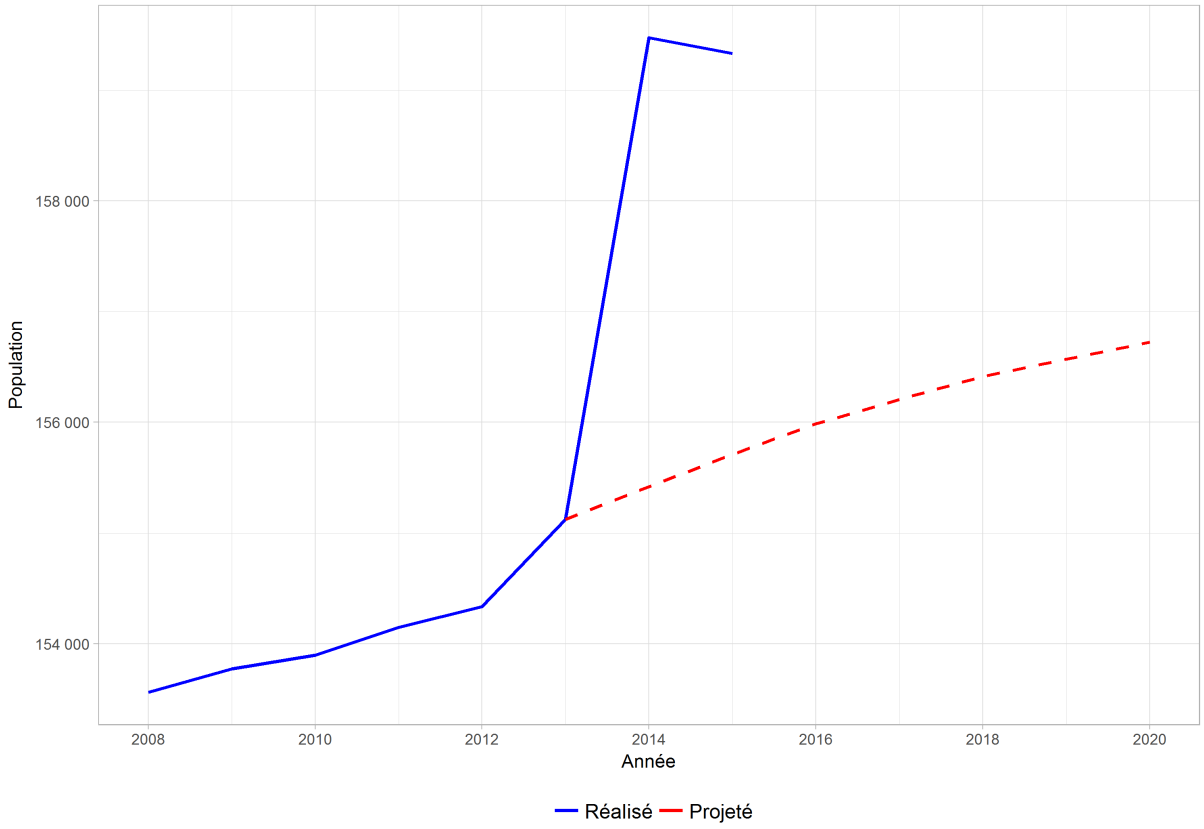


FIGURE 3 – Projection de population Omphale et recensement de la population dans l'arrondissement de Calais - Source : Omphale 2017, recensements de la population 2008 à 2013

graphique sur le territoire. C'est notamment le rôle des acteurs publiques de générer certaines ruptures, par exemple en rendant attractif un territoire en perte de vitesse démographique. Dès lors, le scénario tendanciel d'Omphale lui fournit une vision de son territoire où les tendances du passé se prolongent, par conséquent en amont de toute politique publique ou plus précisément en prolongeant les effets des dernières politiques mises en œuvre.

In fine, on peut essayer de déterminer la robustesse d'une projection de population dans Omphale comme sa capacité à retranscrire correctement les hypothèses du modèle, ou plus précisément, dans le cadre d'une projection du scénario central, à prolonger de manière satisfaisante les dernières tendances observées. Les statisticiens de l'Insee réalisent d'ors et déjà ces travaux de manière très empirique, « à dire d'expert » en observant les courbes de population sur le passé et en projection pour évaluer d'éventuelles ruptures. Des discontinuités trop fortes entre le passé et la projection peuvent attester d'un manque de robustesse d'Omphale sur un territoire. Dans de nombreux cas, ces difficultés du modèle viennent de la non connaissance de la localisation des émigrés vers l'étranger et de l'hypothèse de ventilation exposée précédemment. Si sur un territoire, il existe un déséquilibre fort entre les entrées et les sorties de France, Omphale pourrait sous-estimer sa population en générant tout de même des départs importants : c'est sans doute le cas sur l'arrondissement de Soisson (Figure 4).

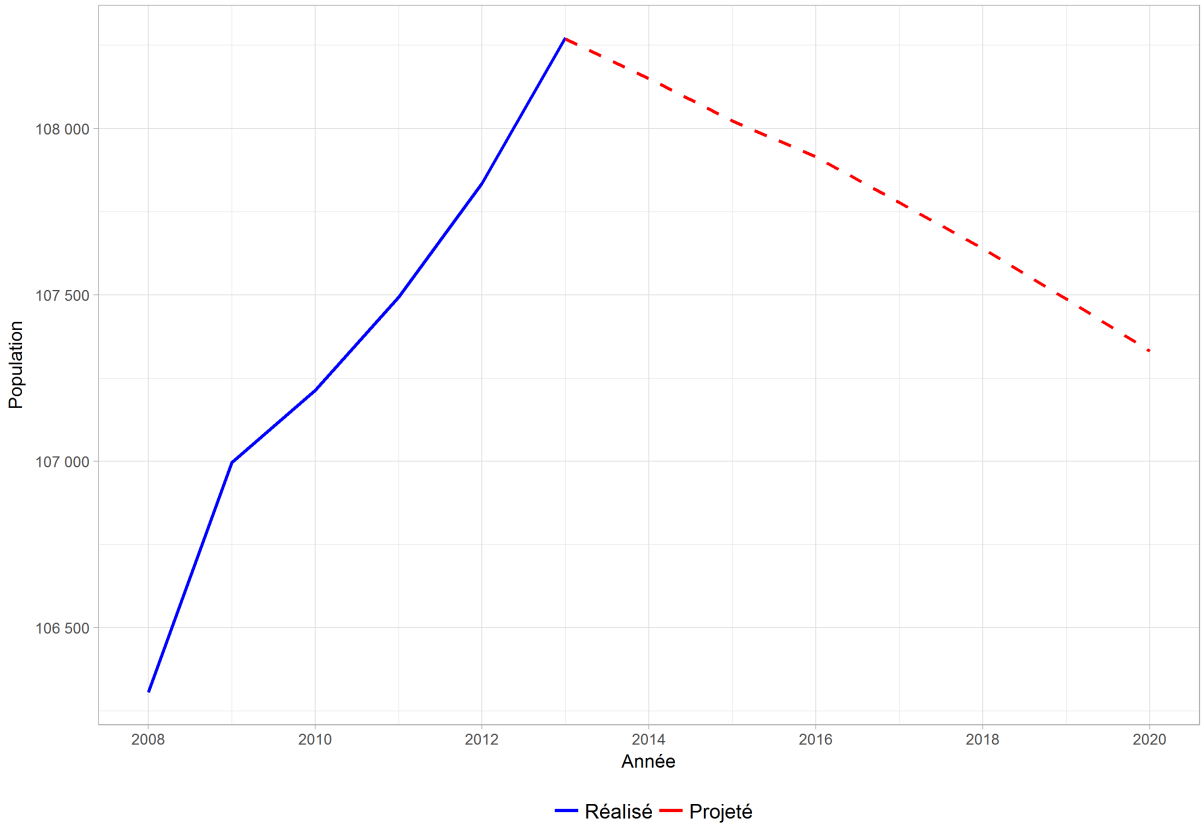


FIGURE 4 – Projection de population Omphale et recensement de la population dans l'arrondissement de Soisson - Source : Omphale 2017, recensements de la population 2008 à 2013

Ainsi, un premier indice de la robustesse d'Omphale est bien la prolongation des tendances passées en comparant les rythmes de croissance des populations observées dans le recensement et celles projetées. Pour autant, cette première méthode peut présenter au moins trois limites :

— il est difficile de déterminer quel recul historique retenir pour estimer la tendance démographique passée du territoire. En toute rigueur, la mécanique du recensement devrait conduire à privilégier des évolutions quinquennales. Dans ce cas, on fait fi des ruptures éventuelles au cours des cinq dernières années comme pour l'arrondissement de Château-Thierry (Figure 5). Le mode de collecte glissant du recensement devrait en théorie conduire à éviter les évolutions annuelles : pourtant la population a-t-elle réellement augmenté puis stagné ou a-t-elle augmenté en continue au fil de la période ? Selon que l'on choisisse l'une ou l'autre, l'évaluation que l'on pourra faire de la projection Omphale associée sera différente.

— il n'existe pas de différentiel de croissance estimé statistiquement à partir duquel la projection peut s'avérer problématique. Dans le cas de Soisson (graphique 7), la rupture est franche mais pour d'autres cas elle est plus légère et la décision sera forcément arbitraire. Qui plus est, il faut avoir en tête que la mécanique d'Omphale a pour objectif de prolonger les tendances passées mais que le résultat n'est pas forcément une adéquation parfaite entre taux de croissance projetés et observés. Un territoire où les comportements démographiques restent inchangés verra tout de même sa croissance de population affectée par des phénomènes tels que le vieillissement de la population ou la dynamique démographique des zones avec lesquelles il échange. Une projection robuste ne reproduit donc pas à l'identique les derniers taux de croissance observés : si les évolutions trop

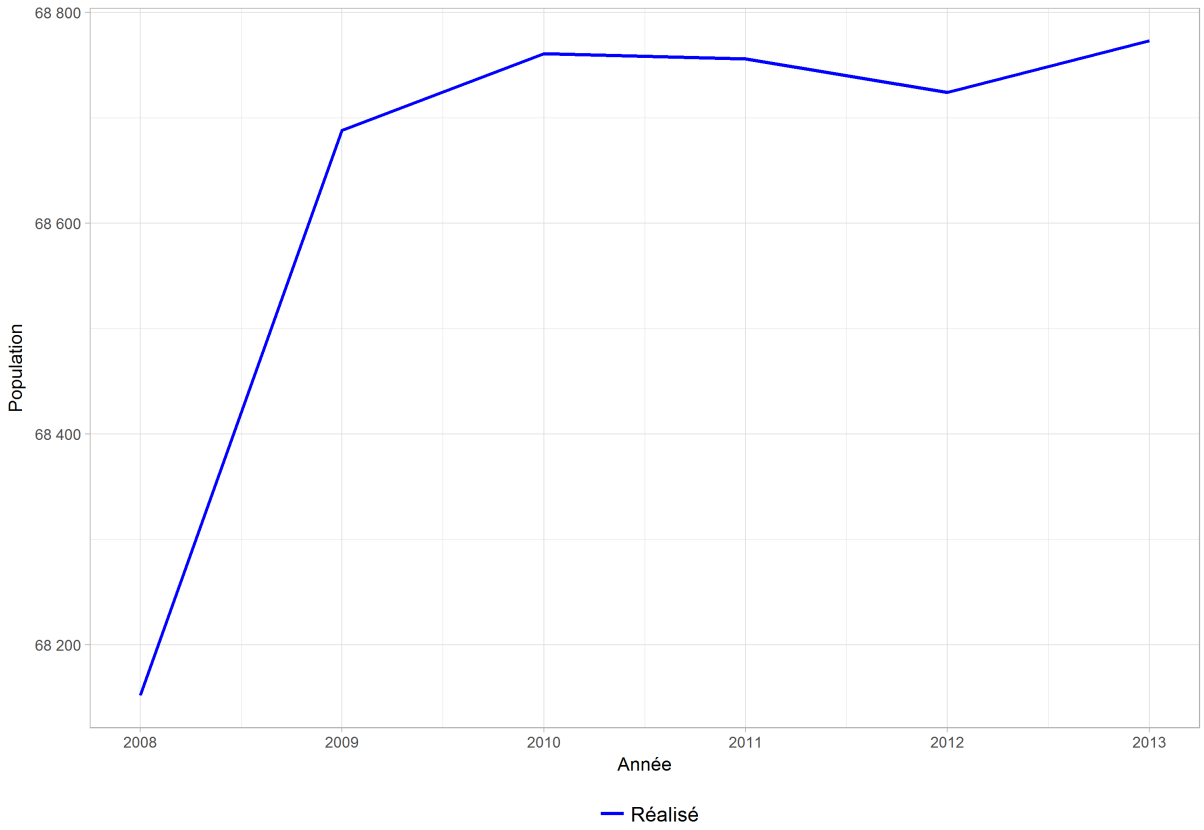


FIGURE 5 – Évolutions du recensement de la population dans l'arrondissement de Château-Thierry - Source : recensements de la population 2008 à 2013

brutales sont suspectes, des changements assez profonds peuvent intervenir sur des périodes plus longues. La projection du scénario central sur l'arrondissement de Saint-Omer (Figure 6) témoigne des infléchissements progressifs à la dynamique démographique du fait du vieillissement de la population. Le taux de croissance se réduit progressivement pour laisser finalement la place à une diminution de la population à partir des années 2040.

— l'intérêt d'Omphale n'est pas uniquement de donner des informations sur la population totale mais également de la ventiler par sexe et âge. Les politiques d'aménagement du territoire à mener ne seront pas les mêmes selon que la croissance de la population est portée par des étudiants, des couples avec enfants ou des séniors. De plus, l'application Omphale est la base d'autres outils méthodologiques Insee déclinant les projections sur d'autres populations d'intérêt : actifs, ménages, élèves, personnes âgées dépendantes ... Par conséquent, la robustesse d'une projection ne se limite pas à l'analyse de la population totale mais doit également valider les évolutions par âge. Cette analyse est plus complexe car la présence de classes d'âge pleines et de classes d'âges creuses conduise à des évolutions très erratiques du nombre de personnes au sein d'une classe d'âge.

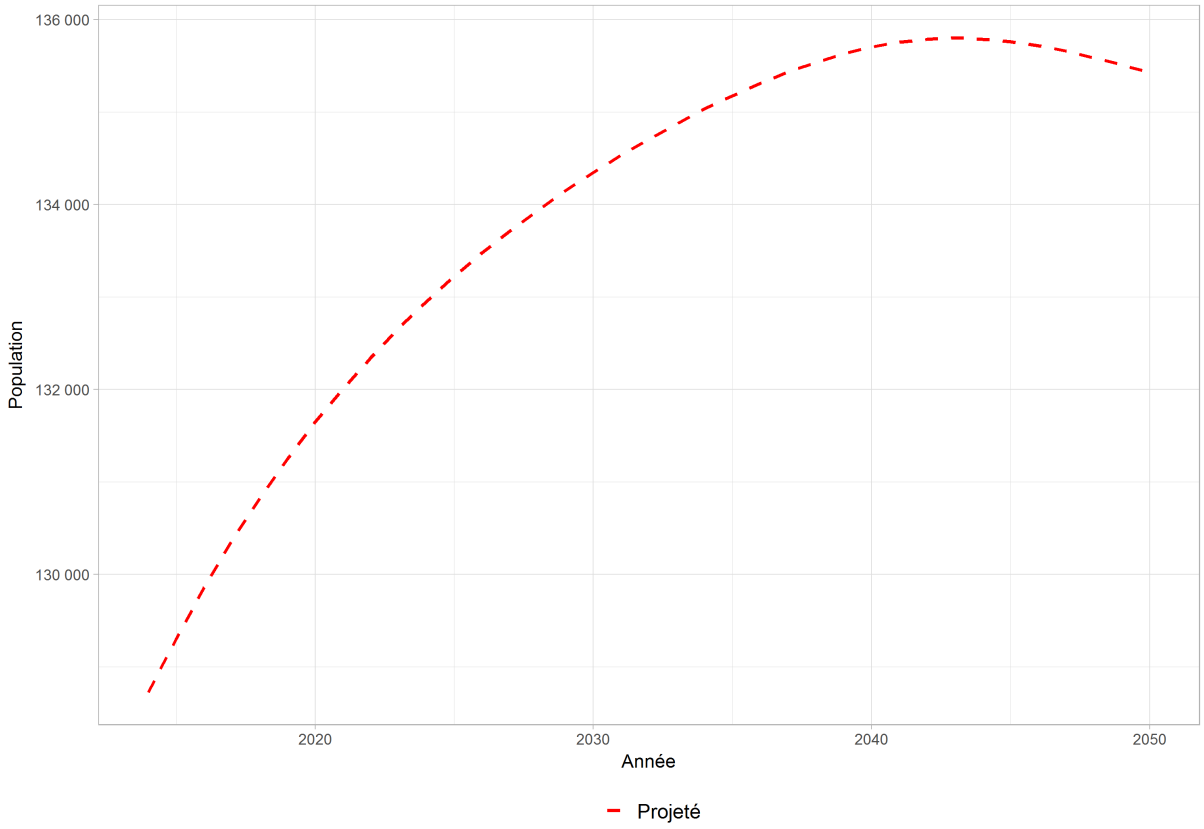


FIGURE 6 – Projection de population Omphale sur l'arrondissement de Saint-Omer - Source : Omphale 2017

1.3 La prise en compte des effets âge et des effets de générations dans le temps

Pour tenter de répondre à ces trois critiques de l'analyse à dire d'expert de la robustesse d'une projection, une méthode de complément a été proposée aux utilisateurs d'Omphale. Elle vise à permettre de détecter des projections potentiellement problématiques en intégrant notamment l'importance de l'analyse par âge.

Le principe de base d'Omphale est de déformer séquentiellement une pyramide des âges dans le temps. Deux grands types de phénomènes entrent en jeu pour expliquer les formes constatées et projetées d'une pyramide des âges :

- les « effets de générations » sont constitués par des différences d'effectifs entre les classes d'âge liées à la génération à laquelle appartiennent les individus. Ils sont donc la résultante de phénomènes relativement structurant, nationaux ou internationaux dont un bon exemple est le baby boom. Dans la pyramide des âges de la France en 2013 et dans la plupart des pyramides des âges locales, le baby boom se traduit par des effectifs importants entre 40 et 65 ans ;
- les « effets âge » tiennent aux caractéristiques par âge du territoire observé, notamment en lien avec les différents comportements en matières de migrations résidentielles au long du cycle de vie. Les « effets âge » peuvent exister au niveau national mais ils sont bien plus visibles à l'échelle locale : les pôles universitaires ont une pyramide des âges très caractéristique avec des effectifs nombreux depuis l'entrée dans l'enseignement supérieur jusqu'aux premières années d'activité et des classes plus creuses ensuite liées notamment au phénomène de périurbanisation.

Chacun de ces deux phénomènes a des répercussions différentes en projection : les effets de générations sont soumis au vieillissement de la population (les baby boomers vont continuer de vieillir dans les prochaines années sur tout le territoire national) tandis que les effets âges sont plus constants dans le temps (un territoire étudiant le reste quand bien même les étudiants de 2013 vieillissent, ils sont remplacés par ceux des générations suivantes). Cette distinction entre effet âge et effet de générations se fait dans Omphale par le biais des quotients de migration : les classes pleines liées au baby boom ne sont pas liées à des arrivées sur la zone d'intérêt, elles vont continuer à s'élever dans la pyramide des âges. À l'inverse, celles liées aux étudiants s'expliquent par des entrées et des sorties modélisées dans Omphale ce qui permet un maintien de la classe pleine à partir de 17 ans. Un nouvel élément pour juger de la robustesse d'une projection est donc que cette dernière respecte dans le temps les effets âge et les effets de générations.

1.4 Comment évaluer la déformation d'une pyramide des âges ?

Il n'est pourtant pas simple de distinguer de manière systématique et automatique les effets de générations et les effets âge. Un territoire comme Paris connaît un déficit migratoire important des jeunes retraités par conséquent, l'effet âge lié à ce phénomène va influencer l'effet lié au vieillissement des baby boomers. Comment dès lors qu'on ne peut pas déterminer *a priori* pour tout zonage à façon les effets âges et de générations s'assurer que la déformation de la pyramide des âges respecte les caractéristiques du territoire ?

La méthode retenue ici vise à déterminer des types de pyramides des âges en 2013 par le biais d'une typologie et d'observer pour chacun de ces types une déformation traditionnelle de la pyramide des âges entre 2013 et 2050 : autrement dit associer aux territoires étudiants un maintien dans le temps du pic de population chez les 17 – 25 ans. Ainsi pour toute nouvelle projection d'Omphale, le territoire peut être regroupé avec un type de pyramide des âges. Les évolutions de sa population par âge peuvent alors être comparées à celles des autres zones du même type. Dès lors, des éventuels problèmes de robustesse pourront être détectés dans les cas où ces évolutions s'avéreraient « extrêmes » au regard de la distribution des autres zones pourtant comparables en termes de structure en 2013.

L'intérêt de cette méthode est double :

- elle ne présuppose pas les effets âge et les effets de générations. Elle postule que sur un grand nombre de zones du même type, ces effets sont respectés. Dès lors, des déformations de la pyramide des âges différentes de la déformation typique peuvent permettre d'isoler des cas problématiques ;
- en se basant sur l'observation de distributions, elle ne nécessite pas de déterminer a priori des seuils au-delà desquels des particularités sont à prendre en considération.

À l'inverse, il faut garder en tête qu'en ciblant les individus « extrêmes » au sens de la distribution, la méthode conduit, par définition, à cibler pour chaque type de territoire des cas potentiellement problématiques, quand bien même les différences seraient faibles. De plus, des évolutions atypiques peuvent très bien se justifier par des facteurs non intégrés dans les critères de constitution des groupes de territoire. Être atypique ne signifie donc pas être automatiquement non robuste : il ne s'agit que d'une aide à l'analyse de la robustesse venant compléter une expertise plus qualitative des projections.

2 Évaluation de la robustesse à l'aide d'une typologie des structures de population

Afin de mesurer la robustesse de la projection d'une zone, l'évolution de la structure de cette dernière va être comparée à celles de zones présentant des caractéristiques similaires (population jeune, âgée, étudiante ...).

Cependant, le type de structure d'une zone n'est pas connu *a priori*. Nous disposons d'un ensemble de pyramides des âges dont nous ne connaissons pas le type de structure. Il s'agit d'un problème d'apprentissage non supervisé : aucune étiquette n'est associée à une zone. Une typologie des structures de population (en 2013) est créée afin d'identifier différents types de structures puis permettre d'étiqueter de nouvelles zones.

2.1 Typologie des structures de population en 2013

2.1.1 Choix des variables

Pour réaliser la typologie des pyramides des âges, il est nécessaire de choisir comme données initiales des zones à un niveau suffisamment local. En effet, à un niveau agrégé, les pyramides des âges sont lissées et peuvent donc cacher d'éventuels phénomènes démographiques. La typologie a donc été réalisée sur des individus statistiques correspondant à un niveau local assez important pour assurer une relative robustesse d'Omphale à un niveau global mais suffisamment fin pour présenter des structures hétérogènes. Nous avons choisi les zones d'emploi comme individu statistique.

À chaque zone d'emploi, on associe sa population par âge et sexe en 2013. Les populations totales sont normées à 100 afin d'éviter des effets tailles. En somme, chaque zone est décrite par un vecteur à 200 composantes.

Néanmoins, un grand nombre de ces variables sont très corrélées entre elles. Afin de réduire le nombre de variables décrivant chaque zone, les populations sont regroupées indépendamment du sexe. De plus, au sein de certaines tranches d'âges, les distributions de population sont très similaires. Par exemple, la proportion de la population ayant 70 ans est assez similaire à celle ayant 69 ans ou 71 ans. Par ailleurs, certaines tranches s'opposent à d'autres. Par exemple, la part d'individus jeunes sera généralement très corrélée négativement avec celle des individus âgés.

Des classes d'âges ont donc été créées en utilisant l'algorithme ClustOfVar développé par Chavent *et al.*. Onze classes d'âges ont été retenues correspondant à des populations facilement identifiables (0 an à 9 ans − 10 ans à 17 ans − 18 ans à 24 ans − 25 ans à 31 ans − 32 ans à 40 ans − 41 ans à 49 ans − 50 ans à 54 ans − 55 ans à 59 ans − 60 ans à 72 ans − 73 ans à 91 ans − 91 ans à 99 ans).

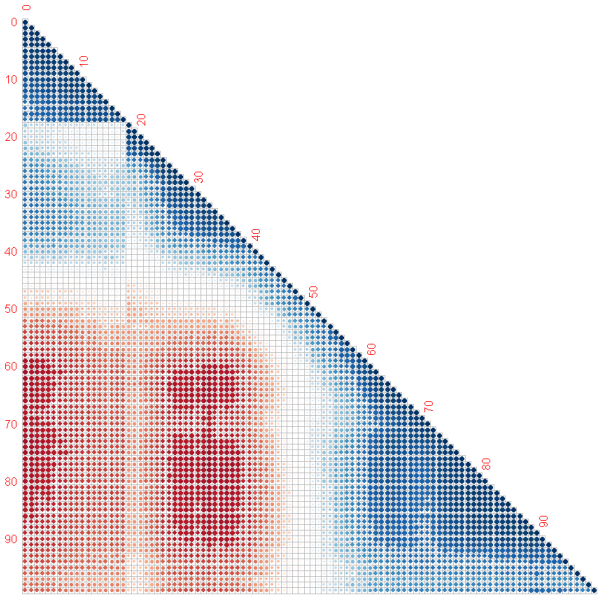


FIGURE 7 – Corrélation entre âges

NOTE DE LECTURE : Chaque point correspond à la corrélation entre les populations de chaque zone d'emploi en 2013 pour deux âges. Si le point est bleu alors il s'agit d'une corrélation positive et si le point est rouge alors il s'agit d'une corrélation négative. Plus le point est coloré, plus la corrélation est forte. Les points blancs démontrent une absence de corrélation. On remarque que certaines tranches d'âges sont très fortement corrélées entre elles.

Ainsi chaque zone d'emploi se présente comme un vecteur à onze composantes :

Zone	[0,10)	[10,18)	[18,25)	[25,32)	[32,41)
Lille	12.81	9.71	13.04	10.79	12.18
[41,50)	[50,55)	[55,60)	[60,73)	[73,92)	[92,99]
11.41	6.05	5.70	10.54	7.43	0.28

TABLE 1 – Exemple de structure de population

2.1.2 Choix de la méthode et résultats

Plusieurs méthodes de classification non supervisée ont été comparées pour réaliser cette typologie : des méthodes de classification hierarchique (CAH) et de partitionnement (*k-means*). Le choix du nombre de classes est délicat car il répond à une double contrainte. D'une part, le nombre de groupes doit être faible afin d'avoir suffisement d'éléments dans chaque groupe et ainsi pouvoir inférer. D'autres part, il doit être assez grand pour observer des comportements démographiques particuliers.

Une typologie en cinq classe par la méthode des *k-means* a été retenue. Chacune de ses classes correspond à un phénomène :

— Une classe ayant une forte proportion de personnes de plus de 50 ans (Population âgée).
— Une classe ayant une proportion plus faible de personne ayant plus de 50 ans que la première classe mais plus forte que la moyenne des zones d'emploi. La part des individus

âgés entre 41 et 55 ans y est similaire à la moyenne sur les zones d'emploi (Population légèrement plus âgée que la moyenne).
— Une classe ayant une structure proche de la structure moyenne des zones d'emploi (Population ayant une structure proche de la structure moyenne).
— Une classe ayant une structure plus jeune que la structure moyenne des zones d'emploi (Population jeune).
— Une classe ayant une forte proportion de personnes âgées de 18 à 25 ans ou une forte proportion proportion de personnes âgées de 0 à 10 ans et de 41 à 50 ans.

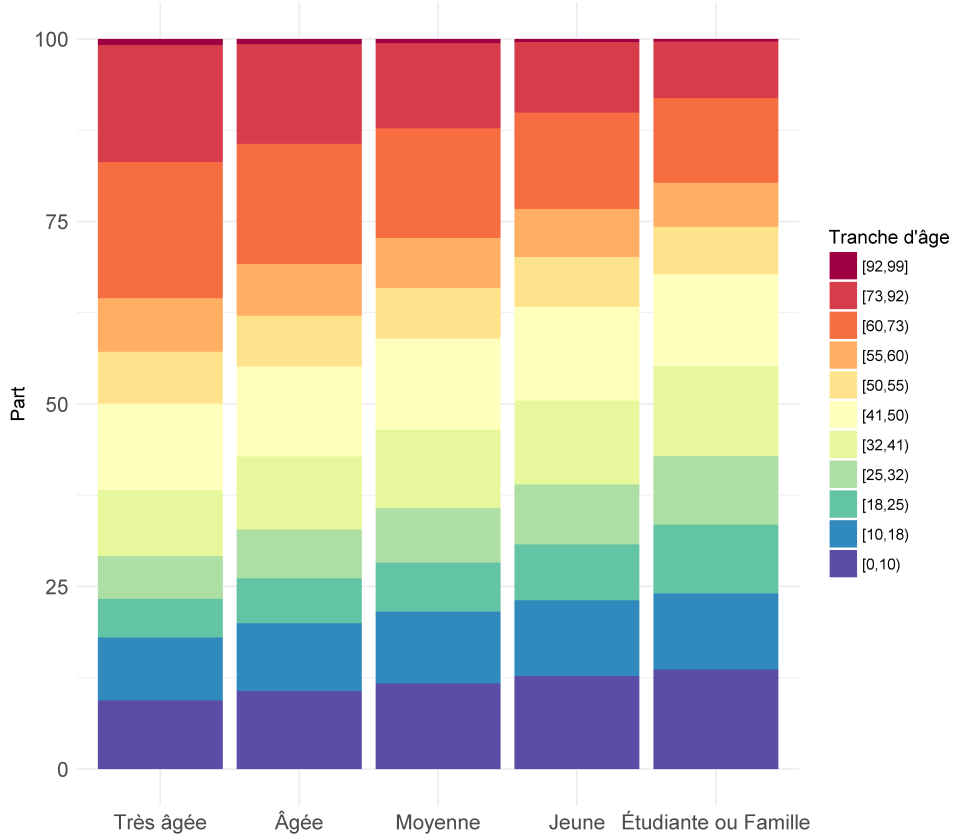


FIGURE 8 – Part moyenne de chaque tranche d'âge par type de structure

2.2 Étiquetage de nouvelles zones

Grâce à la typologie réalisée précédemment, chaque zone d'emploi dispose d'une étiquette décrivant sa structure de population. Notons n, le nombre de zones d'emploi utilisée pour réaliser la typologie, $\mathbf{pop}_i^p \in [0;100]^{11}$ le vecteur décrivant la structure de population de la zone i à l'année p et enfin $\mathbf{type}_i$ le type de structure de la zone.

Le but est de trouver une fonction $\hat{\phi}$ qui a un vecteur $\mathbf{pop}_i^{2013}$ lui associe son type $\mathbf{type}_i$ à l'aide des données étiquetées grâce à la typologie. Ensuite $\hat{\phi}$ sera utilisée pour déduire le type de structure d'une nouvelle zone (cette nouvelle zone n'étant pas nécessairement une zone d'emploi).

2.2.1 Étiquetage par le plus proche voisin sur les barycentres

Une première proposition est de déterminer, pour une nouvelle zone dont la structure de population est $\mathbf{pop}_p^{2013}$, son type de structure à l'aide de celui dont elle est la plus "proche". Pour se

faire, on définit le barycentre d'un type de structure **type**, noté $\overline{\mathbf{pop}}_{\mathbf{type}}$, comme étant la structure moyenne des zones ayant ce type de structure. Autrement dit, $\overline{\mathbf{pop}}_{\mathbf{type}} = \frac{1}{n_{type}} \sum_{i \in \mathbf{type}} \mathbf{pop}_i$. Ces zones forment l'échantillon d'apprentissage.

Le type de la structure de la zone p sera celui qui minimise la distance euclidienne canonique entre $\mathbf{pop}_p^{2013}$ et $\{\overline{\mathbf{pop}}_{\mathbf{type}}\}_{\mathbf{type} \in 1;5}$. Il en vient que $\hat{\phi}(\mathbf{pop}_p^{2013}) = \arg\min_{\mathbf{type}} ||\mathbf{pop}_p^{2013} - \overline{\mathbf{pop}}_{\mathbf{type}}||^2$. Cette règle de décision correspond à l'utilisation de l'algorithme des 1-plus proches voisins où l'échantillon d'apprentissage correspond aux barycentres des structures de population.

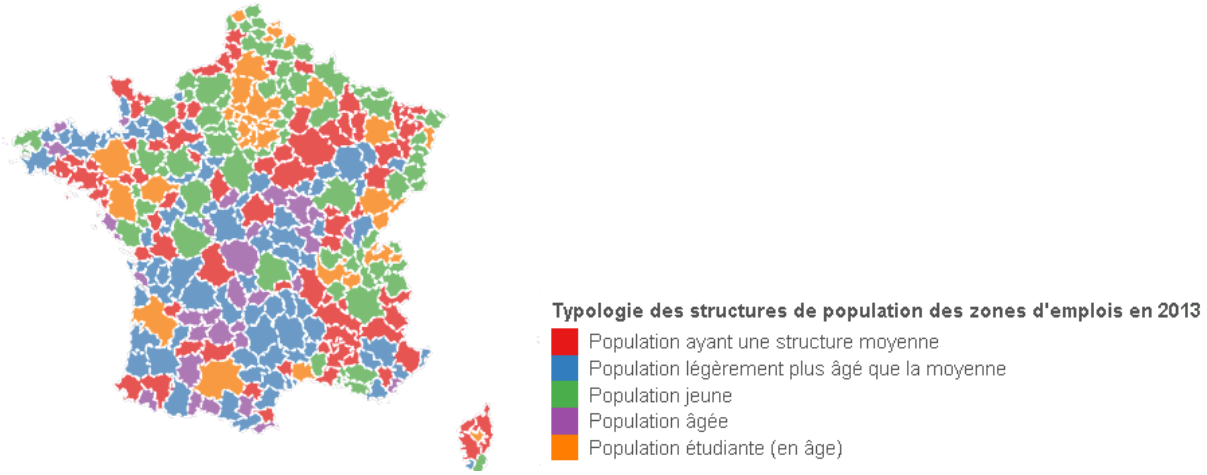


FIGURE 9 – Typologie des structures de population en 2013

2.3 Test général

L'idée du test général est de comparer l'évolution de la structure globale de la zone par rapport à celles des zones ayant le même type de structure.

Dans un premier temps, le type de structure de la zone à tester va être déterminé à l'aide du classifieur $\hat{\phi}$. Puis l'évolution de la structure de la population de la zone va être comparée à celles des zones partageant le même type de structure de population (et appartenant à l'ensemble d'apprentissage). Si l'évolution est *sensiblement* différente de l'évolution moyenne des zones partageant le même type de structure alors on conclura que la zone a une évolution extrême par rapport à son type de structure en 2013.

2.3.1 Mesure de l'évolution de la structure de population

Pour chaque zone i ayant un type de structure $\mathbf{type}_i$, on introduit l'évolution de la zone i comme étant le vecteur $\mathbf{evol}_i$ avec $\mathbf{evol}_{i_j} = \frac{\mathbf{pop}_{i_j}^{2050} - \mathbf{pop}_{i_j}^{2013}}{\mathbf{pop}_{i_j}^{2013}}$ où j décrit les classes d'âge. On note respectivement $\overline{\mathbf{evol}}_{\mathbf{type}}$ et $\sigma_{\mathbf{type}}$, les vecteurs dont chacune des composante sont respectivement la moyenne et l'écart type des composantes des évolutions des zones de l'échantillon d'apprentissage (zones d'emploi) ayant un type de structure **type**. On a $\overline{\mathbf{evol}}_{\mathbf{type}} = \frac{1}{n_{\mathbf{type}}} \sum_{i \in \mathbf{type}} \mathbf{evol}_i$

et $\sigma_{\mathbf{type}} = \frac{1}{n_{\mathbf{type}}} \left(\sum_{i \in \mathbf{type}} \mathbf{evol}_j - \overline{\mathbf{evol}}_{\mathbf{type}} \right)^2$

L'écart d'une zone à tester p à l'évolution moyenne du type de structure auquel elle appartient, noté $\mathbf{ecart}_p$, correspond à une distance entre l'évolution $\mathbf{evol}_p$ et l'évolution moyenne du type de structure $\overline{\mathbf{evol}}_{\mathbf{type}}$ d'où $\mathbf{ecart}_p = d(\mathbf{evol}_p, \overline{\mathbf{evol}}_{\hat{\phi}(pop_p^{2013})})$ avec d une distance.

Dans la suite, on considérera que $\mathbf{ecart}_p = d(\mathbf{evol}_p, \overline{\mathbf{evol}}_{\hat{\phi}(pop_p^{2013})}) =$

$$\left(\sum_{j=1}^{11}\left(\frac{\mathbf{evol}_{pj} - \overline{\mathbf{evol}}_{\hat{\phi}(pop_p^{2013})j}}{\sigma_{\mathbf{type}_j}}\right)^2\right)^{\frac{1}{2}}.$$

2.3.2 Seuils

Une projection sera considérée comme extrême si l'écart $\mathbf{ecart}_p$ est supérieur à un seuil. Mais comment choisir ce seuil ? Supposons que l'on dispose d'une zone à tester dont l'évolution est donnée par $\mathbf{evol}_p$ et dont le type de structure de population est $\mathbf{type} = \hat{\phi}(pop_p^{2013})$. On considérera que la projection de la population de la zone p est extrême si l'écart $\mathbf{ecart}_p$ prend des valeurs extrêmes par rapport à la distribution des écarts des zones de l'échantillon d'apprentissage (zone d'emploi) partageant le même type de structure.

Nous faisons l'hypothèse que chaque écart de zone d'un même type de structure $\mathbf{type}_j$ est la réalisation d'une variable aléatoire de loi $\mathbb{P}_{\mathbf{type}}$. Concrètement, on dira que la projection est extrême si $\mathbf{ecart}_p > Q_{95}(\mathbb{P}_{\mathbf{type}})$ où $Q_\alpha(\mathbb{P}_{\mathbf{type}})$ est le quantile d'ordre α de $\mathbb{P}_{\mathbf{type}}$ (i.e $Q_\alpha(\mathbb{P}_{\mathbf{type}}) = \inf\{x | F_{\mathbb{P}_{\mathbf{type}}}(x) < \alpha\}$) où $F_{\mathbb{P}_{\mathbf{type}}}$ est la fonction de répartition de la loi $\mathbb{P}_{\mathbf{type}}$.

Néanmoins la loi $\mathbb{P}_{\mathbf{type}}$ est inconnue et donc $Q_{95}(\mathbb{P}_{\mathbf{type}})$ n'est pas disponible à l'aide des seules données. $Q_{95}(\mathbb{P}_{\mathbf{type}})$ est estimé à l'aide du quantile empirique $\hat{Q}_{95}(\mathbb{P}_{\mathbf{type}})$ observé sur les écarts des zones de l'échantillon d'apprentissage (les zones d'emploi) dont le type de structure est de type $\mathbf{type}$ (i.e $Q_\alpha(\mathbb{P}_{\mathbf{type}}) = \inf\{x | \hat{F}_{\mathbb{P}_{\mathbf{type}}}(x) < \alpha\}$) où $\hat{F}_{\mathbb{P}_{\mathbf{type}}}$ est la fonction de répartition empirique basée sur les écarts observés.

On dira donc que la projection de population d'une zone p est extrême si $\mathbf{ecart_p} > \hat{Q}_{95}(\mathbb{P}_{\mathbf{type}})$.

Ceci dit, l'estimation du quantile empirique présente de la variabilité. Pour chaque type de structure, on se propose de fournir un intervalle de confiance du quantile par bootstrap. Pour se faire, nous tirons B = 500 échantillons bootstrapés (tirage équiprobable avec remise) sur lesquels on calcule des réplications de $\hat{Q}_{95}(\mathbb{P}_{\mathbf{type}})$. Ainsi on obtient un échantillon de B réplications boostrapés $(\hat{Q}_{95}(\mathbb{P}_{\mathbf{type}})_1^*, ..., \hat{Q}_{95}(\mathbb{P}_{\mathbf{type}})_B^*)$. Un intervalle de confiance à 95 % est donné par $[q_1^*, q_2^*]$ où q_1^* et q_2^* correspondent respectivement aux quantiles empiriques d'ordre 2.5 % et 97.5% de l'échantillon des réplications. On considérera que si un ecart appartient à cet intervalle alors il n'est pas significativement différent de $\hat{Q}_{95}(\mathbb{P}_{\mathbf{type}})$ d'où une deuxième proposition de test : on dira que la projection de population d'une zone p est extrême si $\mathbf{ecart_p} > q_2^*$

2.3.3 Résultats sur les départements

Le test a été utilisé sur l'ensemble des départements de France. Globalement les DOM ont une évolution extrême par rapport à leurs types de structures en 2013. Ceci peut s'expliquer par une démographie différente par rapport à celle de France métropolitaine. On retrouve peu de différences entre les deux tests : le Puy-de-Dôme a une évolution extrême avec le test général avec bootstrap mais non extrême avec la deuxième version. Les autres départements sont classés comme non extrêmes.

Départements	Avec bootstrap	Sans bootstrap
Guadeloupe	Extrême	Extrême
Guyane	Extrême	Extrême
La Réunion	Extrême	Extrême
Martinique	Extrême	Extrême
Mayotte	Extrême	Extrême
Corse du Sud	Extrême	Extrême
Puy de Dome	Extrême	Non extrême

TABLE 2 – Résultats du test général sur les départements

2.4 Test par classe d'âge

2.4.1 Principe

Dans cette section, nous introduisons une déclinaison du test général. Le test par classe d'âge reprend le même principe que le test général mais s'applique à chaque tranche d'âge. On considère une zone à tester p dont le type de structure de population est $\mathbf{type} = \hat{\phi}(pop_p^{2013})$. On dira que l'évolution *pour une classe d'âge j* est extrême si l'évolution de la population pour cette tranche d'âge est en dehors d'un intervalle propre à chaque type de structure et à chaque tranche d'âge. Dans le cas présent, l'évolution *pour une classe d'âge j* est extrême si $evol_{p,j} \notin [q^1_{type,j}, q^2_{type,j}]$ où $q^1_{type,j}$ et $q^2_{type,j}$ correspondent respectivement aux quantiles empiriques à 2.5 et 97.5 % de l'échantillon $\{\text{evol}_{i,j}\}_{i \in \mathbf{type}}$.

Comme pour le test général, on propose une deuxième version où la variabilité des estimations est prise en compte grâce au bootstrap. À chaque quantile estimé, on associe un intervalle de confiance à 95 % par bootstrap construit comme précédemment : $[q^{1,*}_{type,j,1}, q^{1,*}_{type,j,2}]$ pour $q^1_{type,j}$ et $[q^{2,*}_{type,j,1}, q^{2,*}_{type,j,2}]$ pour $q^2_{type,j}$. On dira ici que l'évolution pour une classe d'âge j est extrême si $evol_{p,j} \notin [q^{1,*}_{type,j,2}, q^{2,*}_{type,j,1}]$.

On remarque que, dans notre cas, pour tous les types de structures **type** et pour toutes les tranches d'âges j, on a $[q^{1,*}_{type,j,2}, q^{2,*}_{type,j,1}] \subset [q^1_{type,j}, q^2_{type,j}]$: il en vient que le nombre de classes extrêmes sera plus grand avec la méthode par bootstrap.

2.4.2 Résultat sur les départements

Le test a été ensuite utilisé sur l'ensemble des départements. Les résultats pour les dix départements présentant le plus de classes extrêmes sont disponibles dans la table suivante pour les deux versions du test. Comme prévu, la version "avec bootstrap" compte plus de classes extrêmes. Il y a peu de différences entre les deux versions si le nombre de classes extrêmes est grand sauf pour la Corse du Sud. En effet, sur l'ensemble des départements, on trouve un coefficient de corrélation de 0.9 entre le nombre de classes extrêmes dans la version "avec bootstrap" et le nombre de classes extrêmes dans la version "sans bootstrap".

On remarque que tous les DOM ont un grand nombre de classes d'âge extrêmes. D'une part, les DOM disposent de caractéristiques démographiques différentes du reste de la France et d'autres part, le nombre de classes extrêmes dans ces départements est augmenté car les DOM n'ont pas contribué à la conception du classifieur $\hat{\phi}$.

Départements	Avec bootstrap	Sans bootstrap
Martinique	10	9
Guyane	8	7
La Reunion	8	6
Mayotte	8	6
Guadeloupe	7	7
Puy de Dome	7	7
Corse du Sud	6	2
Loire	5	4
Vienne	5	2
Paris	4	2

TABLE 3 – Résultats du test sur les départements

NOTE DE LECTURE : La colonne "avec bootstrap" donne le nombre de classes d'âge extrêmes au sens du test comparant les évolutions aux quantiles des évolutions du groupe en prenant en compte la variabilité due à l'estimation de ce dernier. Le colonne "sans bootstrap" donne le nombre de classes extrêmes au sens du test comparant les évolutions directement aux quantiles des évolutions du groupe. Pour rappel, il y a onze classes d'âge.

Conclusion

Dès lors que la robustesse d'une projection de population ne répond pas à une définition purement statistique, il apparaît qu'aucun outil n'arrivera à proposer un test systématiquement fiable. Les outils proposés ici se basent sur les pratiques informelles des utilisateurs d'Omphale (observations des courbes de croissance ou des déformations des pyramides des âges) et en proposent une automatisation. Certains concepts comme les effets âges et les effets de générations ne pouvant être détectés de manière standardisée, la démarche s'est donc tournée sur des comparaisons empiriques entre territoires ayant des caractéristiques proches.

Ce choix d'analyse des distributions conduit donc par nature à
- une diversité au sein des classes constituées ;
- la présence de territoires considérés comme extrêmes.

S'ensuit une différenciation entre le caractère atypique d'une projection et celle de sa robustesse, le premier étant un élément nécessaire mais non suffisant pour juger du second.

De plus, les utilisations d'Omphale sont nombreuses. Les suites à donner à une projection jugée non robuste également : il est possible de modifier certaines hypothèses (flux migratoires internes ou avec l'étranger ...) afin de les faire mieux coller à la réalité. Dès lors, compte tenu de ces possibilités offertes par l'outil, il devient primordial d'adapter le diagnostic et la réponse apportée à la problématique : une déformation atypique de la pyramide des âges pour les 18 – 25 ans n'aura pas d'impact sur un exercice de projections de collégiens.

C'est pour ces raisons que cet outil ne se veut qu'un appui à l'utilisateur dans le cadre d'une analyse à la fois quantitative et qualitative. Il existe des raisons légitimes pour lesquelles une projection peut être jugée extrême mais robuste que seule une analyse de la problématique et la connaissance fine du territoire, de ses caractéristiques et de ses dynamiques permet de comprendre.

Références

[1] Chernick, Michael R., L.-R. A. *An Introduction to Bootstrap Methods with Applications to R*, 1st ed. Wiley Publishing, 2011.

[2] Marie Chavent, Vanessa Kuentz-Simonet, B. L. J. S. Clustofvar : An r package for the clustering of variables. *Journal of Statistical Software, Articles 50*, 13 (2012), 1–16.

[3] Nathalie Blanpain, G. B. *Projection de population 2013-2070 pour la France : méthode et principaux résultats*. No. F1606. 2016.

[4] Olivier Léon, D. D. Le modèle de projection démographique omphale 2010. *Insee Methodes*, 124 (2011), 5–44.