COMMISSARIAT
GÉNÉRAL AU
DÉVELOPPEMENT
DURABLE

SERVICE
DES DONNÉES ET
ÉTUDES STATISTIQUES

JANVIER
2020

Document de travail n° 46

Développement durable

Méthode d'estimation sur petits domaines : l'exemple de la régionalisation d'indicateurs de bien-être subjectif

Service des données et études statistiques
www.statistiques.developpement-durable.gouv.fr

Résumé : *deux grandes familles d'indicateurs de bien-être peuvent être distinguées : les indicateurs subjectifs d'une part, qui s'appuient pour l'essentiel sur des enquêtes d'opinion, et les indicateurs objectifs d'autre part, qui, de manière plus normative, identifient les déterminants du bien-être et les traduisent au travers de variables tangibles telles que le taux de chômage, la qualité de l'air ou encore la vitalité associative. À l'ère du Big Data, et pour mieux représenter les réalités locales du terrain, les indicateurs de bien-être sont calculés sur des unités géographiques de plus en plus fines. Si les indicateurs objectifs se prêtent bien à cet exercice de régionalisation, les indicateurs subjectifs restent très peu disponibles à des niveaux infra-nationaux. Dans ce contexte, une tentative de régionalisation des variables de satisfaction de l'enquête Statistiques sur les ressources et conditions de vie* (*SRCV) a été réalisée afin de construire un indicateur subjectif composite de bien-être au niveau régional. La principale originalité de l'étude est de tenir compte, au niveau local le plus fin, non seulement des niveaux de satisfaction des individus, mais aussi de l'importance que ces derniers accordent aux différents aspects de leurs vies, du travail à la santé en passant par le logement, les relations avec les proches ou encore les loisirs.*

Note préliminaire : *ce document de travail propose de mettre en œuvre deux méthodes d'estimation sur petits domaines. À titre illustratif seulement, il propose d'aborder la question de la déclinaison d'indicateurs de bien-être au niveau régional à partir des données de l'enquête SRCV. L'interprétation des indicateurs régionaux de bien-être ainsi obtenus y est donc limitée, et ce d'autant plus qu'une rupture de série, liée à un changement de l'ordre des questions de l'enquête, a été identifiée pour l'édition 2013 de l'enquête (nous gardons toutefois l'ensemble des millésimes pour conserver une taille d'échantillon suffisante). La lecture devra donc se concentrer sur les éléments méthodologiques, qui pourront reposer sur d'autres sources de données, et être appliqués à la régionalisation d'autres indicateurs.*

Auteurs : Adam Baïz (SDES au moment de l'étude) et Pierre Villedieu (Ensae).
Maquettage : Claude Baudu-Baret

Remerciements : Pascal Ardilly (Insee), Henri Martin (SDES) et Frédéric Vey (SDES) pour leurs relectures et apports

Sommaire

Introduction

Face au ralentissement de la croissance et les crises économiques, sociales et environnementales, de nombreux indicateurs alternatifs au PIB ont émergé *(source 1)*. En particulier, et parmi les indicateurs de bien-être, se distinguent les indicateurs subjectifs d'une part, qui s'appuient pour l'essentiel sur des enquêtes d'opinion, et les indicateurs objectifs d'autre part, qui, de manière plus normative, identifient les déterminants du bien-être et les traduisent au travers de variables tangibles telles que le taux de chômage, la qualité de l'air ou encore la vitalité associative. À l'aune des données massives, et pour mieux représenter les réalités locales du terrain, les indicateurs de bien-être sont calculés sur des unités géographiques de plus en plus fines. Si les indicateurs objectifs se prêtent bien à cet exercice de régionalisation, les indicateurs subjectifs restent très peu disponibles à des niveaux infra-nationaux. Dans une visée méthodologique, la présente étude propose de régionaliser les variables de satisfaction de l'enquête Statistique sur les ressources et conditions de vie (SRCV) : cette enquête intègre en effet, et depuis 2010, des questions relatives à la satisfaction générale des individus ainsi que des questions relatives à leur satisfaction sur un aspect spécifique de leur vie (travail, logement, loisirs, etc.) ; (*source 2*).

Afin d'estimer au mieux les différents niveaux de satisfaction à l'échelle régionale, et étant donné que la représentativité de l'échantillon interrogé n'est assurée qu'à un niveau national, sont mobilisées et comparées deux méthodes d'estimation dites *sur petits domaines.* Un panorama méthodologique proposé par Pascal Ardilly sur la régionalisation des taux de pauvreté à partir des données SRCV avait déjà mis en évidence à la fois les faibles performances des estimateurs classiques et les gains potentiels apportés par ces méthodes (*source 3)*. La première méthode utilisée ici est une estimation synthétique réalisée via un calage sur marges, tandis que la seconde fait intervenir une modélisation explicite via l'approche de *Fay-Herriot (voir l'encadré méthodologique)*. Nous avons pour cela utilisé les macros CALMAR (CALage sur MARges) de l'Insee d'une part et les macros STAT CANADA de Statistics Canada d'autre part.

Ces deux méthodes d'estimation reposent sur des variables dites « *auxiliaires* » dont les « vraies » valeurs sont connues au niveau géographique considéré (ici la région) et qui sont liées à la variable à estimer : dans la présente étude, ont été introduites diverses variables auxiliaires ayant trait non seulement aux caractéristiques socio-démographiques des individus (âge, sexe, catégorie sociale, diplôme, type de famille, niveau de vie), mais aussi à divers aspects de leur vie (travail, santé, lieu de vie, lien social et vie civique). De plus, et comme ces deux méthodes perdent en robustesse ou en précision à mesure que la taille de l'échantillon diminue, ce sont les anciennes régions françaises qui sont apparues comme le niveau géographique le plus fin et le plus pertinent pour l'estimation *sur petits domaines*, chacune d'elles comportant jusqu'à plusieurs centaines d'individus dans notre échantillon. Enfin, de par leurs spécificités méthodologiques respectives, la méthode de calage sur marges a été mobilisée pour estimer la part de personnes déclarant une satisfaction supérieure à 8 sur 10, tandis que la méthode de modélisation explicite de Fay-Herriot l'a été pour estimer des niveaux de satisfaction moyenne. Ainsi, chacune de ces méthodes pourra apporter un éclairage spécifique, l'une portant sur le haut de la distribution et l'autre sur la moyenne des niveaux de satisfaction.

1 – Présentation des données utilisées

Les variables auxiliaires, d'une part, sont issues de diverses sources de données tandis que les données relatives au bien-être subjectif proviennent de l'enquête SRCV *(tableau 1)*. Le dispositif SRCV correspond à la déclinaison française du système européen EU-SILC (European union-Statistics on income and living conditions). Il s'agit d'un panel rotatif renouvelé au neuvième : ainsi, 12 000 ménages sont interrogés chaque année dont environ 10 000 l'ont déjà été l'année précédente. Les questions relatives au bien-être subjectif (BES) n'ayant été introduites qu'en 2010, l'étendue temporelle à notre disposition est 2010-2014. Les variables d'intérêt de cette étude ont été d'une part, et principalement, les variables dites de satisfaction *(tableau 2)* et, d'autre part, les variables d'importance *(tableau 3)*. Ces variables correspondent aux réponses à des questions qui sont posées de manière strictement identique dans chacune des éditions de l'enquête SRCV. Néanmoins en 2013, l'ordre des questions posées aux enquêtés a été légèrement modifié, avec pour conséquence une rupture de série sur ces variables.

Tableau 1 : variables auxiliaires utilisées dans l'étude

Variable (nombre de modalités)	Champs	Source/Base de donnée
Sexe (2)	Tous les individus	Recensement (2010-2013)
Âge (6)	Tous les individus	Recensement (2010-2013)
CSP (8)	15 ans et plus	Recensement (2010-2013)
Type de famille (5)	Tous les ménages	Recensement (2010-2013)
Diplôme (4)	Non scolarisés de 15 ans et plus	Recensement (2010-2013)
Situation vis-à-vis du travail (8)	15 ans et plus	Recensement (2010-2013)
Niveau de vie (10)	Tous les individus	Filosofi (2010-2013)
Part des 75 ans et plus seuls (3)	Tous les individus	Recensement (2011)
Indice de mortalité comparatif (4)	Tous les individus	État-civil – Recensement (2012)
Participation présidentielles 2012 (4)	Tous les individus	Ministère de l'intérieur (2012)
Accès aux équipements (3)	Tous les individus	Base permanente des équipements-Recensement de la population (2013)

Tableau 2 : variables de satisfaction de l'enquête SRCV

Domaine	Description
Travail	Note de 0 à 10
Santé	Note de 1 ("très bon") à 5 ("très mauvais")
Relation avec les proches	Note de 0 à 10
Logement	Note de 0 à 10
Loisirs	Note de 0 à 10
Satisfaction générale	Note de 0 à 10

Tableau 3 : Variables d'importance de l'enquête SRCV

Domaine	Description
Travail et études Santé Relation avec les proches Lieu de vie Loisirs	note de 1 "très important" à 5 "pas du tout important"

***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

2 – Méthodologie de l'estimation sur petits domaines

Produire des statistiques régionales à partir de l'enquête SRCV ne va pas de soi. En effet, le plan de sondage de l'enquête SRCV n'assure la représentativité de l'échantillon qu'au niveau national ; de plus, la taille de l'échantillon total est en général trop faible pour obtenir des estimations locales suffisamment stables avec les méthodes classiques de type Horvitz-Thompson. En particulier, ces estimateurs classiques, dits « directs » au sens où ils ne s'appuient que sur les observations propres au domaine considéré, ont une variance inversement proportionnelle à la taille du sous-échantillon considéré. Ainsi, si la taille de la population d'intérêt est divisée par 10, alors la variance de l'estimateur associé est multipliée d'autant. Cet écueil se présente lorsqu'il est question de passer, comme ici, d'une estimation nationale (20 000 individus dans l'échantillon SRCV) à son équivalent régional (moins de 500 individus pour certaines régions).

Face à ce problème, les méthodes dites *sur petits domaines* proposent de s'appuyer sur de l'information auxiliaire qui doit être d'une part corrélée à la variable d'intérêt, et qui doit, d'autre part, être connue au niveau du domaine (la région). L'estimation locale sera d'autant plus robuste que les variables auxiliaires sont corrélées à la variable d'intérêt. En outre, les deux méthodes dites « indirectes » utilisées ici font l'hypothèse que la relation qui lie la variable d'intérêt aux variables auxiliaires reste vraie au-delà du simple domaine en question. Cela signifie par exemple que le niveau de vie ou l'accès aux équipements influencent le bien-être déclaré de la même façon en Bretagne ou en Alsace. Cette hypothèse peut se traduire dans un modèle explicite posé entre la variable d'intérêt et les variables auxiliaires (méthode 1), ou dans le cadre d'une modélisation implicite (méthode 2).

2.1. Calage sur marge régionales

La première méthode correspond à un calage sur marges sur les régions. Soit X1, X2, …, Xk des variables auxiliaires dont les valeurs xi,k sont connues pour tous les individus i de l'échantillon {1, …, n} de l'enquête et qui sont liées à la variable d'intérêt Y. Pour chacune de ces variables, est également connue une valeur fiable du total X^{tot} au niveau du domaine *a*. Le calage utilise cette information pour re-pondérer les observations de l'échantillon total de façon à retrouver les marges associées aux différentes variables auxiliaires. Formellement, cela signifie qu'il est question de minimiser la distance entre les poids initiaux de l'enquête, car ils possèdent *a priori* de bonnes propriétés statistiques, et les nouveaux poids de façon à ce que ces derniers respectent les équations de calage :

$$\forall k \in \{1, ..., K\}, \sum_{i=1}^{n} x_{i,k} * w_{i,a} = X_{k,a}^{tot}$$

Il est important de rappeler que la procédure est réalisée autant de fois qu'il y a de domaines. À partir du jeu de poids ainsi obtenu au niveau régional, l'estimation du total de la variable d'intérêt *Y* dans la région *a* est obtenue par la somme de l'ensemble des *y* de l'échantillon, pondérées par leur poids associés à cette région :

$$Y_a^{est} = \sum_{i=1}^{n} y_{i,a} * w_{i,a}$$

2.2. Méthode Fay-Herriot

Comme évoqué plus haut, la première méthode repose sur une modélisation implicite du lien entre les variables auxiliaires et la variable d'intérêt. La seconde méthode repose elle sur une modélisation explicite, ici au niveau individuel. En plus des aléas classiques d'un modèle économétrique (terme ea,i), le modèle fait ici intervenir un effet aléatoire ou « effet domaine » qui permet de capter de potentielles spécificités régionales (terme va) :

$$Y_{a,i} = X_{a,i}^T \beta + v_a + e_{a,i}$$

À partir de ce modèle appartenant à la classe des modèles linéaires mixtes, l'estimateur final peut s'écrire sous la forme suivante :

$$\hat{\bar{Y}}_a^{FH} = \gamma_a [\bar{y}_a + (\bar{X}_a - \bar{x}_a)^T \tilde{\beta}] + (1 - \gamma_a) \bar{X}_a^T \tilde{\beta}$$

Cette forme fait apparaître la nature hybride de l'estimateur puisqu'il s'agit d'une moyenne pondérée entre un estimateur synthétique (équivalent à la première méthode) et d'un estimateur plus « direct » privilégiant les observations du domaine considéré. Le poids (noté gamma « γ ») accordé à cette partie directe sera d'autant plus grand que la taille du domaine est grande dans l'échantillon, et que cette région se distingue effectivement des autres. Enfin, cette méthode présente l'intérêt de permettre le calcul de variance des estimateurs dans le cas de plan de sondage suffisamment simples.

Pour plus d'informations sur les méthodes *sur petits domaines*, le lecteur est invité à se référer à la note méthodologique de l'Insee *(source 3)*.

3 – Mise en œuvre de l'estimation

Mettre en œuvre une procédure d'estimation *sur petits domaines* demande un effort particulier dans la préparation des données et la construction de variables auxiliaires pertinentes. Il s'agit non seulement de s'assurer que chacune des variables auxiliaires relève du même *concept* et du même *champ* que sa contrepartie dans l'enquête, mais aussi de vérifier que chacune d'elles est bien corrélée à la variable d'intérêt du modèle.

3.1. Construction et cohérence des variables auxiliaires

Deux types de variables auxiliaires ont été utilisées dans l'étude : les variables définies au niveau individuel d'une part (sexe, âge, CSP, type de famille, diplôme, situation vis-à-vis du travail et niveau de vie) et les variables définies au niveau infra-régional d'autre part (proportion des plus de 75 ans vivant seuls, indice de mortalité comparatif ICM, participation électorale, accès aux équipements). Contrairement aux variables du premier type, les variables du second type consistent à appliquer une même valeur de référence à l'ensemble des individus d'un même territoire de vie[1], ce qui peut causer des erreurs d'attributions : ainsi, un individu peut par exemple être en parfaite santé et pourtant vivre dans un territoire où l'ICM est relativement mauvais. Mais en raison de la multiplicité des territoires à l'échelle d'une région, ces erreurs se compensent globalement.

De plus, il a été question de vérifier que chacune des variables auxiliaires est définie de la même façon dans l'enquête SRCV et dans la source exhaustive dont elle est issue. Pour certaines variables comme le sexe ou l'âge, cette condition est garantie. Pour d'autres variables (diplôme, situation vis-à-vis du travail, etc.), les modalités qui s'y rattachent peuvent varier d'une source à l'autre. Des études comparatives spécifiques ont permis de conclure à la cohérence entre les différentes sources de données.

3.2. Pertinence de la modélisation

Pour réellement accroître la qualité des estimateurs, les variables auxiliaires mobilisées doivent être suffisamment corrélées à la variable d'intérêt, et doivent en l'occurrence constituer autant de déterminants potentiels du bien-être.

À l'instar de l'étude sur le bien-être du Commissariat général à l'égalité des territoires (CGET), les variables auxiliaires ont été choisies parmi des variables socio-démographiques : âge, sexe, CSP, type de famille, diplôme, et niveau de vie[2] *(source 4).* Pour les compléter, il a été décidé de représenter les cinq dimensions usuelles du bien-être (la situation vis-à-vis du travail, la santé, le lien social, le lieu de vie et la vie civique) à travers un jeu complémentaire de variables auxiliaires. Ainsi, le

[1]) S'affranchissant des limites des unités urbaines, les territoires de vie sont définis découpent les bassins de vie de plus de 50 000 habitants pour mieux rendre compte de la diversité de la qualité de vie au sein des territoires les plus urbanisés. La France métropolitaine est ainsi constituée de 2 677 territoires de vie, les bassins de vie de moins de 50 000 habitants étant conservés tels quels.

[2]) Le niveau de vie correspond au revenu disponible par unité de consommation. Il est donc relatif au ménage et est identique pour chacun des individus du ménage.

travail par exemple, a été approché par le croisement de quatre variables, afin de rendre compte de l'ensemble des situations possibles : actif/inactif * emploi * type de contrat * temps complet/partiel.

Ces variables ont été choisies en vertu de la disponibilité des données correspondantes, et en vertu de leur lien statistique avec le bien-être, tel que documenté par la littérature académique. Par exemple Dolan et al. (2008) montrent que les variables de revenu, d'âge, de santé, de chômage et de relations sociales et familiales ont un lien relativement fort et robuste avec le niveau de satisfaction déclaré par les individus *(source 6).*

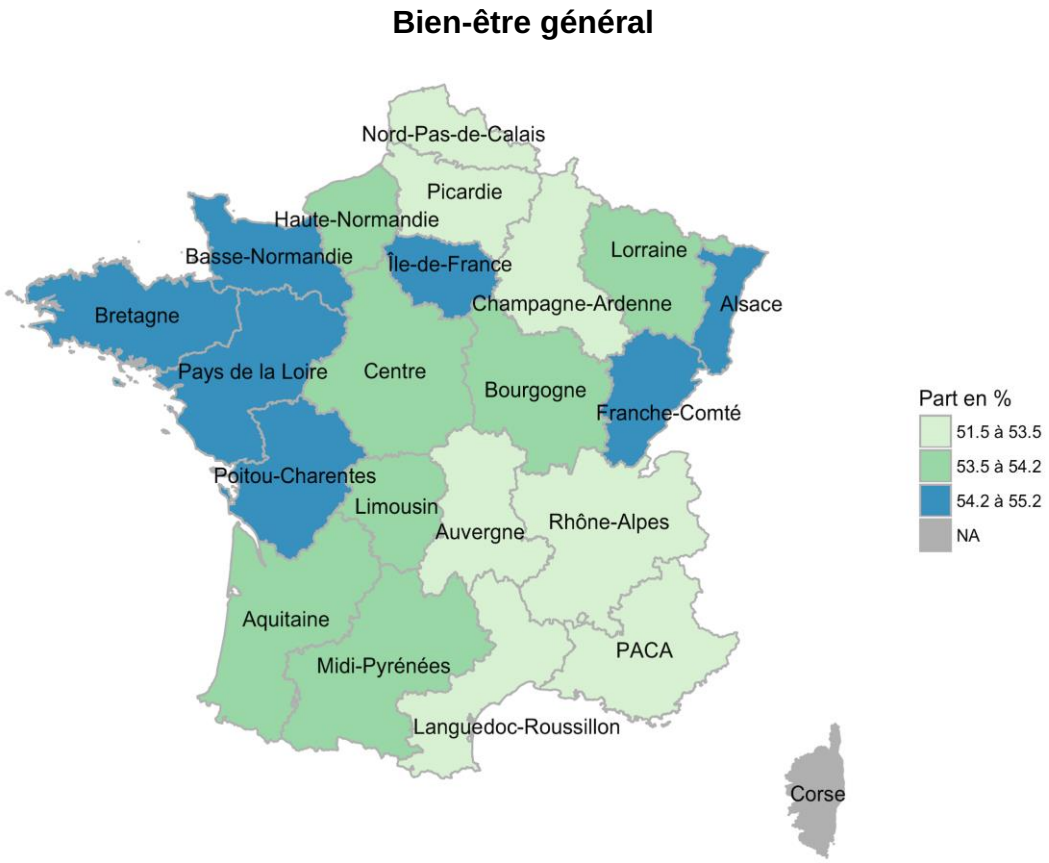
4 – Principaux résultats

Cette partie présente les résultats obtenus avec chacune des deux méthodes présentées plus haut : la méthode de calages sur marges d'une part et la méthode de Fay-Herriot d'autre part.

4.1. Analyse des résultats obtenus avec la méthode de calages sur marges

Il est ici question d'estimer la part de personnes relativement satisfaites, soit plus précisément la part des personnes déclarant une satisfaction supérieure ou égale à 8 sur 10. La *carte 1* présente les estimations régionales pour *la satisfaction dans la vie en général* et la *carte 2* celles pour les sous-indices étudiés[3]. Les estimations correspondent à une moyenne des résultats annuels sur la période de disponibilité des données (2010-2013).

Carte 1 : part des personnes déclarant une satisfaction dans la vie supérieure ou égale à 8 sur 10 (satisfaction générale) – Estimation via un calage sur marges



***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

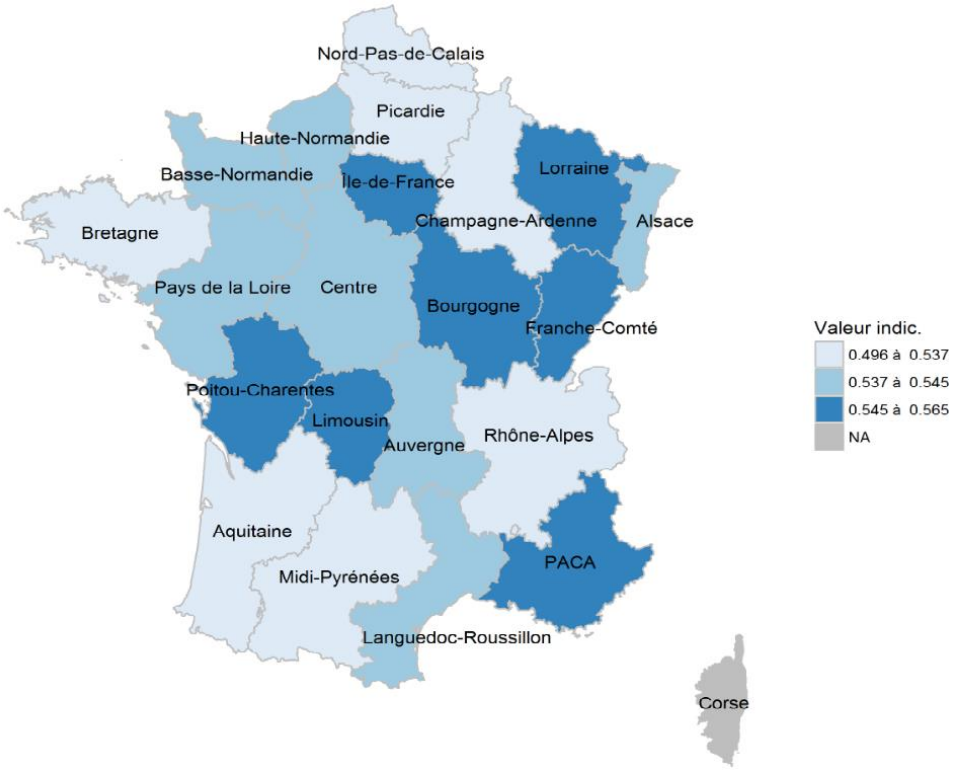
[3]) Il est à noter que la Corse n'a pas donné lieu à une estimation à cause de la très faible taille du sous-échantillon associé à cette région.

Concernant la *satisfaction dans la vie* en général, il apparaît une relative proximité des taux de satisfaction (51 % à 55 %), qui s'explique notamment par le tassement des disparités intra-régionales découlant de l'agrégation des données. Il ressort néanmoins que les régions où la satisfaction générale est la plus faible sont celles du Nord et du Sud-Est, alors que c'est en l'Île-de-France et dans les régions de l'Ouest que l'on trouve les parts de personnes satisfaites les plus élevées.

Concernant les *satisfactions thématiques* (liées au travail, à la santé, etc.), il apparaît que les taux de satisfaction estimés pour l'Île-de-France sont relativement élevés pour les loisirs, l'emploi, ou la santé, mais plus faibles pour les relations avec les proches ou le logement. Dans l'ensemble, il ressort que les régions (Nord – Pas-de-Calais, Picardie, Champagne-Ardenne) dont la satisfaction dans la vie est relativement faible cumulent aussi les taux de satisfaction faibles dans les différentes dimensions du bien-être étudiées et inversement pour les régions où les taux sont relativement élevés (Île-de-France, Pays de la Loire, Poitou-Charentes). Ceci indique une corrélation, au niveau régional, marquée entre la satisfaction que les individus déclarent dans la vie en général et les satisfactions thématiques qui sont liées à des dimensions concrètes de la vie.

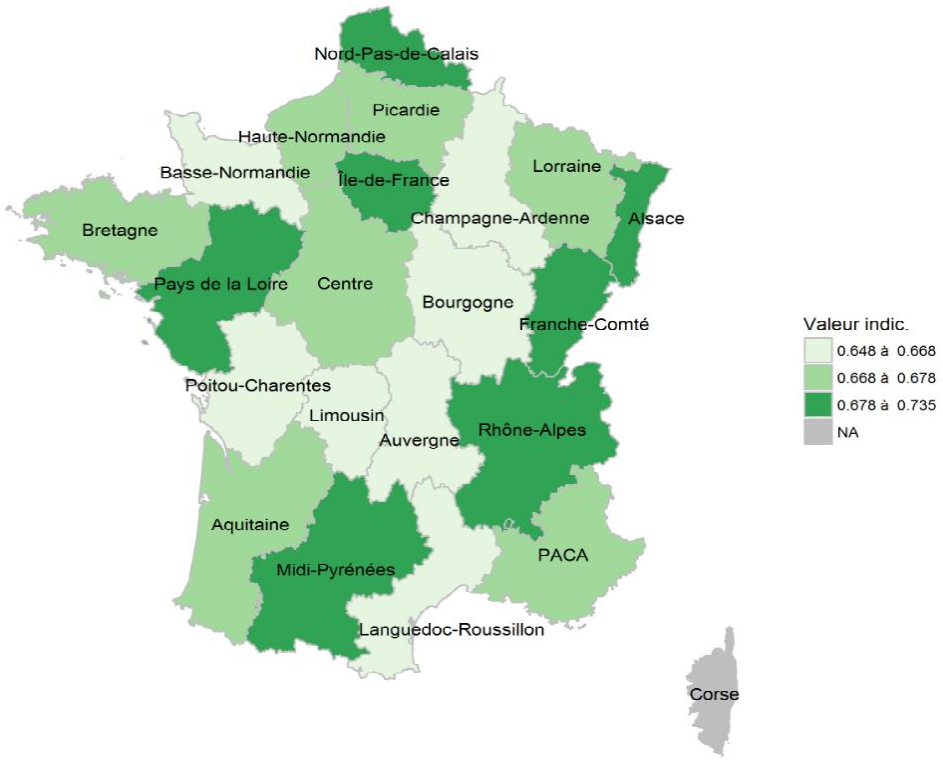
Cartes 2 : part des personnes déclarant une satisfaction dans la vie supérieure ou égale à 8 sur 10 (satisfactions thématiques) – Estimation via un calage sur marges

Travail

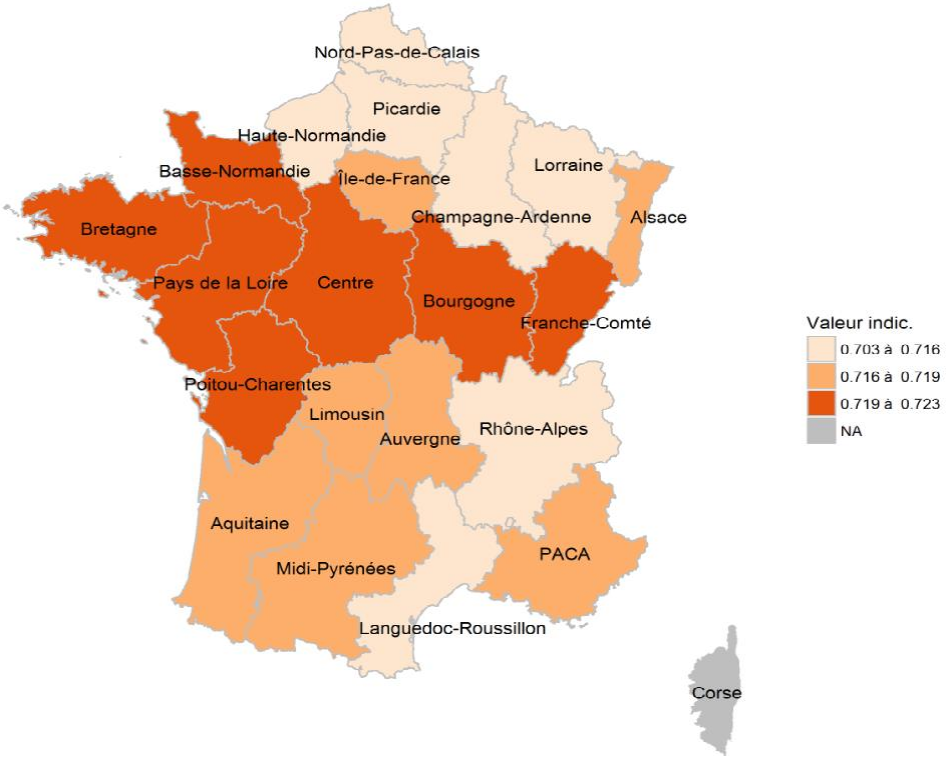


***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

Santé

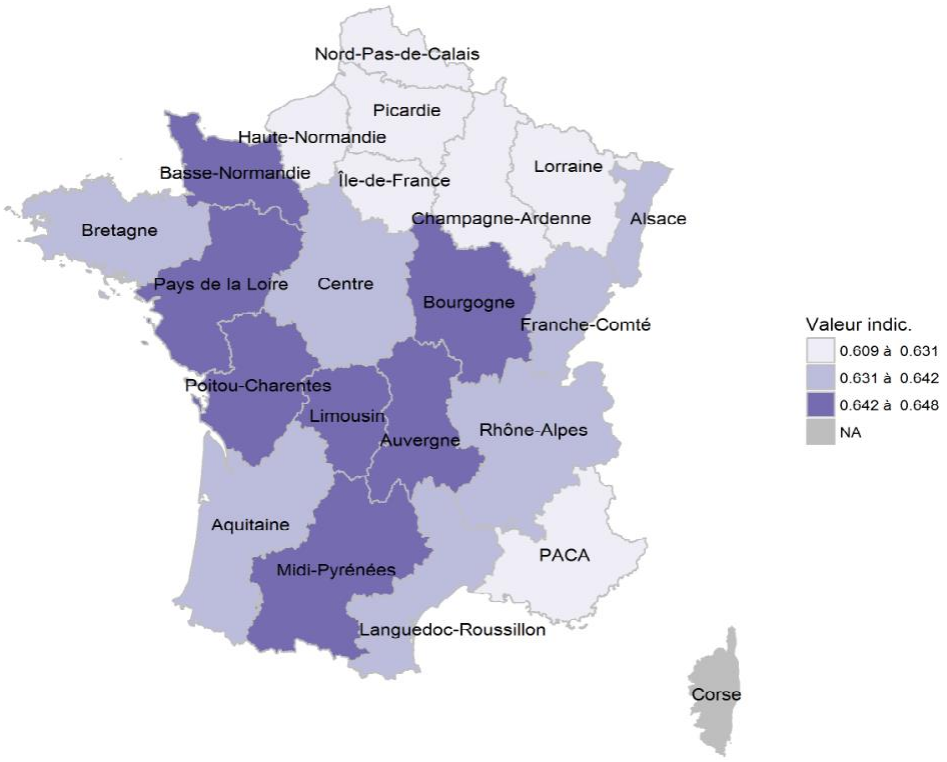


Relations avec les proches

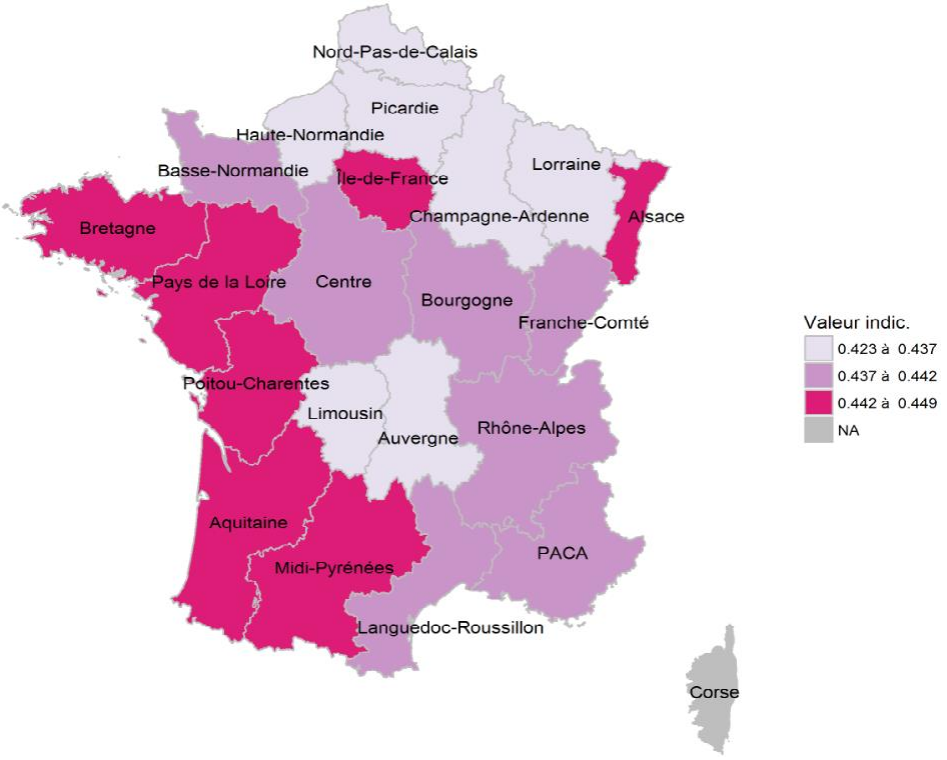


***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

Logement



Loisirs

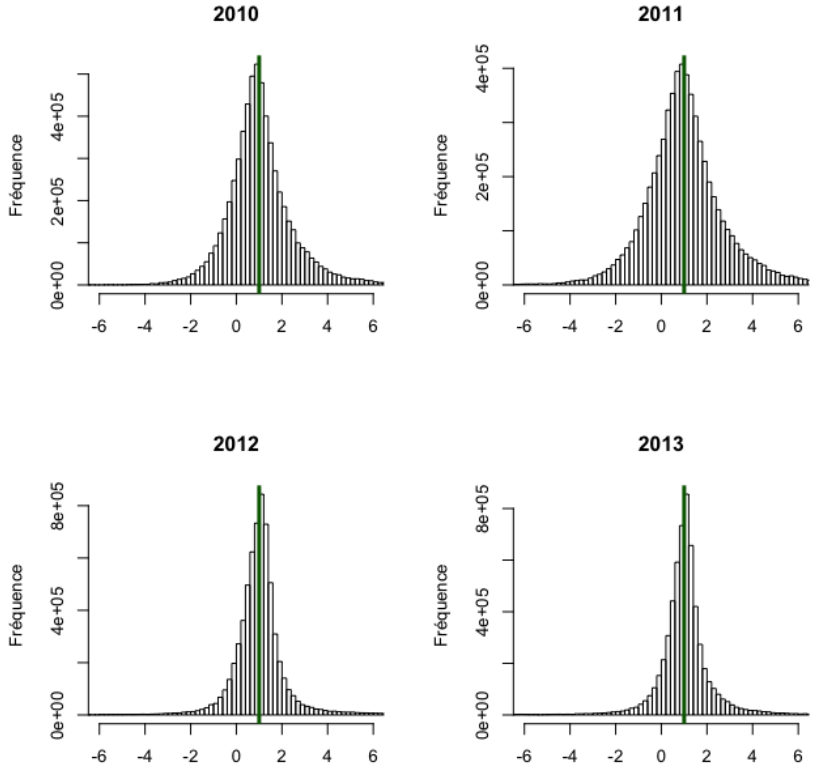


***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

Afin de mieux appréhender la fiabilité de ces résultats, plusieurs tests ont été envisagés.

Un premier test usuel consiste à inspecter l'histogramme des rapports de poids initiaux avec les poids calés *(figure 1).* La distribution doit être centrée en 1, puisque le principe du calage sur marges est d'obtenir de nouveaux poids qui soient les plus proches possibles des anciens. Ainsi la forme de la distribution et l'épaisseur des queues nous renseigne sur la difficulté à concilier cette contrainte avec les équations de calage.

Figure 1 : appréciation du biais : rapport des poids calés sur les poids non calés



***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

Dans le cas présent, les distributions associées aux différentes années sont bien centrées en 1 et les rapports de poids élevés sont relativement peu fréquents.

Il est également possible d'apprécier si l'estimateur synthétique a introduit un biais manifeste, soit numériquement, soit par une méthode graphique.

Numériquement, il s'agit de comparer les estimations nationales des variables d'intérêt en utilisant les poids initiaux d'une part, ce qui aboutit à des estimations non biaisées *a priori*, et celles issues des poids calés d'autre part. Si les estimations diffèrent trop, il est probable qu'un biais ait été introduit. Le *tableau 4* présente les résultats relatifs à la variable de *satisfaction dans la vie*. Il apparaît qu'en dehors des modalités extrêmes (i.e. qui ne concernent que très peu d'individus), le biais relatif est faible, inférieur à 5 % en général. Les résultats sont comparables pour les sous-indices considérés dans cette étude (travail, loisirs, logement, etc.).

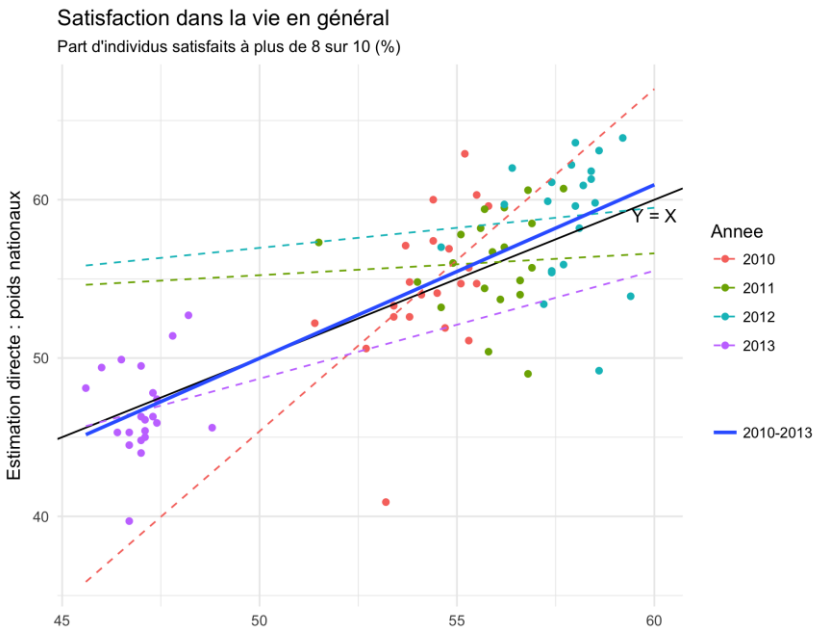
Tableau 4 : appréciation numérique du biais

Modalité	2010	2011	2012	2013
0	- 1,7	6,3	6,6	7,5
1	- 6,4	10,4	32,3	40,3
2	- 1,9	- 1,7	9	- 12,1
3	0,9	10,6	- 9	0,6
4	3,8	4,4	6	12,5
5	2,3	4	6,8	3,7
6	- 0,6	4,3	4,6	2,1
7	1,4	3	3,3	- 2
8	0,4	3,7	1,1	3,9
9	- 0,7	5	2,4	1,6
10	- 2,7	1,8	0,1	7,3

***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

Enfin, il est aussi possible d'apprécier ce biais de façon graphique. Dans le repère formé par les estimations de chaque type, il s'agit de comparer la droite de régression associée au nuage de points avec la droite Y = X. Une différence importante entre leurs pentes respectives signale le plus souvent l'introduction d'un biais. La *figure 2* montre que si l'on considère la période 2010-2013, aucun biais substantiel n'apparaît, les deux droites étant très proches. En revanche, pour une année donnée, les écarts de pentes sont nettement plus grands. L'agrégation des différentes périodes apparaît donc utile pour aboutir à une estimation plus fiable.

Figure 2 : appréciation graphique du biais

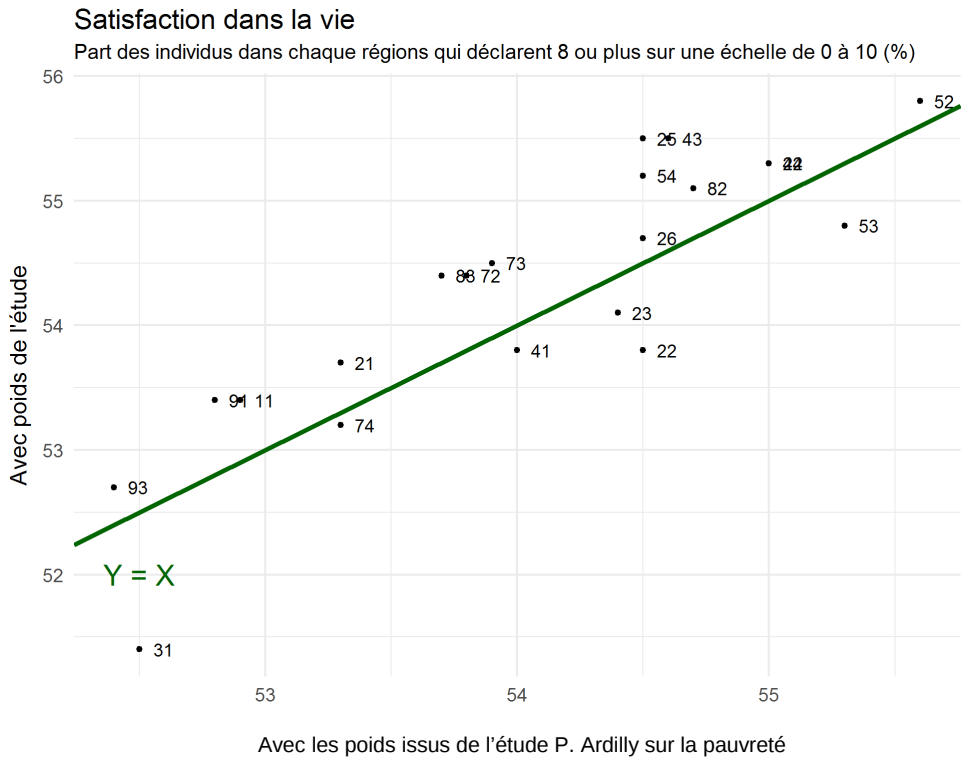


Estimation synthétique : poids régionaux

Note : les estimations relatives à l'édition 2013 de l'enquête SRCV (points violets) sont éloignées des estimations relatives aux autres millésimes. Cela s'explique par une modification de l'ordre des questions au sein du questionnaire intervenue cette année-là.
***Source :** éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

Enfin, les résultats ont été comparés avec les estimations obtenues par P. Ardilly sur la régionalisation des taux de pauvreté en France à partir de la même enquête SRCV *(source 5).* Les démarches sont identiques à la différence près que la présente étude introduit davantage de variables auxiliaires (en l'occurrence celles relatives aux différentes dimensions du bien-être). La *figure 3* représente les estimations régionales selon le jeu de poids obtenu dans la présente étude et le jeu de poids obtenu par P. Ardilly. Il apparaît de même une répartition resserrée du nuage de point autour de la droite Y = X, ce qui permet d'infirmer l'introduction d'un biais.

Figure 3 : comparaison avec les estimations issues des poids d'Ardilly (2016)



***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

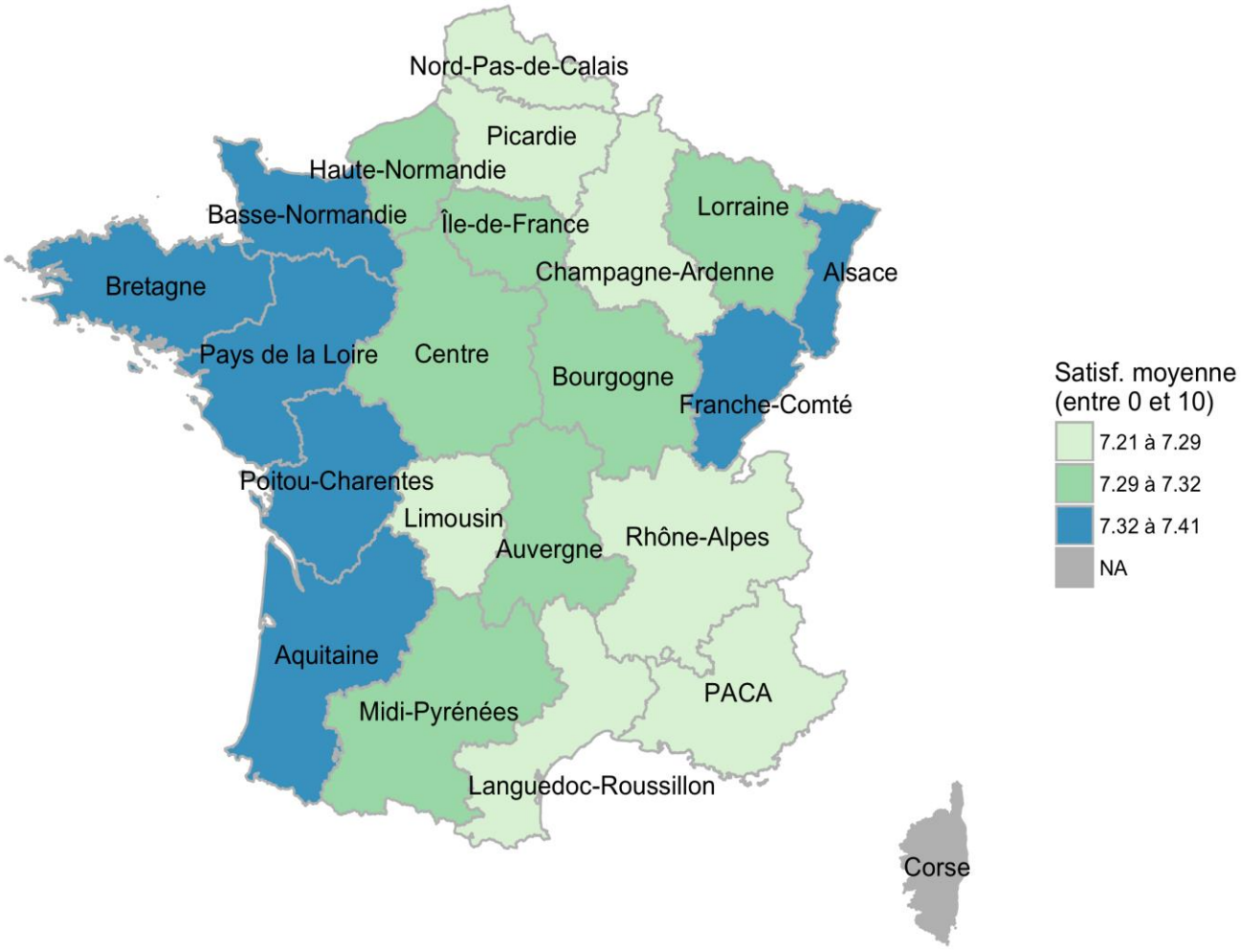
4.2. Analyse des résultats obtenus avec la modélisation explicite Fay-Herriot

Pour des raisons techniques notamment[4], la méthode de Fay-Herriot a été mobilisée pour estimer la satisfaction moyenne des individus. La *carte 3* présente les résultats obtenus concernant la satisfaction dans la vie en général[5]. Il est à noter, une fois encore, que la dispersion des estimations est restreinte, même en tenant compte des spécificités régionales (via l'effet aléatoire introduit dans le modèle) : ce qui est remarquable, c'est que la hiérarchie entre les régions est quasiment la même selon la première méthode ou cette seconde, comme indiqué dans la prochaine section.

[4]) Les macros de Stat Canada utilisées pour mettre en œuvre ces estimations permettent en effet de gérer uniquement des variables d'intérêt quantitatives.

[5]) Afin d'alléger la présentation, les cartes associées aux satisfactions thématiques ne sont pas présentées ici et sont disponibles sur demandes.

Carte 3 : niveau moyen de satisfaction déclarée dans la vie (sur une échelle de 1 à 10) - Estimation avec la méthode Fay-Herriot



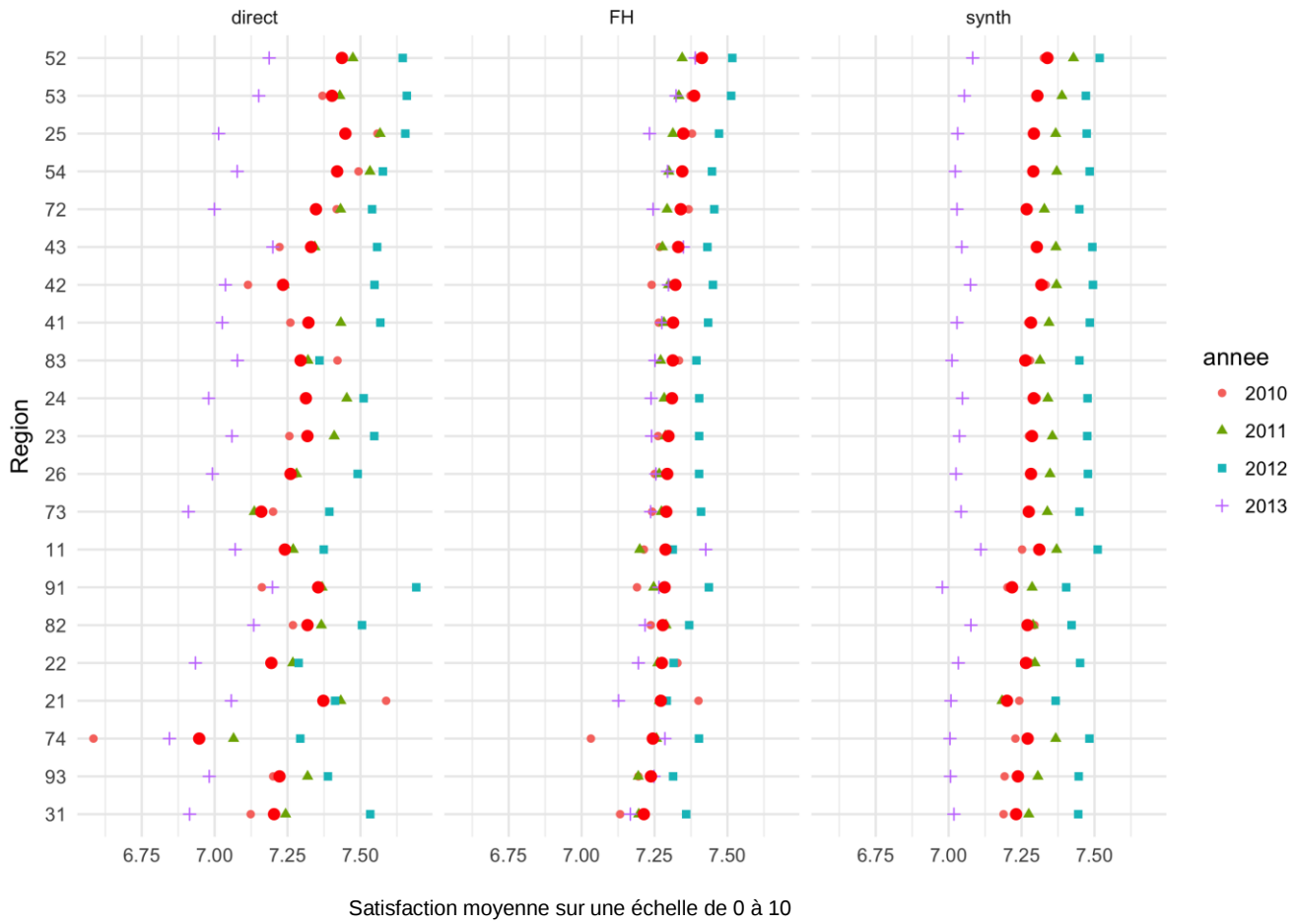
***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

4.3. Comparaison des méthodes d'estimation

Partant d'estimations assez instables et peu représentatives du domaine en question, la présente étude a mobilisé deux méthodes distinctes afin d'améliorer la qualité des estimations. En particulier, comme indiqué plus haut, ces méthodes permettent de régler le problème de l'instabilité (variance) de l'estimateur. En contrepartie, et en ce sens qu'elles constituent des *modélisations*, elles présentent toutes les deux le risque de l'introduction d'un biais. Pour en juger, sont dans cette partie comparés les résultats de satisfactions moyennes à travers les deux méthodes.

La *figure 4* présente les valeurs estimées de satisfaction moyenne la *satisfaction dans la vie en général*. Les estimations sont rangées verticalement selon l'ordre obtenu par la méthode F.H..

Figure 4 : comparaison des estimations pour la variable satisfaction dans la vie



***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

Le lien statistique entre les estimations d'une méthode à l'autre apparaît relativement fort. Le *tableau 5* donne, pour chaque variable étudiée, les corrélations (Pearson et Spearman) entre les estimations issues de chaque estimateur. Les résultats démontrent une certaine robustesse des méthodes : autrement dit, une région bien classée selon l'une des méthodes à des grandes chances de l'être également avec l'autre méthode.

Tableau 5 : corrélations entre les estimations issues des deux méthodes : Fay-Herriot et synthétique

		Vie en général	Santé	Rel. humaines	Travail	Loisirs	Logement
Pearson	coeff.	0.69	0.87	0.59	0.59	0.75	0.87
	p-val. (5%)	0.000472	2.992e-07	0.00493	0.005246	9.817e-05	3.852e-07
	Int. de confiance	[0.38;0.87]	[0.70;0.95]	[0.21;0.81]	[0.21;0.81]	[0.47;0.89]	[0.69;0.94]
Spearman	coeff.	0.71	0.71	0.53	0.49	0.64	0.71
	p-val. (5%)	0.00048	0.00045	0.01419	0.0244	0.00225	0.00048

Note : les corrélations ont été calculées à partir de valeurs moyennées sur la période 2010-2013 pour chaque dimension.
***Source** : éditions 2010, 2011, 2012, et 2013 de l'enquête SRCV, Insee. Calculs : SDES*

La corrélation de Pearson est comprise entre 60 % et 87 % selon la dimension considérée et est toujours significative (à 1 %). Le fait que ce lien reste significatif et du même ordre de grandeur pour la corrélation de Spearman (environ 49 % à 70 %) nous permet de conclure que cela n'est pas lié à des valeurs extrêmes.

Qu'il s'agisse du niveau de satisfaction moyen ou de la part de personnes déclarant une satisfaction supérieure à 8 sur 10, la régionalisation des variables de satisfaction révèle des différences circonscrites entre les régions, mais presque toujours la même hiérarchie entre ces dernières, quelle que soit la dimension considérée et quelle que soit la méthode d'estimation retenue.

5 – Conclusion

Afin de régionaliser des indicateurs de bien-être subjectif, aujourd'hui uniquement disponibles au niveau national, cette étude met en œuvre deux méthodes d'estimations *sur petits domaines* : l'estimation synthétique (par calage sur marge) et la modélisation de type Fay-Herriot. Ainsi, ce travail a vocation à alimenter la littérature mettant en application de façon concrète les méthodes d'estimations *sur petits domaines*. Il semble en effet que le recours à ces méthodes puisse être une approche intéressante dans un contexte où l'on cherche à produire des indicateurs de plus en plus localisés sans pour autant pouvoir augmenter suffisamment la taille des échantillons.

L'intérêt de l'étude réside également dans la production de statistiques. Les résultats obtenus ont notamment mis en évidence certaines régions qui semblent cumuler des taux de satisfactions plus ou moins importantes selon les dimensions étudiées (travail, relations humaines, santé, loisirs, logement). L'amplitude de ces différences reste néanmoins relativement faible. La méthode d'estimation contribue mécaniquement à ce tassement des disparités entre les domaines. Néanmoins, cela suggère aussi que le niveau régional est sans doute trop agrégé pour faire ressortir des différences notables en termes de bien-être subjectif. Les hiérarchies régionales découlant de cette étude doivent donc être prises avec précaution. Et ce d'autant plus qu'une rupture de série, liée à un changement de l'ordre des questions de l'enquête, a été identifiée pour l'édition 2013. D'une façon générale, le calcul de précision dans le cadre d'enquête complexe comme SRCV est déjà difficile pour des estimateurs classiques, il devient donc encore plus complexe dans le cadre d'estimation *sur petits domaines*. Pour cette raison, tester la significativité des différences inter-régionales nécessiterait un travail spécifique qui n'a pu être mené dans le cadre de cette étude.

Les pistes d'amélioration concernant cette étude sont de trois types. D'abord, des variables auxiliaires supplémentaires pourraient être ajoutées pour accroître la capacité prédictive du modèle. Il serait également intéressant de comparer différents scénarios d'estimations basés sur plusieurs groupes de variables auxiliaires couvrant les mêmes déterminants et ainsi tester la robustesse des résultats au choix des variables auxiliaires. Sur le plan de la méthode utilisée, une troisième approche est possible : il s'agit des méthodes basées sur des estimateurs optimums, dits « bayésiens empiriques », qui permet d'envisager des estimateurs non linéaires. Enfin, l'ajout des vagues de l'enquête SRCV encore indisponibles au moment de l'étude permettrait de gagner en robustesse.

Sources

1 – *Comment expliquer la longévité de l'indicateur du PIB face aux indicateurs alternatifs de richesse ?*, Adam Baïz et Pierre Villedieu, Document de travail n° 37 CGDD/SDES, février 2018.

2 – Insee www.insee.fr/fr/metadonnees/definition/c1749.

3 – *Panorama des principales méthodes d'estimation sur les petits domaines*, Pascal Ardilly, Document de travail, Insee, 2006.

4 – *Qualité de vie, habitants, territoires*, Rapport de l'Observatoire des territoires, CGET, 2014.

5 – *Estimation régionale de taux de pauvreté par une méthode de calage*, Pascal Ardilly, Insee, 2016.

6 – *Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being*, Dolan et al., Journal of Economic Psychology, 2008.

Ministère de la Transition écologique et solidaire
Commissariat général au Développement durable
Service des données et études statistiques
Tour Séquoia
92055 La Défense cedex
Courriel : diffusion.sdes.cgdd@developpement-durable.gouv.fr

www.statistiques.developpement-durable.gouv.fr