ESTIMATIONS RÉGIONALISÉES DE L'ENQUÊTE TIC 2016 ENTREPRISES PAR DES MÉTHODES SUR PETITS DOMAINES

Nadège PRADINES (*)

(*) Insee, Direction des Statistiques d'Entreprises

nadege.pradines@insee.fr

Mots-clés *:* petits domaines, tic, modèles, calage, entreprises.

Résumé

L'enquête sur les technologies de l'information et de la communication et le commerce électronique (TIC) auprès des entreprises apprécie la place des outils nouveaux dans les relations externes de l'entreprise et dans leur fonctionnement interne. Ses résultats sont obtenus au niveau national. Toutefois, des acteurs locaux manifestent de plus en plus d'intérêt pour ces sujets, sans qu'il soit, pour l'heure, possible de leur fournir des indicateurs équivalents aux indicateurs nationaux de qualité suffisante.

Plusieurs outils permettant des estimations sur petits domaines tentent de remédier à cette insuffisance d'information sur les régions.

D'une part, des estimateurs directs et des estimateurs indirects synthétiques à partir de l'information d'enquête sont produits par des calages sur marges régionales, au moyen de la macro SAS Calmar. La variabilité des estimateurs directs peut être calculée au moyen de la macro Everest développée par l'Insee. Si les calages sur marges en vue de produire des estimateurs directs donnent des résultats appréciables, les estimateurs indirects ainsi produits semblent fortement biaisés.

D'autre part, des estimateurs composites sont produits par un modèle linéaire général au niveau des régions (modèle de « Fay-Herriot »). Ces estimateurs sont une moyenne pondérée d'estimateurs directs et d'estimateurs modélisés à partir d'une information auxiliaire sur les régions, donnant la préférence à l'estimation directe quand elle est de bonne qualité et à l'estimation modélisée sinon. Les modèles de Fay-Herriot sont produits au moyen d'une macro SAS développée par l'institut Statistics Canada. Ces estimateurs sont théoriquement sans biais : une approche graphique confirme que c'est le cas pour la plupart des paramètres estimés. La variabilité des estimateurs est réduite par rapport aux estimateurs directs.

Toutes ces méthodes supposent une information auxiliaire à l'information TIC. Pour les opérations de calage, cette information doit être disponible au niveau de l'individu. Il a donc fallu, en amont, rassembler et résumer cette information, depuis des sources de diverses origines (Insee : Sirus, DADS, Fare, Fare localisé, Lifi, Liasses fiscales ; Arcep : données de couverture mobile et données de couverture très haut débit domiciliée). La quantité d'information auxiliaire et d'enquête disponible sur les DROM a contraint les travaux sur le champ métropolitain.

Dans l'ensemble, la structure relative des résultats régionaux est conservée entre les estimateurs directs et les estimateurs composites, à l'exception notable de la Corse, dont les estimateurs directs étaient de très mauvaise qualité. La précision des estimations a été améliorée par le recours à des méthodes d'estimation sur petits domaines. Toutefois, elle ne permet pas de mettre en évidence de différences significatives entre régions dans les usages des TIC par les sociétés.

Abstract

Local policy makers show an increasing interest in the ICT usage in their territory. However, the survey on information and communication technologies (ICT) usage and e-Commerce in enterprises provides information on ICT usage at a national level only. The sampling design of the survey does not allow regional estimations with a sufficient quality. Three small-area-estimation methods are used to make up for this lack of information: first, a new calibration of the database on regional margins improves the quality of direct estimations; secondly, a Fay-Herriot composite estimation based on the previous calibrated estimations improves their quality. Those two methods help a lot to improve the precision of the estimations. A third method, a model-based calibration, produces a synthetic estimation, but this method suffers from a strong bias. No method proves precise enough to show significative differences between regions in ICT usage in enterprises.

1. Introduction

En 2016, en France, 93 % des sociétés de 10 personnes ou plus disposent d'une connexion fixe à internet en haut débit. 17 % d'entre elles achètent des services de cloud computing payant et 53 % ont effectué des achats sur le web.

Ces résultats sont obtenus grâce à l'enquête sur les technologies de l'information et de la communication et le commerce électronique (TIC) auprès des entreprises, menée par l'Institut national des statistiques et des études économiques (Insee). L'enquête TIC vise à mieux connaître l'informatisation et la diffusion des technologies de l'information et de la communication dans les entreprises. Elle cherche notamment à apprécier la place des outils nouveaux dans les relations externes de l'entreprise (Internet, commerce électronique) et dans leur fonctionnement interne (réseaux, systèmes intégrés de gestion). Cette enquête, annuelle, répond à un règlement européen imposant la fourniture d'indicateurs nationaux. À la Commission européenne, ces résultats sont utilisés pour un parangonnage des pays européens, sur la foi d'un questionnement *a priori* semblable et d'une méthodologie *a priori* de qualité.

Toutefois, les acteurs européens et nationaux ne sont pas les seuls à s'intéresser à la pénétration des technologies de l'information et de la communication et du commerce électronique dans les sociétés. Des acteurs locaux manifestent de plus en plus d'intérêt pour ces sujets, sans qu'il soit, pour l'heure, possible de leur fournir des indicateurs équivalents aux indicateurs nationaux. Alors que l'équivalence du questionnement semble facile à accomplir, l'équivalence de la qualité est un but plus difficile à atteindre. Deux options se présentent : gérer la problématique en amont, en élargissant l'échantillon de l'enquête et adaptant son plan de sondage pour garantir une représentativité des résultats au niveau local ; gérer la problématique en aval, bon gré mal gré, en ajustant les résultats obtenus avec l'échantillon classique *via* plusieurs modèles.

Si la première option serait de loin la plus satisfaisante en matière de précision des estimateurs, c'est aussi la plus coûteuse pour l'Insee. Une seule extension de l'échantillon a été réalisée à ce jour pour l'enquête TIC, en 2013, pour la région Basse-Normandie. Les résultats nationaux et bas-normands n'étaient pas significativement différents. Il a été jugé peu rentable le surcoût d'une telle extension de l'échantillon, *a fortiori* sur l'ensemble des régions de France.

Laissé à la seconde option, à savoir traiter les données déjà disponibles dans une approche géographique, de nombreux obstacles se dressent encore sur le chemin du statisticien. L'objet de cet article est de recenser ces obstacles et, moyennant des arbitrages récurrents, les franchir et aboutir à des estimateurs locaux des principales variables de l'enquête TIC menée en 2016.

Après une description du contexte méthodologique, une seconde partie présente les arbitrages faits sur le champ et l'enrichissement de la base avec de l'information auxiliaire. Dans la troisième partie, les estimations directes sont présentées, y compris les arbitrages nécessaires au choix des marges de calage pour la production d'un estimateur calé sur des marges régionales. La dernière partie déroule les estimations issues de modèles : en premier lieu, les estimations synthétiques issues d'un modèle descriptif ; en second lieu, des estimations issues de modèles de Fay-Herriot, alimentés par les estimations régionales directes produites. Dans chaque partie, une étude de la qualité des estimateurs obtenus complète l'analyse, lorsque c'est possible. La conclusion sera l'occasion de discuter les résultats et leurs limites.

2. Le contexte

2.1. Objectifs des travaux

2.1.1. L'enquête TIC

L'enquête TIC 2016, qui s'adresse aux entreprises occupant au moins 10 personnes, est une enquête européenne annuelle obligatoire. L'unité d'interrogation est l'unité légale, identifiée par son numéro SIREN. Les unités légales enquêtées sont des unités françaises, actives, marchandes et exploitantes au 31 octobre 2015. Les unités légales enquêtées sont localisées en métropole ou dans les départements et régions d'outre-mer (DROM).

Le matériel de collecte (questionnaire et notice) et les métadonnées de l'enquête sont disponibles sur le site de l'Insee[1]. La liste des paramètres que l'on souhaite estimer à un niveau local figure en Figure annexe 1.

2.1.2. L'échantillonnage de l'enquête TIC

L'échantillon compte 12 658 unités légales (UL) et 13 entreprises profilées[2].

L'échantillon est sélectionné selon un tirage aléatoire simple stratifié. La stratification adoptée est un croisement entre le secteur d'activité de l'entreprise, sa tranche de personnes occupées et sa tranche de chiffre d'affaires. Les modalités des secteurs d'activité ont des niveaux d'agrégation très divers (de la classe au regroupement de sections). Pour chaque modalité de taille, un seuil de chiffre d'affaires est défini au-delà duquel toutes les unités sont interrogées. Au final, la strate de tirage est la concaténation des modalités des différents critères de stratification décrits ci-dessous.

Les allocations de l'échantillon sont issues d'une allocation mixte, correspondant à une moyenne entre une allocation proportionnelle au nombre d'unités et une allocation proportionnelle au nombre de personnes occupées.

Ces allocations sont contraintes par : un nombre minimum de 10 unités par strate, chaque fois que c'est possible ; une majoration, pour chaque activité, de la demi-longueur de l'intervalle de confiance à 10 % sur les proportions.

L'échantillon national de l'enquête est un panel, renouvelé par moitié tous les ans (hors unités du champ exhaustif qui sont interrogées tous les ans, la moitié de l'échantillon de TIC 2015 continue d'être interrogée pour TIC 2016, à condition d'être toujours dans le champ). Les allocations sont calculées sur l'ensemble de la base, sans distinguer partie renouvelée et partie conservée, puis elles sont converties en taux de sondage, appliqués uniquement sur la partie renouvelée de la base de sondages et les unités nouvelles. Pour cette raison, à chaque strate sont associés deux poids de sondage : le poids « historique » de TIC 2015 pour les unités qui étaient déjà échantillonnées en 2015, le poids de TIC 2016 pour les unités nouvelles.

La base de sondage comporte 191 405 unités.

2.2. Petit domaine : question technique

2.2.1. Le problème de la taille

Un domaine est un sous-ensemble du champ. Des petits domaines (*small area*) sont des sous-ensembles contenant un faible nombre *n* d'unités : secteur fin, découpage géographique fin, etc. Dans le cas de l'enquête TIC, on souhaite travailler sur des domaines de niveau régional.

La théorie des sondages nous assure que l'erreur d'échantillonnage, quand on la mesure par le coefficient de variation, sera d'autant plus grande que le domaine est petit. Il revient au statisticien

[1] https://insee.fr/fr/metadonnees/source/s1273

[2] Ces entreprises profilées seront exclues de l'analyse. Toutefois, leur présence dans l'échantillon signifie que les UL qui les composent ne figurent pas dans la base de sondage, ce qui introduit un léger biais sur le champ.

de définir le seuil de coefficient de variation au-delà duquel une estimation classique ne suffira plus. Pour les présents travaux, faute d'obtenir un coefficient de variation le plus bas possible, la cible d'un coefficient de variation au moins inférieur à 17,5 %, semble raisonnable.

2.2.2. Estimations directes, synthétiques, composites

La classification ci-dessous est une classification simplifiée et non exhaustive des méthodes d'estimation sur petits domaines. Elle s'attache surtout à décrire succinctement les méthodes mises en œuvre au cours du projet (pour un panorama complet, on peut se reporter à (Ardilly, 2006)).

On utilise les notations suivantes :

U est le champ (la population globale)
a est un domaine de U
s est l'échantillon issu de U, de taille *n*
r est l'échantillon répondant, de taille *ñ*
s_a est l'échantillon issu de U et inclus dans *a*, de taille n_a
r_a est l'échantillon répondant inclus dans a, de taille $ñ_a$

Par construction, $0 \leq n_a \leq n$

Y est une le total de la variable d'intérêt sur le champ
Ya est le total de la variable d'intérêt *Y* sur le domaine *a*.

2.2.2.1. Estimations directes

L'estimation directe ne fait appel à aucune information (collectée) relative à des individus se situant hors du domaine. L'estimation la plus simple consiste à calculer pour chaque domaine l'estimateur avec les unités répondantes du domaine et les poids issus du redressement (correction de la non-réponse partielle et totale et calage sur marges nationales). On peut résumer l'estimateur de *Y* sur le domaine *a* à :

$$\hat{Y}_a = \sum_{i \in r_a} \omega_i^{Pise} Y_i$$

avec ω_i^{Pise} le poids « Pise » de l'unité *i*, issu du redressement effectué par le Pôle d'ingénierie statistique d'enquête (Pise). Comme $ñ_a$ est petit, la variance d'échantillonnage est grande et les totaux estimés sont instables. De plus, les totaux ne seront pas forcément cohérents avec le nombre d'unités connues dans le domaine puisque le calage n'a pas été réalisé sur des marges propres au domaine.

Une seconde approche consiste donc à caler les données sur des marges régionales puis calculer l'estimateur calé sur le domaine. Cette approche a surtout du sens si on choisit des variables de calage pour leur lien avec les variables d'intérêt, de façon à diminuer la variance des estimateurs en même temps qu'on assure la cohérence des totaux avec le nombre d'unités connues dans le domaine.

$$\hat{Y}_a = \sum_{i \in r_a} \omega_i^{cal} Y_i$$

Avec un tel estimateur, les moyennes comme les totaux sont cohérents avec les informations connues par ailleurs (nombre d'unités dans le domaine, par exemple). Toutefois, la mise en œuvre d'un calage et sa réussite dépendent de la taille *n* du domaine. Si *n* est trop petit, le calage a peu de chances de converger vers un résultat satisfaisant. Cet estimateur a donc du sens lorsque *n* est « petit mais pas trop ». Malgré le soin apporté à un tel calage, la variance restera liée à n_a.

2.2.2.2. Estimations synthétiques descriptives

Au contraire de l'estimation directe, l'estimation synthétique se fonde sur des hypothèses de comportement reliant le domaine au reste de la population. L'exemple trivial d'une estimation synthétique d'une moyenne est de considérer que « le résultat national est un bon estimateur du résultat dans chacun des domaines ». Il faut de solides convictions pour mettre en œuvre ce modèle : il aboutit à un estimateur « synthétique » identique dans tous les domaines et ne traduit donc aucune spécificité locale.

D'autres modèles d'estimation synthétiques descriptives de totaux ou de moyennes peuvent être défendus avec plus d'arguments. Toutes les modélisations envisageables ont pour point commun de reposer entièrement sur des paramètres non aléatoires. Leur qualité dépend de la pertinence des hypothèses formulées.

Les présents travaux vont produire un estimateur synthétique selon la modélisation suivante : on considère qu'il n'existe pas d'effet propre au domaine dans la réponse que fournit une unité *i* au-delà de ce qui est compris dans l'information auxiliaire. La structure de la population du domaine (sur des informations auxiliaires choisies pertinemment : secteurs d'activité, ancienneté, chiffre d'affaires, nombre d'employés…) serait seule responsable des différences entre domaines. Dans ce cas, on aurait :

$$\hat{Y}_a = \sum_{i \in r} \omega_i^a Y_i$$

avec ω_i^a le poids de l'unité *i* calé sur des marges propres au domaine *a*. Comme la réponse d'une unité n'est pas liée à son domaine, toutes les unités *i* de l'échantillon national participent au calage. Il faut alors procéder à autant de calages qu'il existe de domaines (Ardilly, 2016) (Ardilly, 2015). Les calages sont réalisés sur des marges de type « total ».

2.2.2.3. Estimations composites avec modèle stochastique

Une alternative aux estimateurs descriptifs réside dans l'ajout d'une composante stochastique. On s'intéresse dans la suite à l'estimation d'une moyenne.

Le paramètre à estimer $\acute{Y}_a$ est une variable aléatoire à prédire. L'estimateur de $\acute{Y}_a$ est un estimateur composite d'un estimateur direct et d'un estimateur synthétique issu d'un modèle :

$$\hat{\bar{Y}}_{a,COMP} = \phi_a . \hat{\bar{Y}}_a^D + (1-\phi_a) . \hat{\bar{Y}}_a^{SYNTH} \qquad \text{où } \phi_a \in [0,1]$$

On module ainsi les poids des estimateurs direct (biais faible ou nul mais forte variance) et synthétique (faible variance mais biaisé).

2.2.3. La macro de Statistique Canada

L'institut Statistique Canada a développé des macros SAS qui produisent des estimateurs composites à partir de modèles stochastiques de type modèle linéaire général :

$$Y = X \cdot \beta + \varepsilon$$

$$E(\varepsilon) = 0 \text{ et } V(\varepsilon) = \Sigma$$

avec ε aléatoire (effet propre au domaine), *X* vecteur de variables auxiliaires déterministes connues, β paramètre inconnu et Σ non aléatoire. L'estimation des paramètres β et Σ dans le cadre de ce modèle est réalisée sur l'échantillon complet *r*, ce qui la rend très stable.

Les macros utilisent, pour la prédiction, une stratégie linéaire sans biais optimale (*best linear unbiased predictor* « BLUP). Les deux macros fournissent des aides à l'interprétation sur le modèle et le calcul d'une erreur quadratique moyenne. L'une des macros propose des estimations au niveau de l'individu et est adaptée aux aux variables d'intérêt quantitatives, ou qualitatives avec un grand

nombre de modalités. Les variables de l'enquête TIC étant principalement dichotomiques, c'est la macro de niveau domaine qui est utilisée.

Cette macro réalise un modèle au niveau du domaine (d'après (Fay, Herriot, 1979)). L'information auxiliaire doit être fournie au niveau du domaine et le nombre *m* de domaines doit être suffisamment grand[3].

$$\overline{Y}_a = X_a^t \cdot \beta + b_a \cdot v_a$$

Avec b_a un réel connu et v_a l'effet aléatoire propre au domaine. Le modèle suppose que $\overline{Y}_a$ est quantitative et de nature continue, mais peut être appliqué pour des proportions P_a ou des dénombrements, si la population du domaine est assez grande. La modélisation prend en compte l'erreur d'échantillonnage e_a supposée sans biais et de vraie variance Ψ_a supposée connue (en pratique, estimée) qui représente la variabilité au sein d'un domaine. De plus, les e_a sont supposés deux à deux indépendants. Le modèle est alors :

$$\widehat{\overline{Y}}_a = X_a^t \cdot \beta + b_a \cdot v_a + e_a$$

Ce modèle réunit deux aléas : l'aléa e_a du sondeur et l'aléa v_a de l'économètre. On considère ces deux aléas comme indépendants. L'estimation BLUP n'est sans biais que si on considère conjointement les deux aléas.

La stratégie BLUP fournit un estimateur composite dont le coefficient ϕ_a est petit quand la variabilité entre les domaines σ_v^2 est petite et/ou b_a est petit : l'importance donnée à l'estimateur synthétique est plus grande. De même, si Ψ_a est faible, le coefficient tend vers 1 et l'estimateur direct reprend l'avantage.

En pratique, l'estimation de la variance d'échantillonnage locale est instable. Une option de la macro permet de lisser les $\widehat{\Psi}_a$ pour éviter que les estimations de l'effet local σ_v^2 ne soient nulles.

La macro offre plusieurs options pour estimer σ_v^2 et σ_e^2 :

- la méthode de Fay-Herriot (une sorte de « méthode des moments ») ;
- le maximum de vraisemblance restreint (*Restricted Maximum Likelihood – REML)* qui requiert une normalité des v_a (cette méthode est intéressante lorsque *m* est plutôt petit) ;
- la méthode de Wang-Fuller, pour laquelle l'option de lissage des $\widehat{\Psi}_a$ est impossible ;
- le maximum de densité ajusté (prémunit contre le risque d'avoir un trop fort resserrement de la distribution des estimateurs – *shrinkage* – dû à un risque de σ_v^2 nul).

2.2.4. Conclusion

Les estimations sur petits domaines dépendent de l'information disponible sur les individus, sur les domaines, du plan de sondage, et du type même des paramètres à estimer. Les macros de Statistique Canada permettent de prendre en considération plusieurs cas de figure. Il est également possible de se tourner vers d'autres outils, comme le paquet R *sae* (*small area estimation*) (Molina, Marhuenda, 2015). Ce paquet n'a pas été utilisé dans le cadre des présents travaux.

Un certain nombre d'estimations directes et synthétiques peuvent également être obtenues avec les outils habituels du statisticien sous SAS (macro CALMAR pour les calages d'estimateurs directs et synthétiques, par exemple).

[3]Plus le nombre de domaines augmente, plus on peut se permettre de solliciter un grand nombre de régresseurs.

3. Le champ et les données

3.1. Localiser les unités légales

La localisation de l'information concernant une entreprise ne va pas de soi. Produire des statistiques d'entreprise localisées, comme on cherche à le faire pour TIC, c'est en premier lieu se demander quel sens *local* donner à une information sur une unité légale, en particulier sur son système d'information. Comment refléter la réalité des acteurs économiques en région ? Est-ce possible ? Faut-il attribuer l'information TIC à la région où le nombre d'établissements de l'UL est le plus grand ? Où se situe le plus grand établissement – selon quel critère de « grandeur » ? Ou encore, est-il raisonnable de supposer que l'information qualitative connue sur l'unité légale vaut pour chacun de ses établissements, et ainsi doublonner l'information TIC pour chaque établissement : si cette solution a pour mérite d'aboutir à des n_a grands (réduisant la problématique « petits domaines »), elle pèche par d'autres aspects. L'organisation d'une société est telle que l'information sur les usages des TIC sera forcément différenciée entre, par exemple, l'établissement administratif et l'établissement productif d'une même société.

3.1.1. Le Fare localisé 2014

Le Pôle de services d'action régionale « Études économiques régionales » (PSAR EER) a développé un outil, le Fare localisé, issu du *Fichier approché des résultats d'Esane* (Fare), contenant les informations comptables issues des liasses fiscales mises en cohérence avec les informations provenant de l'enquête Enquête sectorielle annuelle (ESA). Le Fare localisé est disponible sur l'année n-2 (en l'occurrence, 2014).

Il indique pour chaque unité légale connue à la date de référence du fichier sa localisation régionale selon différents critères : le siège ; l'implantation des établissements (nombre d'établissements et effectif pour chaque région) ; l'activité (effectif salarié moyen).

Le cas le plus simple est alors de définir une entreprise comme régionale si tous ses établissements sont dans la région. Elle est alors qualifiée de « monorégionale ». Cela revient à exclure de l'analyse beaucoup d'unités qui ne sont pas exclusivement dans une région. On peut également graduer l'importance de l'activité de l'entreprise présente dans la région, entre les seuils extrêmes que sont les monorégionales et les entreprises implantées dans la région. Trois seuils ont été sélectionnés : si l'activité est à plus de 80 % dans la région (quasimonorégionales), si elle est à plus de 50 % (majoritairement régionales) et enfin si l'activité dans la région d'intérêt est plus importante que dans n'importe quelle autre région (principalement dans la région)[4].

9 961 observations de l'enquête TIC 2016 sont retrouvées dans le Fare localisé 2014[5]. Elles disposent toutes d'une régionalité (« principalement dans la région ») (Figure 1). 7 096 unités sont monorégionales (implantées uniquement dans une région). 8 224 unités sont au moins quasimonorégionales.

3.1.2. Géographie officielle : quelques complications

Le Fare 2014 est localisé sur la géographie en vigueur en 2014. Concrètement, cela veut dire qu'aucune localisation n'est disponible sur les nouvelles régions (en vigueur depuis le 1er janvier 2016). La Figure 1 ne fait que concaténer, pour les nouvelles régions, les informations disponibles sur les anciennes régions dans le Fare 2014, avec la perte d'information que cela implique.

En effet, le critère d'activité est très sensible aux contours des territoires. Ainsi, pour la région Nouvelle Aquitaine, 127 unités de la base de sondage n'étaient pas quasimonorégionales dans l'une

[4]Par construction, ces terminologies sont imbriquées : les entreprises monorégionales sont, *a fortiori*, quasimonorégionales ; les entreprises quasimonorégionales sont aussi majoritairement régionales ; etc.

[5]147 observations de TIC ne sont pas retrouvées dans le Fare localisé 2014 (57 sont des associations). Elles seront exclues de l'analyse.

des trois anciennes régions, donc exclues du champ, mais sont monorégionales dans la Nouvelle Aquitaine, donc intéressant les statistiques régionales. S'il est facile de retrouver les monorégionales, il est plus compliqué, sans les algorithmes du Fare, de recalculer le critère d'activité et donc la quasimonorégionalité. Par souci de clarté des critères d'inclusion dans le champ, les travaux se cantonneront donc à utiliser les critères de régionalité calculés dans le Fare localisé.

Région	Siège	Activité			
		Monorégionales	Quasi monorégionales	Majoritairement régionales	Principalement dans la région
Île-de-France	2 867	1 657	1 956	2 250	2 585
Auvergne-Rhône-Alpes	1 309	981	1 122	1 236	1 334
Grand Est	777	594	696	767	802
Hauts-de-France	706	518	615	673	742
Nouvelle-Aquitaine	660	516	599	647	697
Provence-Alpes-Côte d'Azur	645	505	569	617	669
Occitanie	639	484	561	619	663
Pays de la Loire	600	453	523	584	621
Bretagne	434	330	384	417	447
Normandie	407	309	351	397	429
Bourgogne-Franche-Comté	378	299	344	382	396
Centre-Val de Loire	307	237	283	315	340
La Réunion	94	94	95	95	95
Corse	47	45	46	46	47
Martinique	41	33	36	38	41
Guadeloupe	39	32	34	41	43
Guyane	11	9	10	10	10
Total	**9 961**	**7 096**	**8 224**	**9 134**	**9 961**

Figure 1 – Localisation régionale des unités exploitables de TIC 2016 retrouvées dans le Fare 2014, selon le critère[6]

3.1.3. Conclusion sur le champ et choix des domaines

Les analyses porteront sur les unités quasimonorégionales (QMR) telles que présentes dans le Fare 2014 (ces unités comprennent, par définition, les unités monorégionales). Les domaines d'estimation seront les contours régionaux en vigueur depuis 2016. On admet une légère perte d'information liée au changement des contours régionaux.

Les départements et régions d'outre-mer (DROM) sont exclus de l'analyse. Il nous semble raisonnable de craindre que l'information concernant les DROM ne puisse être modélisée à partir d'une information majoritairement collectée sur la métropole. En effet, leurs faibles $\widetilde{n}_a$ laisse présager que l'estimateur composite sera proche d'un estimateur synthétique. Si le modèle laisse à penser que l'estimateur synthétique est dans l'ensemble bon, il peut ne pas l'être pour ces cas particuliers.

La Corse, dont le $\widetilde{n}_a$ est également faible, demeure et fera l'objet d'une expertise vigilante.

Au final, la base de sondage restreinte aux unités métropolitaines QMR comporte 170 828 unités. La somme des poids[7] de l'échantillon métropolitain QMR de TIC (répondantes et non-répondantes) vaut

[6]Mayotte ne figure pas dans le tableau : bien que le DROM soit dans le champ théorique, dans la base de sondage, seulement deux unités mahoraises étaient présentes, aucune n'ayant été tirée dans l'échantillon.
[7]Poids de sondage, qui n'ont pas de raison *a priori* d'être calés sur les critères de régionalité.

170 829,7. C'est une très bonne adéquation initiale, d'autant que même sur la base complète, il n'y a pas égalité entre les deux totaux, à cause du renouvellement par moitié de l'échantillon.

3.2. L'information auxiliaire

Les modèles d'estimation de niveau individu supposent l'existence d'une information auxiliaire individuelle et disponible pour toutes les unités du champ. L'identifiant d'appariement est dans la plupart des cas le SIREN.

Les modèles d'estimation de niveau domaine supposent l'existence d'une information auxiliaire disponible pour tous les domaines. Dans le cadre des présents travaux, elle peut le plus souvent être déduite de l'information individuelle.

3.2.1. Les données de la base de sondage

La base de sondage est une première source de données auxiliaires, disponibles par définition sur l'ensemble des unités. Elle comporte les informations suivantes : chiffre d'affaires et total de bilan ; catégorie d'entreprise ; nombre d'établissements actifs ; strate de sondage ; activité principale exercée ; catégorie juridique ; nombre de personnes occupées ; code commune du siège de l'unité légale.

La plupart de ces variables ont participé au dessin du plan de sondage.

3.2.2. DADS 2014

La déclaration annuelle des données sociales (DADS) est une formalité déclarative que doit accomplir toute entreprise employant des salariés[8].

Dans les DADS 2014, chaque ligne correspond à un employé d'une unité légale. Les agrégats suivants ont été calculés sur les salariés de chaque UL : part ayant de 15 à 24 ans, de 25 à 49 ans, 50 ans ou plus ; part dont la PCS est : inconnue ; artisans, commerçants, chefs d'entreprises ; cadres et professions intellectuelles supérieures ; professions intermédiaires ; employés ; ouvriers ; part d'hommes et part de femmes ; part d'ingénieurs-cadres (PCS commençant par 38) ; part de personnes occupant une profession numérique.

La liste des professions numériques a été fixée par le PSAR EER de l'Insee à l'occasion de la création de l'investissement « Economie du numérique des territoires » (voir Figure annexe 2).

3.2.3. Liasses fiscales 2015

Les liasses fiscales sont un ensemble de déclarations fiscales remises par les professionnels et les sociétés. Deux montants de la comptabilité des sociétés paraissent intéressants : les « autres immobilisations corporelles : matériel de bureau et mobilier informatique, au début de l'exercice » (LB) et « autres augmentations d'immobilisations corporelles : matériel de bureau et mobilier informatique (par acquisitions, créations, apports et virements, de poste à poste) » (LD).

Pour permettre une comparaison de ces deux informations entre les unités de la base de sondage, LB et LD ont été normalisés par le chiffre d'affaire issu de Sirus et exprimés en pourcentages. Quelques règles déterministes ont été mises en œuvre pour limiter les valeurs manquantes.

On récupère également dans les liasses fiscales le total de bilan et le chiffre d'affaires, qui sont plus complets que les informations contenues dans la base de sondage (18 valeurs manquantes contre quelques centaines avant traitements).

3.2.4. Fare 2014

Le Fichier approché des résultats d'Esane contient les informations comptables issues des liasses fiscales mises en cohérence avec les informations provenant de l'enquête Enquête sectorielle

[8]Depuis janvier 2017, les DADS ont été remplacées par les déclarations sociales nominatives (DSN).

annuelle (ESA). On récupère dans ce fichier le taux d'endettement, le taux d'exportations et le taux d'investissement en 2014.

3.2.5. Données de couverture mobile par zone d'emploi

Les données ouvertes de l'Autorité de régulation des communications électroniques et des postes (Arcep) fournissent une information sur la couverture internet mobile des communes de France en avril 2016[9] (date de fin de collecte de l'enquête TIC). Ces données ne sont disponibles que pour la France métropolitaine.

Les taux de couverture par commune reflètent la disponibilité, à l'extérieur des bâtiments, d'accès à un service, tel que les opérateurs l'affichent sur leurs cartes de couverture. Ces cartes sont le résultat d'une modélisation informatique produite par les opérateurs.

Les données n'étant disponibles que par opérateur, pour chaque commune, on considère que la couverture internet mobile (tous opérateurs confondus) est égale au taux de couverture de l'opérateur le plus performant. Par exemple, pour la couverture 3G et pour une commune *i*, le taux de couverture est estimé par :

$$T_i^{3G} = Max\left(T_i^{3G,Bouygues}, T_i^{3G,Free}, T_i^{3G,Orange}, T_i^{3G,SFR}\right)$$

Cet estimateur est, par nature, un minorant du taux de couverture réel, puisqu'en théorie, les intersections entre les zones couvertes par chaque opérateur prises deux à deux pourraient être nulles et la couverture communale réelle serait alors la somme des couvertures de chaque opérateur. Toutefois, les intersections entre opérateurs sont inconnues.

Enfin, un appariement de la base de couverture mobile avec la base communale[10] permet le calcul d'un taux de couverture internet mobile par zone d'emploi 2010 (ZE). Pour chaque ZE :

$$T_{ZE}^{3G} = \sum_{i \in ZE} \frac{\gamma_i}{\sum_{i \in ZE} \gamma_i} T_i^{3G}$$

avec γ_i la surface de la commune *i*.

Il a semblé en effet plus pertinent d'aborder la question de la couverture mobile sur un territoire plus vaste que la seule commune, puisque la connexion internet mobile concerne des employés qui, a priori, se déplacent sur le territoire autour de leur établissement d'affectation. La zone d'emploi a semblé le territoire le plus approprié. Une zone d'emploi est un espace géographique à l'intérieur duquel la plupart des actifs résident et travaillent, et dans lequel les établissements peuvent trouver l'essentiel de la main d'œuvre nécessaire pour occuper les emplois offerts. Des calculs similaires ont été effectués pour obtenir les taux de couverture en 4G et tous types de réseaux sur les ZE par chacun des quatre opérateurs, ainsi que le meilleur taux de couverture 4G et le meilleur taux de couverture, tous types de réseau.

On affecte à chaque unité le taux de couverture de la zone d'emploi dans laquelle est implanté l'établissement de plus haut effectif.

3.2.6. Données de couverture THD domiciliée

L'Observatoire France Très Haut Débit (piloté par l'Agence du Numérique, rattachée au ministère de l'Économie) produit des données sous licence ouverte sur les connexions internet fixes en très haut débit (THD) en France, au niveau de la commune[11]. Ces données sont disponibles pour la France

[9] https://www.arcep.fr/index.php?id=13272

[10] Fichier Insee récapitulant pour chaque commune tous les zonages administratifs et d'étude auxquels elle appartient, ainsi que sa surface et sa population issue du recensement de la population à la date de référence.

[11] http://www.francethd.fr/l-observatoire/l-observatoire-france-tres-haut-debit.html

entière[12]. Quelques ajustements ont été nécessaires pour s'accorder avec la géographie officielle 2016. Nous nous intéressons aux données du premier trimestre 2016.

Les statistiques sont calculées en fonction des débits atteignables à partir des réseaux de communications électroniques filaires. Les statistiques de débit correspondent à une valeur théorique, c'est-à-dire qu'elles correspondent au débit maximal descendant que la ligne peut effectivement atteindre. Les données sont disponibles pour chaque type de technologie (DSL sur cuivre, câble coaxial et fibre optique) et pour l'ensemble des technologies. Elles sont exprimées en part des locaux éligibles à la technologie décrite et selon les seuils suivants : éligibles au THD, 3Mbits/s et plus, 8Mbits/s et plus, 30Mbits/s et plus, 100Mbits/s et plus.

Une classification rapide des communes selon les caractéristiques de leur réseau THD fixe résume également l'information en 17 classes.

On retient pour la base finale les indicateurs localisés à la commune du siège. L'hypothèse est que le niveau de connectivité du siège peut avoir une influence non négligeable sur l'organisation du système d'information de l'unité, en particulier lorsqu'il s'agit de recourir à des outils d'accès à distance (mails, *cloud computing*, etc.) ou de commercer en ligne.

3.2.7. Lifi 2014

La base Lifi (Liaisons financières) permet d'identifier les groupes de sociétés opérant en France et de déterminer leur contour. Pour chaque unité, Lifi permet de connaître si c'est une tête de groupe, une unité du noyau dur d'un groupe ou une unité du contour élargi d'un groupe.

Par convention, une unité qui n'est pas retrouvée dans le contour d'un groupe est considérée comme n'appartenant à aucun groupe.

3.2.8. Conclusions

Les modèles de niveau individu supposent une information auxiliaire exhaustive pour chaque unité. Par construction, les données de couverture internet fixe et mobile sont complètes pour la France métropolitaine. L'indisponibilité des données de couverture internet mobile sur les DROM confirme le choix de ne pas travailler sur ces territoires. Les données du Fare sont également exhaustives pour la base QMR, puisque les régionalités sont calculées sur le Fare. Les données issues de Lifi sont également exhaustives par construction. 18 observations issues des liasses fiscales manquent.

Enfin, 1 185 observations ne sont pas retrouvées dans les DADS (sur la base de sondage QMR métropolitaine qui fait 170 828 unités). Les conséquences de ces trous dans la base dépendront de l'utilité de ces variables dans la suite des travaux. Il est d'ores et déjà exclu de procéder à une imputation, qui ajouterait de l'aléa au modèle.

Pour conclure ce chapitre, notons que l'appariement avec des données non-Insee se révèle très coûteux, faute de référentiels géographiques cohérents — et dans un cas où le code géographique servait de clef d'appariement. L'appariement avec des données Insee suppose un décalage assumé de millésime entre l'information d'intérêt et l'information auxiliaire (souvent plus ancienne).

[12]Y compris les cinq DROM, Saint-Pierre-et-Miquelon, Saint-Barthélémy et Saint-Martin.

4. Estimations régionales directes

4.1. Estimations directes

De premières estimations directes sont réalisées avec les poids redressés de la base TIC 2016 (Figure 2), sur le sous-champ propre aux analyses sur petits domaines (métropolitain, unités QMR dans le Fare 2014). Les poids redressés tiennent compte d'une part d'une correction de la non-réponse totale par repondération au sein de groupes de réponse homogènes (GRH), d'autre part d'un calage sur des marges nationales calculées sur la base de sondage, sur les variables suivantes.

Variable	Ile de France	Centre - Val de Loire	Bourgogne - Franche Comté	Normandie	Hauts de France	Grand Est	Pays de la Loire	Bretagne	Nouvelle Aquitaine	Occitanie	Auvergne - Rhône Alpes	Provence Alpes Côte d'Azur	Corse	**Ensemble**
B1	20,3	11,1	14,7	8,2	13,8	12,5	11,8	11,6	10,4	13,8	13,7	13,0	11,7	14,3
B3	11,7	5,0	4,7	1,7	5,3	5,9	4,9	6,8	3,6	6,5	6,1	6,3	4,6	6,8
C2	94,4	95,1	92,4	96,6	92,5	92,7	93,4	94,0	92,6	93,4	91,0	92,9	93,1	93,3
C4	63,6	62,9	59,5	60,9	57,0	62,3	63,7	66,9	60,4	57,5	57,8	61,1	51,0	61,0
C7	62,1	50,2	61,2	60,0	58,1	57,9	57,7	62,9	56,4	52,7	56,0	56,8	44,4	58,1
C8	68,1	67,7	71,5	68,1	62,8	73,5	67,5	74,2	67,0	67,6	71,3	65,3	41,3	68,4
C9b	20,3	14,4	16,6	14,7	17,9	14,3	16,0	20,4	18,6	18,3	15,6	21,2	23,2	17,9
C10	20,5	14,8	15,6	16,9	14,6	14,2	15,2	17,8	15,7	15,7	15,7	16,9	23,0	16,8
C12	43,0	31,9	32,7	30,1	32,2	33,5	32,4	36,0	35,5	36,0	33,5	33,8	27,2	35,7
C13	52,5	32,9	40,0	35,4	40,0	38,3	37,8	36,8	37,6	40,3	40,3	35,4	21,0	41,3
C15	20,0	20,9	21,5	15,1	17,4	13,5	16,8	15,3	20,9	17,9	16,8	16,1	13,8	17,8
D1	22,3	14,3	12,1	11,9	15,1	10,8	16,6	13,0	13,4	17,5	16,1	16,1	5,5	16,3
E1	13,0	11,3	9,7	10,8	12,2	11,0	10,3	11,0	9,8	8,9	8,5	12,1	12,7	11,0
G1	17,4	14,1	16,9	15,7	13,8	13,4	13,2	18,3	14,0	15,9	14,5	17,9	19,5	15,6
G5	5,4	10,1	8,8	5,2	6,4	10,5	9,6	10,1	9,8	6,3	8,6	4,4	2,7	7,4
G7	59,3	45,4	55,7	53,5	47,8	51,0	49,6	54,5	48,7	55,6	53,4	52,8	44,9	53,4
G8	8,6	10,5	6,4	7,8	7,4	8,0	7,5	10,3	8,7	10,5	9,3	8,7	4,3	8,6

Figure 2 - Estimateurs directs, calage historique

4.1.1. Qualité des estimateurs

La qualité des estimateurs directs est obtenue grâce à la macro Everest (Estimation de Variance dans les Enquêtes Redressées à Echantillon Stratifié) développée au Département des méthodes statistiques de l'Insee. Les coefficients de variation obtenus sont renseignés en Figure annexe 3.

Le coefficient de variation de la variable C2 (Connexion fixe à haut débit) oscille selon la région entre 0,7 % (Île-de-France) et 1,9 % (Bourgogne-Franche-Comté), exception faite de la Corse (4,3 %). Ces coefficients sont satisfaisants et suggèrent de ne pas chercher à améliorer cet estimateur.

La variable C4 (Connexion à internet mobile) donne une relative satisfaction avec l'estimateur direct : le coefficient de variation oscille entre 2,2 % (Île-de-France) et 5,3 % (Centre-Val de Loire), exception faire de la Corse (15,3 %). De même les variables C7 et C8 donnent-elles dans l'ensemble satisfaction.

La variable G7 (achats sur le web) a aussi des coefficients de variation raisonnablement bas pour un certain nombre de régions.

Un schéma se répète de façon régulière : le coefficient de variation de la Corse dépasse, de loin, le coefficient des autres régions. Ainsi, pour 12 des 17 variables, le coefficient de la Corse est supérieur ou égal à 24,5 %. Il atteint plus de 40 % pour cinq d'entre elles. Les trois coefficients suivants sont presque toujours dans l'une des trois régions de plus bas effectifs après la Corse : Normandie, Bourgogne-Franche-Comté, Centre-Val de Loire. À l'inverse, celui de l'Île-de-France est la plupart du temps le plus bas. Ce résultat ne peut pas nous surprendre : il découle naturellement du nombre d'unités dans chaque domaine.

4.2. Régressions préalables à un calage

Dans la perspective d'un calage, la sélection de l'information la plus explicative est compliquée par le fait que l'on souhaite travailler sur un grand nombre de variables d'intérêt qui n'ont pas nécessairement les mêmes liens avec l'information auxiliaire.

Nous recherchons ici les informations auxiliaires qui sont des régresseurs communs au plus grand nombre de variables d'intérêt. Pour chaque champ et jeu de régresseurs potentiels, 17 régressions logistiques sont réalisées, sur 17 variables d'intérêt. On modélise la probabilité que la variable vaille « Oui ». Différents champs sont envisagés, selon la disponibilité de l'information auxiliaire.

4.2.1. Régressions sur l'ensemble du champ

Aucun régresseur issu des DADS ne participe à ce modèle. 22 régresseurs issus des autres sources participent au modèle. Les régresseurs significatifs au seuil 5 % sont récapitulés en Figure annexe 4.

Les variables issues de la base de sondage et de Lifi sont majoritairement significatives : catégorie d'entreprise, nombre de personnes occupées, strate d'activité, appartenance au contour d'un groupe. Notons ici que la variable de secteur d'activité sur huit positions (sect_act_8) n'est pas significative pour la raison qu'elle est redondante : ses modalités sont des regroupements des modalités de la strate d'activité (strate_a), qui est, elle, systématiquement significative. Toutefois, le trop grand nombre de modalités de strate_a empêchera son utilisation dans les calages sur marges. La suppression de strate_a dans le modèle aboutit à sect_act_8 quasiment toujours significatif.

La variable de nombre d'établissements est significative pour la moitié des variables, en particulier celles relatives aux usages d'internet.

L'accès à une connexion internet fixe à plus de 100 Mbits/s est également significatif pour une majorité de variables d'intérêt ; dans l'ensemble, la couverture en très haut débit fournit au moins un régresseur significatif pour 14 variables d'intérêt, alors que la couverture en internet mobile est moins souvent significative (mais elle l'est pour des variables pour lesquelles cela a du sens, comme l'accès à distance ou le recours à des services de *cloud* payant, qui incluent des accès à distance au système d'information de l'entreprise).

Le taux d'exportations en 2014 est significatif pour plus de la moitié des variables d'intérêt. Les données issues des liasses fiscales sont rarement explicatives.

4.2.2. Régressions sur le champ tronqué des non-réponses aux DADS

Le même exercice est réalisé sur le champ amputé des unités non retrouvées dans les DADS, afin de tester le lien entre les variables d'intérêt et les informations tirées des DADS.

4.2.2.1. Avec les variables DADS

Les variables issues des DADS sont les variables de proportions des salariés de l'entreprise selon trois classes d'âge, selon le sexe, selon la PCS ; sont également connues la proportion d'ingénieurs et la proportions de salariés des professions numériques.

Dans ces modèles, les variables issues de la base de sondage et de Lifi sont majoritairement significatives, tout à fait comme dans les régressions portant sur l'ensemble du champ, avec de menues variations. Au contraire, dans ce modèle avec des régresseurs issus des DADS, les données de couverture THD cessent d'être significatives pour un grand nombre de variables d'intérêt (l'accès à un THD à 100 Mbits/s n'est plus significatif que pour trois variables d'intérêt). Comme dans les régressions précédentes, les données des liasses fiscales sont rarement utiles au modèle.

Toutes les variables issues des DADS sont explicatives pour au moins une variable d'intérêt. Celles qui se détachent toutefois sont la répartition par âge de la population (en particulier, la part de salariés de plus de 50 ans) et la répartition par PCS de la population. La proportion de professions du numérique est rarement significative, mais elle est notablement significative pour les variables « Emploi de personnel spécialisé dans le domaine des TIC » et « Recrutement de personnel pour des postes dans le domaine des TIC » et pour la gestion du commerce sur internet (achats et ventes).

4.2.2.2. Sans les variables DADS

A titre de comparaison entre les deux champs (avec et sans les unités connues dans les DADS 2014), des régressions sont faites sur le champ tronqué des non-réponses DADS, mais avec uniquement les régresseurs connus sur le champ complet (aucune variable issue des DADS).

Les régresseurs significatifs sur le champ tronqué sont les mêmes, variable par variable, que ceux qui étaient significatifs sur le champ complet. On peut donc considérer que la modification du champ ne modifie pas, ou de manière négligeable, les liens entre variables.

4.2.2.3. Comparaison des deux modèles

Sur le champ tronqué, les modélisations incluant des variables issues des DADS sont toujours meilleures que les modélisations n'utilisant pas ces variables, selon le critère d'information d'Akaike et le critère d'information bayésien (Figure 3).

	Sans variables DADS		**Avec variables DADS**	
	AIC	**BIC**	**AIC**	**BIC**
B1_PERS_SPEC	6 399	6 685	6 149	6 477
B3_RECRUT	4 908	5 299	4 689	5 011
C2_FIXE_HAUT_DEBIT	3 012	3 291	2 965	3 259
C4_CONNEX_MOB_3G	8 921	9 215	8 824	9 139
C7	7 953	8 246	7 772	8 094
etc.	etc.	etc.	etc.	etc.

Figure 3 - Qualité des régressions réalisées sur le champ tronqué (sélection de variables)

Il apparaît immédiatement que les DADS représentent une source d'information augmentant la qualité des modèles. Toutefois, N2 unités (1 185) sont inconnues dans les DADS (contre N1=170 643 unités connues).

La diminution du nombre d'unités inhérente à l'utilisation d'un champ tronqué pose toutefois problème dans le cadre d'estimations sur petits domaines. Il semble de plus exclu de restreindre dans l'absolu le champ de l'analyse : on préfère adapter les paramètres au champ que le champ aux paramètres. Le calage sera donc réalisé sur le champ complet, sans utiliser d'information des DADS.

Les liens entre variables ne sont pas modifiés par le passage du champ complet (N1+N2) au champ tronqué (N1). D'autre part, les N2 unités non trouvées dans les DADS n'ont pas de profil particulier en matière de secteur d'activité, taille d'effectif, ancienneté de l'UL ou région de localisation. En postulant une équivalence des champs, on pourrait alors utiliser de l'information auxiliaire DADS sur les domaines (calculée à partir de la base tronquée) dans un modèle de Fay-Herriot.

4.3. Estimations directes avec calage sur marges régionales

4.3.1. Choix des variables de calage

Un calage sur marges est envisagé au moyen de la macro Calmar développée par l'Insee. Chaque calage est réalisé sur les unités d'une seule région, sur des marges régionales, avec *a priori* des variables de calage identiques pour chaque région. Les variables retenues *a priori* pour les calages sont des variables significatives pour une majorité de variables d'intérêt d'après les régressions réalisées : Secteur d'activité (huit positions) ; Catégorie d'entreprise ; Contour Lifi ; Taux d'exportations en 2014 (discrétisé en cinq modalités selon les bornes : le troisième quartile, le neuvième décile, le 95^{e} percentile et le 99^{e} percentile) ; Eligibilité au THD à 100 Mbits/s (discrétisé en 3 modalités selon les bornes : la médiane, le troisième quartile) ; Nombre de personnes occupées.

4.3.2. Mise en œuvre du calage

Le calage sur marges est réalisé en partant des poids de TIC corrigés de la non-réponse selon des groupes de réponse homogènes tels que calculés lors de la production de l'enquête. Le calage converge pour toutes les régions sauf la Corse, avec la méthode logit ; pour certaines régions (Auvergne-Rhône-Alpes et Grand Est), la convergence est imparfaite avec la méthode logit. La méthode linéaire tronquée converge dans ce cas. La Figure 4 présente les rapports de poids minimum, moyen et maximum obtenus pour chaque région.

Région	Rapport des poids			Méthode
	Min	Moyenne	Max	
Ile-de-France	0,6820	1,0225	1,3871	logit
Centre-Val de Loire	0,1580	1,0447	4,5364	logit
Bourgogne-Franche-Comté	0,5398	0,9705	1,8412	logit
Normandie	0,3925	1,0280	3,0964	logit
Hauts de France	0,5369	1,0713	4,5525	logit
Grand Est	0,1765	0,9857	1,9902	linéaire tronqué
Pays de la Loire	0,2913	0,9906	2,7476	logit
Bretagne	0,1144	1,0604	2,0900	logit
Nouvelle Aquitaine	0,6392	1,0908	2,8413	logit
Occitanie	0,4378	1,0685	2,7067	logit
Auvergne-Rhône-Alpes	0	1,0002	1,7435	linéaire tronqué
Provence-Alpes-Côte d'Azur	0,0084	0,9864	2,1992	logit

Figure 4 - Rapports des poids (poids calé / poids d'origine) par région

Le projet de caler les unités corses sur les marges régionales avec les mêmes variables et modalités de calage que les autres régions est abandonné. Plusieurs calages spécifiques à la Corse ont été tentés, sur un nombre restreint de marges. S'il est possible d'obtenir un calage convergent avec des marges sur les effectifs et sur le secteur d'activité, ce calage est suspect, car il augmente les coefficients de variation[13]. Par conséquent, aucun estimateur direct calé sur des marges locales ne sera communiqué pour la Corse.

4.3.3. Estimations calées et qualité

Les estimations sur les variables participant encore à l'analyse et sur les régions métropolitaines (hors Corse) sont présentées en Figure 5.

[13]Les conditions asymptotiques pour un bon calage ne sont pas remplies par le faible nombre d'unités corses.

La qualité des estimations est produite grâce à la macro Everest, déjà présentée. Dans l'ensemble, le calage a permis de réduire le coefficient de variation des estimations. Toutefois, sur les 208 estimations produites, 7 sont victimes d'une augmentation du coefficient de variation :

- C9b : le coefficient de variation pour les Pays de la Loire passe de 11,0 % à 11,3 %. Pour cette variable, les régions ont en moyenne gagné 1,3 points de pourcentage de coefficient de variation.
- C15 : le coefficient de variation pour les Pays de la Loire passe de 11,4 % à 11,9 % et pour Centre-Val de Loire de 13,4 % à 13,7 %. Pour cette variable, les régions ont en moyenne gagné 0,5 points de pourcentage de coefficient de variation.
- G1 : le coefficient de variation pour les Pays de la Loire passe de 12,0 % à 12,4 %. Pour cette variable, les régions ont en moyenne gagné 0,9 points de pourcentage de coefficient de variation.
- G5 : le coefficient de variation pour la Bourgogne-Franche Comté passe de 17,0 % à 17,6 % et pour le Grand Est de 12,0 % à 12,3 %. Pour cette variable, les régions ont en moyenne gagné 1,3 points de pourcentage de coefficient de variation.
- G7 : le coefficient de variation pour Provence-Alpes-Côte-d'Azur passe de 4,5 % à 4,8 %. Pour cette variable, les régions ont en moyenne gagné 0,3 points de pourcentage de coefficient de variation.

Variable	Ile de France	Centre - Val de Loire	Bourgogne - Franche Comté	Normandie	Hauts de France	Grand Est	Pays de la Loire	Bretagne	Nouvelle Aquitaine	Occitanie	Auvergne - Rhône Alpes	Provence Alpes Côte d'Azur	**Ensemble (hors Corse)**
B1	19,7	11,8	14,2	7,4	14,6	12,1	12,9	11,0	10,4	12,8	14,0	13,6	14,2
B3	11,4	4,8	4,8	1,8	5,4	5,7	5,0	6,2	3,6	6,1	6,2	6,3	6,7
C4	63,6	64,2	59,3	57,6	56,6	61,4	65,5	67,7	60,2	56,6	58,3	60,9	61,0
C7	61,7	53,6	60,7	56,2	57,1	57,0	60,2	64,3	55,7	50,7	57,0	57,0	58,0
C8	67,3	69,1	70,8	65,6	63,6	73,4	68,3	74,6	67,4	66,8	72,0	64,9	68,5
C9b	19,9	14,9	17,4	13,9	17,2	14,6	14,9	20,8	20,2	18,6	15,4	21,0	17,8
C10	20,2	14,6	15,5	15,5	14,9	14,5	15,2	17,7	16,9	15,5	15,7	17,3	16,8
C12	42,4	31,2	32,2	29,0	32,8	33,6	31,4	34,5	36,9	34,8	33,7	33,7	35,5
C13	51,8	36,5	39,1	32,3	41,1	37,6	38,9	37,3	37,9	39,0	40,9	36,3	41,5
C15	19,6	19,9	21,2	13,8	17,1	13,7	16,2	15,2	21,6	18,3	17,1	16,6	17,8
D1	22,0	15,1	12,2	11,2	16,0	11,1	17,3	12,8	13,3	16,9	16,3	16,6	16,4
E1	13,1	10,7	9,3	9,8	12,2	11,3	10,8	11,6	9,8	8,2	8,3	12,4	10,9
G1	17,2	14,3	17,1	15,0	13,5	14,0	12,5	18,6	15,2	16,0	14,3	17,7	15,6
G5	5,4	10,9	8,3	5,2	7,2	10,0	9,8	10,4	9,1	6,8	8,4	4,7	7,5
G7	58,9	46,0	54,8	52,5	48,1	51,1	49,9	54,0	49,7	54,6	53,6	52,2	53,4
G8	8,7	11,5	6,7	8,0	7,8	8,0	7,5	10,7	9,6	11,0	9,2	9,2	8,9

Figure 5 - Estimations directes, calage sur marges régionales

Il est possible que les calages aient mal décrit certaines spécificités locales, ou que des calages avec des contraintes différentes sur les rapports de poids auraient abouti à une amélioration de la qualité.

De plus, les variables choisies comme marges étaient des régresseurs décrivant globalement bien une majorité de variables d'intérêt, mais peuvent être inadaptées pour certaines.

Ainsi, pour la variable C15 (payer pour la publicité), par exemple, le gain est très modeste, quel que soit le domaine. Pour deux régions, le coefficient de variation de l'estimation de C15 augmente. On peut lier ces faibles performances au calage sur marges : cette variable avait pour régresseur significatif le taux d'investissement, non pris en compte par le calage ; elle n'avait pas pour régresseur significatif le taux d'exportation et le nombre de personnes, pris en compte par le calage. Pour cette variable, le calage mis en œuvre est donc peu approprié, voire localement contre-productif. Le même effet se produit avec G5 (ventes EDI), pour qui l'éligibilité au THD et le taux d'exportations n'étaient pas significatifs, alors que le taux d'investissement l'était. Deux des treize coefficients de variation augmentent au lieu de diminuer (en moyenne sur toutes les régions, le calage représente tout de même un gain de précision).

Au contraire, certaines variables de mauvaise qualité avec l'estimation utilisant le calage historique sont bien plus précises après le calage. C'est par exemple le cas des variables relatives au personnel, B1 et B3.

En l'état des estimations, seules les estimations en Figure 6 ont encore un coefficient de variation au-dessus du seuil-cible de 17,5 %.

En revanche, les variables C4, C7, C8, C12, C13 et G7 ont toutes des coefficients de variation inférieurs à 10 %.

Pour ces variables, il est raisonnable de considérer que l'estimation avec calage sur des marges régionales donne de bons résultats. Ces variables étaient déjà de bonne qualité avec l'estimation utilisant le calage historique, mais que le calage sur marges régionales les a améliorées.

		Est. direct calage historique		**Est. direct calage régional**	
Var	Domaine	estimation (%)	coefficient de variation (%)	estimation (%)	coefficient de variation (%)
B3	Centre - Val de Loire	5,0	*25,3*	4,8	*21,9*
B3	Bourgogne - Franche Comté	4,7	*24,2*	4,8	*21,4*
B3	Normandie	1,7	*32,2*	1,8	*20,7*
E1	Centre - Val de Loire	11,3	*18,9*	10,7	*18,6*
E1	Bourgogne - Franche Comté	9,7	*18,8*	9,3	*18,8*
G5	Bourgogne - Franche Comté	8,8	*17,0*	8,3	*17,6*
G5	Provence Alpes Côte d'Azur	4,4	*20,1*	4,7	*19,4*
G8	Bourgogne - Franche Comté	6,4	*23,0*	6,7	*20,8*
G8	Normandie	7,8	*19,9*	8,0	*17,6*

Figure 6 - Estimations à faible précision à l'issue du calage régional

4.3.4. Conclusion

Enfin, les calages ont été effectués avec de faibles contraintes sur les rapports de poids. Ils ne représentent pas un calage optimum et pourraient être améliorés. L'absence de calage satisfaisant pour la Corse contraint à renoncer à un estimateur direct pour la Corse, et par conséquent, à un résultat national de référence sur un champ incluant la Corse. Les gains de précision sur les autres régions et pour toutes les variables sont appréciables. Ces estimateurs directs sont de bons candidats à l'application d'un modèle de Fay-Herriot. Un modèle de Fay-Herriot sur la région a toutefois des risques d'échouer, car *m* petit. Le nombre de variables auxiliaires ne pourra pas être trop important.

5. Estimations régionales avec modèle

Dans ce chapitre, on souhaite améliorer la qualité des estimations en ajoutant une part de modélisation. Deux axes sont proposés : d'une part, la modélisation de Fay-Herriot précédemment décrite, qui crée un estimateur composite entre un estimateur direct et un estimateur issu d'un modèle stochastique. D'autre part, un estimateur synthétique, issu d'un modèle entièrement descriptif, sans aléa.

5.1. Estimations synthétiques

5.1.1. Mise en œuvre du calage

La logique du modèle consiste à dire qu'il n'existe aucun effet régional. L'ensemble de la réponse d'une entreprise ne dépend que de ses caractéristiques : en calant l'échantillon national sur ces caractéristiques dans chaque région, chaque estimation régionale se nourrit de l'ensemble des réponses à l'enquête.

On se restreint toutefois aux répondantes appartenant au champ des QMR métropolitaines. La possibilité d'une spécificité ultramarine a déjà été posée ; de même, les entreprises qui ne sont pas quasimonorégionales ont sans doute des caractéristiques et un comportement qui leur sont propres.

Toutes les unités *i* de l'échantillon national participent à un calage linéaire sur les marges du domaine *a*. Il faut alors procéder à autant de calages qu'il existe de domaines et obtenir autant de pondérations de l'échantillon national qu'il existe de domaines. On obtient alors des estimations synthétiques *(indirectes)* calées sur des marges régionales.[14]

Le calage sur marges est réalisé sur les marges précédemment calculées pour les estimations *directes* avec calages sur marges régionales.

5.1.2. Estimateurs et qualité du modèle

Les estimateurs synthétiques sont présentés en Figure 7. On les compare avec l'estimateur direct de l'ensemble avec calage historique sur des marges nationales (le calage régional n'est disponible que sur un ensemble amputé de la Corse[15]).

Pour certaines variables, les estimateurs synthétiques issus de ces calages présentent un visage des régions gommé de la plupart des aléas par rapport aux estimateurs directs. C'est le cas notamment des variables C2 (présence d'une connexion fixe à haut débit ; variable presque saturée), C9b (présence d'un panier virtuel sur le site internet de la société), C15 (la société a payé pour diffuser de la publicité sur Internet - Figure 8), E1 (analyses big data), G1 (ventes sur le web), G5 (achats EDI).

Dans le cas des variables C4 (connexion mobile à internet), C7 (utilisation par certains employés d'appareils connectés) et C8 (site web), seule une valeur atypique empêche la tendance linéaire de s'aplatir vraiment : il s'agit de la Corse. En effet, même si, avec l'estimateur synthétique, la spécificité locale observée avec l'estimateur direct (bas et de mauvaise qualité) s'atténue, la Corse demeure moins dotée de ces technologies TIC que les autres régions.

Au contraire, pour les variables C9b (présence d'un panier virtuel sur le site internet de la société) et C10 (développement d'un site mobile), la Corse avait une valeur atypiquement haute avec l'estimateur direct, possiblement due à un aléa de l'échantillonnage ; à l'issu du calage, sa valeur est la plus basse de toutes. L'ensemble des nuages de points sont disponibles en Figure annexe 5.

[14]Cette méthode et la justification de l'utilisation d'un calage linéaire ont été présentés par (Ardilly, 2016).

[15]On pourrait procéder à un nouveau calage de la base nationale avec les mêmes variables de calage que celles utilisées pour les calages régionaux pour les estimateurs directs et synthétiques, mais ce serait multiplier les estimateurs et, sans doute, s'y perdre un peu.

Variable	Ile de France	Centre - Val de Loire	Bourgogne - Franche Comté	Normandie	Hauts de France	Grand Est	Pays de la Loire	Bretagne	Nouvelle Aquitaine	Occitanie	Auvergne - Rhône Alpes	Provence Alpes Côte d'Azur	Corse	**Ensemble** (est. direct, calage national)
B1	20,1	11,5	12,3	11,1	12,9	12,7	12,6	12,1	11,4	11,9	13,8	12,8	7,2	14,3
B3	11,1	4,7	5,0	4,5	5,7	5,8	5,2	5,1	4,7	5,3	6,2	5,9	2,9	6,8
C2	93,7	93,1	93,1	93,0	93,2	92,8	93,4	93,2	92,9	92,8	93,2	92,8	92,1	93,3
C4	63,0	60,4	60,6	60,2	61,0	60,3	61,2	60,5	59,6	59,2	61,4	60,1	57,1	61,0
C7	61,4	57,0	57,9	56,4	58,0	57,4	58,3	57,5	55,6	54,8	58,7	55,5	50,7	58,1
C8	71,5	67,2	68,4	66,6	68,1	67,6	68,5	67,7	66,6	66,3	69,0	67,2	61,1	68,4
C9b	20,0	16,1	16,7	16,2	16,8	17,6	16,4	17,1	17,0	17,3	17,3	18,6	15,7	17,9
C10	21,3	14,3	14,5	14,3	15,3	15,3	15,1	15,2	14,9	15,4	16,1	17,1	12,8	16,8
C12	43,3	31,2	32,2	31,2	33,2	34,1	32,3	32,2	32,1	33,1	35,1	36,0	27,7	35,7
C13	48,7	37,9	39,5	37,4	40,1	39,6	40,1	39,1	37,2	37,1	41,8	38,9	29,0	41,3
C15	20,2	16,5	16,6	16,4	16,9	17,3	16,9	17,1	16,8	17,0	17,4	18,1	16,2	17,8
D1	21,5	13,7	14,4	13,5	15,1	15,1	14,7	14,4	13,6	14,0	16,1	15,1	10,1	16,3
E1	12,7	10,0	10,0	10,2	10,6	10,5	10,3	10,4	10,1	10,2	10,7	10,9	9,3	11,0
G1	16,7	14,5	15,0	14,6	15,0	15,4	14,8	15,0	15,2	15,4	15,3	16,3	14,1	15,6
G5	6,2	8,2	8,9	7,8	8,3	8,2	8,2	7,8	7,6	7,0	8,0	6,5	5,4	7,4
G7	57,7	50,9	51,4	50,8	52,0	52,4	51,7	51,2	51,3	52,0	53,2	53,6	48,8	53,4
G8	9,2	8,4	8,5	8,5	8,6	8,4	8,8	9,0	8,7	8,7	8,6	8,7	8,4	8,6

Figure 7 - Estimateurs synthétiques régionaux et estimateur direct de l'ensemble

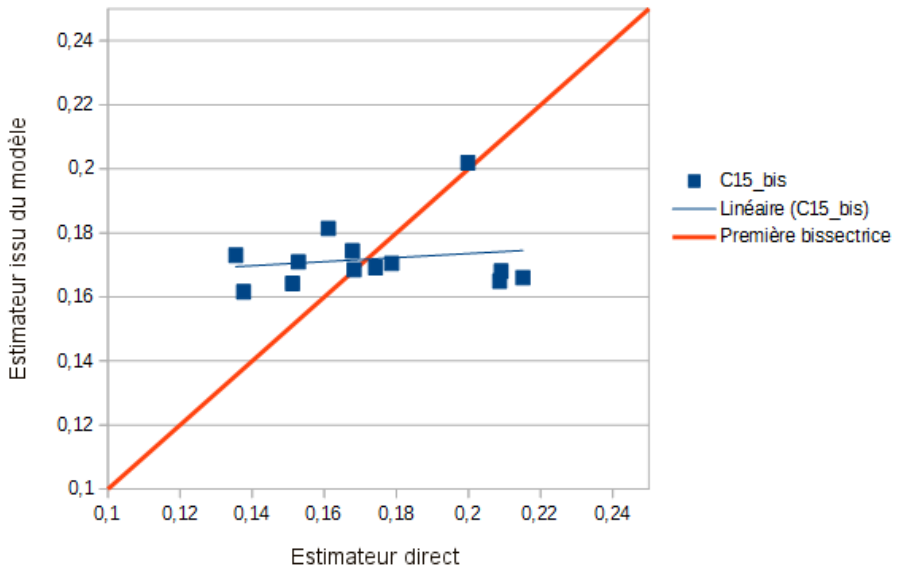


Figure 8 - Comparaison des estimateurs directs et synthétiques et tendance linéaire du nuage (variable C15)

L'utilisation de la macro Everest pour mesurer la précision de ces estimations n'a pas de sens car l'outil ne mesurerait pas la pertinence de la méthode. Il n'y a, de plus, aucune raison d'estimer une

« variance d'échantillonnage » sur des poids calés sur des marges qui n'ont rien en commun avec la base de sondage.

La mise en œuvre de ce modèle a été envisagée dans la perspective d'offrir pour la Corse une meilleure estimation non aléatoire que l'estimation directe avec les poids historiques. Les résultats ont pour mérite d'offrir pour ce territoire une estimation moins atypique que l'estimation directe fondée sur 46 observations. L'estimation demeure, dans sa valeur, tout à fait dépendante du modèle.

5.2. Estimations mixtes, modèle au niveau régional

5.2.1. Préparation des données en entrée des modèles

Un modèle de Fay-Herriot de niveau domaine est réalisé pour l'ensemble des variables. Les estimateurs fournis en entrée sont des estimateurs directs et les variances d'échantillonnage empiriques fournies sont celles calculées par la macro Everest sur ces estimateurs.

Les variables auxiliaires qualitatives sont détaillées en autant de variables qu'elles ont de modalités, chacune valant la part de la modalité dans la région.

Les variables auxiliaires quantitatives sont entrées sous forme de moyenne régionale (part moyenne des entreprises de la région pour la plupart, part moyenne des salariés de la région pour les variables issues des DADS).

En raison du faible nombre de domaines, une sélection des variables du modèle est faite par régressions linéaires : chaque estimation régionale directe (calage régional, sauf pour la Corse) est fonction linéaire des informations auxiliaires agrégées au niveau régional. Pour chaque paramètre, on retient au plus trois régresseurs (dans le cas où un paramètre a davantage de régresseurs significatifs, on retient les trois premiers entrés dans le modèle par la méthode stepwise). Les régressions incluent une constante. Dans certains cas, le modèle avec constante ne donne pas de régresseurs significatifs ou un R^2 trop bas pour les trois premiers régresseurs. Dans ces cas, un modèle sans constante améliore souvent l'ajustement.

Cet exercice a été fait deux fois : d'une part pour les estimations directes régionales avec calage sur marges régionales (la Corse n'étant pas incluse), d'autre part pour ces mêmes estimations auxquelles est ajoutée l'estimation directe pour la Corse fondée sur les poids d'origine de TIC (calage national).

Les modèles de Fay-Herriot sont ensuite mis en œuvre pour chaque variables avec les informations auxiliaires les plus pertinentes.

Par exemple, dans le modèle de Fay-Herriot pour E1 (Corse incluse), les informations auxiliaires sont : la couverture moyenne en 4G, la part moyenne de salariés de plus de 50 ans, le secteur de l'industrie ; il n'y a pas de constante. Le R^2 de la régression linéaire de E1 sur ces paramètres vaut 0,9924 (bon ajustement de la régression).

Les différentes méthodes d'estimation de σ_v^2 et σ_e^2 sont testées : la méthode de Fay-Herriot (FH), le maximum de vraisemblance restreint (REML), le maximum de densité ajusté (ADM) et la méthode de Wang-Fuller (WF). Pour les trois premiers, un lissage de la variance est demandé. De plus, un benchmarking est demandé, pour caler les estimations régionales composites sur les estimations directes nationales connues. Cette étape permet au statisticien de « retomber sur ses pattes » : il sera possible de comparer l'estimation régionale à l'estimation nationale.

5.2.2. Résultats

5.2.2.1. Des résultats appréciables

Ces travaux ont été réalisés pour l'ensemble des variables d'intérêt qualitatives. S'il est impossible de présenter ici tous les résultats de toutes les variables par toutes les méthodes, on peut énoncer quelques généralités et prendre quelques exemples.

Ainsi, à l'issue des modélisations, toutes les régions disposent d'un estimateur composite pour chaque variable : aucun estimateur direct n'était si mauvais qu'il ne puisse contribuer à l'estimateur synthétique – y compris dans les modèles incluant la Corse. Toutefois, selon les variables, certaines méthodes d'estimation de σ_v^2 et σ_e^2 n'ont pas convergé et les estimateurs ne sont pas disponibles. La méthode du maximum de densité ajusté (ADM) a toujours convergé. On donne en Figure 9 les estimateurs obtenus avec cette méthode, pour une sélection de variables.

Variable	Ile de France	Centre - Val de Loire	Bourgogne - Franche Comté	Normandie	Hauts de France	Grand Est	Pays de la Loire	Bretagne	Nouvelle Aquitaine	Occitanie	Auvergne - Rhône Alpes	Provence Alpes Côte d'Azur	Corse
B1	19,8	11,6	13,1	10,1	13,9	12,1	12,6	12,0	11,3	12,6	13,7	13,8	11,9
B3	11,6	4,9	4,7	3,4	4,7	5,4	5,5	5,5	4,4	6,5	6,3	5,9	4,7
C2	93,9	94,5	92,3	94,7	93,6	92,6	94,6	94,7	93,5	93,4	91,5	92,1	91,4
C4	63,8	63,8	61,9	59,6	55,7	60,2	64,7	66,8	60,9	56,9	59,0	59,8	55,0
C7	62,1	55,3	59,5	56,0	56,2	55,7	60,6	63,3	56,2	52,8	57,9	55,7	44,0
C8	67,3	68,3	73,4	64,5	65,5	71,7	70,5	71,8	67,4	67,1	72,1	65,2	44,3
C9b	19,9	16,1	15,8	16,6	16,9	14,8	15,5	19,7	20,6	18,0	15,1	20,6	24,1
C10	20,1	14,6	15,3	15,4	14,9	15,2	15,2	16,4	15,9	16,2	15,8	17,7	18,6
C12	42,2	29,4	32,3	30,7	32,9	33,5	31,3	35,4	36,3	34,7	34,7	35,7	24,3
C13	51,8	36,4	39,2	33,3	39,1	38,9	39,9	37,3	36,9	37,9	40,5	38,1	21,7
C15	19,6	19,2	22,3	13,8	17,1	15,0	16,1	15,7	19,4	17,9	18,1	16,2	15,0
D1	22,2	14,7	11,3	12,2	15,8	11,5	16,7	13,7	14,2	16,4	15,9	14,8	7,4
E1	13,3	10,2	9,6	10,7	11,6	10,7	10,2	10,8	10,1	9,4	8,9	11,6	10,3
G5	5,4	9,6	9,3	6,9	7,7	9,2	9,5	9,4	8,5	6,7	8,5	5,1	3,6
G7	58,9	47,0	54,1	51,5	49,5	49,3	49,9	53,5	52,0	54,5	53,1	53,7	46,6
G8	8,4	9,8	7,9	8,5	8,3	8,3	8,3	9,6	9,0	9,6	8,6	8,5	7,0

Figure 9 - Estimateurs composites SAE, pour les modèles incluant la Corse, méthode ADM

Il arrive que les modèles donnent une part écrasante à l'estimateur synthétique. C'est le cas pour C2 (fixe haut débit) pour les méthodes REML et FH du modèle sans la Corse et les méthodes REML et WF du modèle avec la Corse – la méthode FH n'ayant pas convergé. C'est également le cas pour G8 (achats EDI) pour les méthodes FH et WF – la méthode REML n'ayant pas convergé. La précision des estimateurs composites issus de ces modèles est améliorée par rapport à celle de l'estimateur direct.

Notons ici que la méthode ADM n'aboutit dans aucun cas à un coefficient très bas pour l'estimateur direct. Plus encore, dans la plupart des cas, la méthode ADM donne un poids plus important que les autres méthodes à l'estimateur direct. On le constate par exemple pour la variable B1. Un résumé des résultats du modèle est en Figure 10, d'où seront aussi tirés les exemples ci-après.

Par ailleurs, le coefficient associé à l'estimateur direct est en général plus élevé quand le domaine avait une population répondante plus grande. C'est ainsi que l'estimateur composite de l'Île-de-France accorde une part plus importante à l'estimateur direct (entre 0,74 et 0,81 pour B1, selon la méthode) que ne le fait l'estimateur composite de la Bourgogne-Franche Comté (entre 0,32 et 0,46 pour B1, selon la méthode). Dans les modèles incluant la Corse, le coefficient de l'estimateur direct est le plus bas de toutes les régions, au profit de l'estimateur synthétique (entre 0,08 et 0,11 pour B1, selon la méthode).

Dans l'ensemble, les estimations concernant la Corse ont beaucoup gagné en précision grâce aux modèles (de 41,7 % pour l'estimateur direct de B1 à 14,3 % pour le plus bas des coefficients de variation obtenus – avec la méthode REML). Les estimations concernant la Corse demeurent, pour les variables G1, G5, C15 (pour la méthode ADM), et D1, un peu au-dessus de la cible de précision (17,5 %).

Région		Ile de France	Centre - Val de Loire	Bourgogne - Franche Comté	Normandie	Hauts de France	Grand Est	Pays de la Loire	Bretagne	Nouvelle Aquitaine	Occitanie	Auvergne - Rhône Alpes	Provence Alpes Côte d'Azur	Corse
Estim. Direct (en %)		20	12	14	7	15	12	13	11	10	13	14	14	12
CV échantillonnage (en %)		4,3	13,8	13,0	15,0	8,6	9,3	9,5	9,6	9,9	8,9	6,6	9,9	41,7
Coefficient de l'estimateur direct	**FH**	0,76	0,34	0,38	0,39	0,52	0,54	0,48	0,41	0,51	0,49	0,65	0,50	0,08
	ADM	0,81	0,40	0,45	0,45	0,58	0,61	0,54	0,47	0,58	0,56	0,71	0,56	0,11
	REML	0,74	0,32	0,36	0,36	0,49	0,52	0,46	0,38	0,49	0,47	0,63	0,47	0,08
	WF	0,76	0,46	0,40	0,64	0,59	0,64	0,60	0,67	0,68	0,63	0,73	0,55	0,09
Estimation composite (en %)	**FH**	20	12	13	10	14	12	13	12	11	13	14	14	12
	ADM	20	12	13	10	14	12	13	12	11	13	14	14	12
	REML	20	12	13	11	14	12	13	12	11	13	14	14	12
	WF	20	12	13	9	14	12	13	12	11	13	14	14	12
CV de l'estimation composite (en %)	**FH**	3,7	10,3	9,6	11,0	7,8	8,4	8,5	9,5	9,0	8,5	6,3	7,7	14,6
	ADM	3,6	10,4	9,5	11,4	7,6	8,2	8,4	9,6	9,0	8,3	6,1	7,6	15,3
	REML	3,7	10,0	9,5	10,7	7,7	8,3	8,3	9,3	8,8	8,3	6,3	7,5	14,3
	WF	4,4	10,9	11,0	11,0	8,1	8,4	8,4	8,4	8,5	8,1	6,3	8,3	17,4

Figure 10 - B1 (emploi de personnel spécialisé en TIC) : résultats du modèle de Fay-Herriot

5.2.2.2. Quelques déceptions

Malgré le soin apporté à l'ajustement de l'information auxiliaire pour chaque modèle, certains modèles donnent des résultats décevants, en ce qu'ils accordent une part écrasante à l'estimateur direct ; l'estimateur composite ne gagne pas en précision (il en perd même localement). Cette circonstance se présente pour les modèles de G1 et le modèle de C8 sans la Corse. Le nombre de domaines n'était peut-être pas suffisant pour appliquer la méthode de Fay-Herriot (d'autant meilleure que *m* est grand).

D'autre part, pour certains modèles, la précision n'est pas améliorée autant qu'on l'aurait souhaitée. Par exemple, l'estimation directe de B3 (recrutement de personnel spécialiste des TIC) était imprécise (CV > 17,5 %) pour quatre régions. Les estimations composites ont toutes gagné en précision, mais :

- Avec la méthode ADM, la Normandie continue d'être au-dessus du seuil-cible (CV=20,3 %) ;
- Avec la méthode WF, deux régions continuent d'être au-dessus du seuil-cible. Il s'agit de la Normandie et la Corse, avec des coefficients de variation respectifs de 18,2 % et 18,1 %.

De plus, pour le modèle sur B3 sans la Corse, certains des plus gros coefficients de variation ont augmenté entre l'estimateur direct et l'estimateur composite (celui de la Normandie avec la méthode ADM, celui de la Bourgogne-Franche Comté avec la méthode WF – les autres méthodes

n'ayant pas convergé). Dans ces conditions, il serait préférable de conserver l'estimation directe calée plutôt que l'estimation composite.

Enfin, le coefficient de variation de l'Île-de-France est régulièrement augmenté par le modèle par rapport à l'estimateur direct ; ce coefficient reste toutefois raisonnablement bas.

5.3. Conclusion

Il est malaisé de résumer l'application de ces différentes méthodes d'estimations sur petits domaines pour l'ensemble des variables d'intérêt. L'étude d'un cas permet cependant de rappeler les différentes méthodes mises en œuvre, leurs résultats et leurs limites.

La variable E1 « l'entreprise a procédé à des analyses de données massives en 2015 » fait partie des variables qui suscitent un grand intérêt chez les acteurs publics. Le service d'étude et diffusion des Pays de la Loire a manifesté son intérêt pour cette variable.

C'est une variable que l'on pouvait supposer *a priori* liée à des informations auxiliaires disponibles (le secteur d'activité, la présence dans l'entreprise de professionnelles du numérique, par exemple) et qui avait le défaut de coefficients de variation parfois élevés avec l'estimateur direct issu du calage historique.

Au niveau individu, les variables explicatives de E1 (régression logistique) sans les informations DADS sont : catégorie d'entreprise, personnes occupées, strate d'activité, position dans le contour d'un groupe, éligibilité au THD à 100Mbits/s, total de bilan. À l'exception du total de bilan, toutes ces variables ont été utilisées dans le calage, qui utilisait, en sus, le taux d'exportations en 2014. Il est raisonnable de penser que le calage devrait donner de bons résultats pour E1. Avec les variables issues des DADS, les variables explicatives sont : catégorie d'entreprise, personnes occupées, strate d'activité, part des plus de 50 ans (effet négatif), part des cadres et des professions intermédiaires (effet positif).

Au niveau régional, les bons régresseurs retenus pour les modèles de Fay-Herriot étaient : la couverture 4G, le secteur des transports, l'appartenance au contour élargi d'un groupe, une constante pour le modèle sans la Corse (R^2 = 0,7526) ; la couverture 4G, la part de salariés de 50 ans ou plus, le secteur de l'industrie pour le modèle avec la Corse (R^2=0,9924).

Pour ces deux modèles, seule la méthode ADM a convergé. Les coefficients sont tels que plus de la moitié de l'estimation finale est issue de l'estimateur synthétique. Pour une majorité de régions, le poids de l'estimateur synthétique dépasse même les trois quarts.

La Figure 11 présente l'ensemble des estimations régionales d'E1 produites dans les présents travaux. La Figure 12 présente l'ensemble des coefficients de variation associés à ces estimations. Ces tableaux appellent quelques observations :

Tout d'abord, le calage sur des marges régionales améliore la qualité des estimations. Il a le défaut de ne pas être disponible pour la Corse. Malgré des variables de calage judicieusement choisies, le calage ne permet toutefois pas d'atteindre un niveau de qualité suffisant pour certaines des autres régions.

Ensuite, l'estimateur synthétique, qui suppose que le comportement des unités de l'échantillon national ne dépend pas des régions dans lesquelles elles se situent, gomme la plupart des différences entre régions. Seules deux régions ressortent : l'Île-de-France, dont le résultat est presque inchangé par rapport à l'estimateur historique, et la Corse, dont le résultat est totalement changé (la région passe de seconde derrière l'Île-de-France à lanterne rouge des régions). Ce résultat semble plus conforme à l' « intuition » et dessine un paysage cohérent avec les estimateurs synthétiques obtenus pour les autres variables. La qualité de ces estimations n'est cependant pas mesurée.

Par ailleurs, les estimations composites produites à partir des estimateurs directs historiques et calés sur marges régionales appellent deux commentaires : d'une part, la qualité des estimations est

fortement améliorée par rapport aux estimateurs directs, quel que soit le modèle. D'autre part, la hiérarchie des régions est globalement conservée entre les estimateurs directs et les estimateurs composites, au contraire de l'estimateur synthétique et à l'exception de la Corse.

Région	Estimateur direct historique	Estimateur direct, calage régional	Estimateur synthétique	Estimateur composite FH (avec Corse)	Estimateur composite FH (sans Corse)
Ile de France	13,0%	13,1%	12,7%	13,3%	13,2%
Centre - Val de Loire	11,3%	10,7%	10,0%	10,2%	10,3%
Bourgogne - Franche Comté	9,7%	9,3%	10,0%	9,6%	9,3%
Normandie	10,8%	9,8%	10,2%	10,7%	11,7%
Hauts de France	12,2%	12,2%	10,6%	11,6%	11,7%
Grand Est	11,0%	11,3%	10,5%	10,7%	11,2%
Pays de la Loire	10,3%	10,8%	10,3%	10,2%	10,4%
Bretagne	11,0%	11,6%	10,4%	10,8%	11,4%
Nouvelle Aquitaine	9,8%	9,8%	10,1%	10,1%	9,5%
Occitanie	8,9%	8,2%	10,2%	9,4%	8,4%
Auvergne - Rhône Alpes	8,5%	8,3%	10,7%	8,9%	9,0%
Provence Alpes Côte d'Azur	12,1%	12,4%	10,9%	11,6%	11,8%
Corse	12,7%	n.d.	9,3%	10,3%	n.d.

Figure 11 - Analyses de données massives en 2015 : toutes les estimations

Région	CV Estimateur direct historique	CV Estimateur direct, calage régional	CV Estimateur synthétique	CV Estimateur composite FH (avec Corse)	CV Estimateur composite FH (sans Corse)
Ile de France	7,1%	6,7%	n.d.	6,2%	6,7%
Centre - Val de Loire	18,9%	18,6%	n.d.	6,6%	6,2%
Bourgogne - Franche Comté	18,8%	18,8%	n.d.	9,5%	9,3%
Normandie	16,6%	15,8%	n.d.	5,5%	8,3%
Hauts de France	12,8%	11,4%	n.d.	7,5%	9,7%
Grand Est	12,1%	11,8%	n.d.	7,1%	7,5%
Pays de la Loire	14,2%	13,7%	n.d.	8,6%	7,1%
Bretagne	16,4%	13,4%	n.d.	6,0%	4,9%
Nouvelle Aquitaine	13,6%	11,9%	n.d.	7,3%	7,4%
Occitanie	14,8%	14,5%	n.d.	7,2%	10,3%
Auvergne - Rhône Alpes	10,6%	10,4%	n.d.	8,5%	8,9%
Provence Alpes Côte d'Azur	12,5%	12,4%	n.d.	7,5%	6,9%
Corse	38,8%	n.d.	n.d.	14,8%	n.d.

Figure 12 - Analyses de données massives en 2015 : coefficients de variation des estimations

La tentation est forte de choisir, pour chaque région, de diffuser l'estimateur dont le coefficient de variation est le plus faible. Cela exclut, d'emblée, les estimateurs synthétiques.

Notons, enfin, qu'il est préférable, comme pour les normes de diffusion de TIC au niveau national, de diffuser les résultats arrondis à l'unité, pour éviter de faire croire à une précision qui n'existe pas.

6. Conclusion

Les estimations directes ont le mérite d'être (presque) partout disponibles, mais pas toujours aussi précises qu'on le souhaiterait. À l'inverse, les estimations issues de modèles de Fay-Herriot apportent un gain conséquent de précision, à la condition de converger. L'estimateur synthétique, enfin, corrige certaines aberrations selon l'intuition à dire d'expert, mais sa qualité finale ne peut être jaugée.

Certaines estimations localisées présentent une qualité que l'on pourra juger individuellement suffisante, mais les estimations ne permettent pas de réaliser des comparaisons entre les régions de France métropolitaine.

6.1. Rapport coût-bénéfices

L'apport de données auxiliaires extérieures à la base de sondage est important, notamment pour la réalisation d'un calage sur marges adapté aux variables d'intérêt et pour l'alimentation d'un modèle de Fay-Herriot. Le temps consacré à l'enrichissement de la base inciterait le statisticien à rentabiliser ses efforts en réalisant les estimations de nombreuses variables. Toutefois, l'estimation sur petits domaines ne peut pas être industrialisée et devrait même, idéalement, être gérée variable par variable plutôt qu'en groupant les variables comme ce fut le cas pour le calage sur marges, dans les présents travaux.

Le coût doit également être mis en regard de la qualité : il peut être frustrant d'obtenir des résultats dont on a des raisons de penser qu'ils sont, dans l'ensemble, meilleurs que l'estimateur direct historique, sans pour autant être certains qu'ils sont meilleurs pour chacun des domaines considérés : les estimateurs sur petits domaines issus de méthodes et d'hypothèses *ad hoc* peuvent être, localement, moins justes que l'estimateur direct. Il n'y a aucun moyen de savoir quel domaine est moins juste que tel autre.

6.2. Précautions d'interprétation

Les travaux portent sur un champ restreint, construit d'après les caractéristiques des entreprises et de leurs régions. Le contour des régions 2016 n'a été qu'imparfaitement pris en compte.

Le champ historique de TIC doit par ailleurs être rappelé, en ce qu'il exclut un certain nombre d'unités : les TPE ; les sociétés agricoles et les sociétés financières et d'assurance, alors que certains thèmes pourraient les concerner au premier plan (accès distant, analyses de données massives, etc.) et leur absence pourrait faire défaut à la compréhension d'un territoire.

Le choix d'un calage sur marges sur le seul critère des régresseurs significatifs signifie que le plan de sondage n'est pas entièrement pris en compte ; en particulier, les unités des secteurs ou tranches de chiffre d'affaires tirées exhaustivement ne devraient pas participer au calage.

Enfin, il faut conserver à l'esprit que l'estimation sur petits domaines implique, par essence, une part de croyance. De plus, le seuil choisi pour déterminer qu'une estimation est « suffisamment » précise, inspiré des pratiques d'autres pays, demeure arbitraire. La qualité des estimations finalement produites ne permet pas, la plupart du temps, d'illustrer de différences significatives entre régions.

6.3. Pistes d'amélioration

On l'a dit, les calages sur marges régionales ont été fait *a minima* mais peuvent être améliorés par le choix de bornes plus rapprochées pour les rapports de poids ou par un calage par paramètre d'intérêt. Ces travaux nécessitent du temps mais pourraient améliorer encore la précision des estimations calées sur ces marges.

Cette amélioration appellerait naturellement de nouveaux modèles de Fay-Herriot, donc de nouvelles régressions pour confirmer le choix des informations auxiliaire alimentant chaque modèle : c'est un cycle complet de travaux qui s'ouvrirait.

Bibliographie

[1] Pascal Ardilly, « Panorama des principales méthodes d'estimation sur les petits domaines », *Documents de travail Insee N°M0602*, 2006

[2] Pascal Ardilly, « Estimation régionale de taux de pauvreté par une méthode de calage », *Séminaire de Méthodologie Statistique du département des méthodes statistiques* « *Miscellanées sur le calage »,* Insee, 15 mars 2016.

[3] Pascal Ardilly, « Regional estimates of poverty indicators based on a calibration technique », *Statistical working papers*, Eurostat, 2015

[4] Fay, Herriot, « Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data », *Journal of the American Statistical Association n°74*, 1979

[5] Molina, Marhuenda, « sae: An R Package for Small Area Estimation », *The R Journal Vol. 7/1*, 2015.

Annexes

Figure annexe 1 - Variables d'intérêt issues de TIC 2016

Variable	Description
B1_PERS_SPEC	Emploi de personnel spécialisé dans le domaine des TIC
B3_RECRUT	Recrutement de personnel pour des postes dans le domaine des TIC
C2_FIXE_HAUT_DEBIT	Utilisation d'un connexion fixe haut débit
C4_CONNEX_MOB_3G	Présence de connexion mobile haut débit
C7	Utilisation par certains employés d'appareils portables permettant une connexion mobile à internet pour accéder au système de courrier électronique, aux documents ou aux applications de l'entreprise
C8_SITE_WEB	Présence d'un site web
C9b_COMMANDE_LIGNE	Possibilité de commande en ligne sur le site web de l'entreprise
C10_SMOB	Développement d'un site mobile ou d'une application web pour des appareils portables
C12	Profil utilisateur sur au moins un média social
C13_ACCES_DIST	Accès à distance pour les employés au système de courrier électronique, aux documents ou aux applications de l'entreprise
C15	L'entreprise paye pour diffuser de la publicité ciblée
D1_CLOUD	Achat de services de Cloud computing
E1	Analyses Big Data en 2015
G1_VENTES_WEB	Réception de commandes de biens ou services sur le site web
G5_VENTES_EDI	Réception de commandes de biens ou services via EDI
G7_ACHATS_WEB	Achats de biens et services via un site ou une application web
G8_ACHATS_EDI	Achats électroniques via des messages de type EDI

Champ : Unités légales quasimonorégionales de France métropolitaine, connues dans le Fare 2014

Figure annexe 2 - Professions du numérique (source PSAR EER)

PCS	Libellé
544A	Employés et opérateurs d'exploitation en informatique
478A	Techniciens d'étude et de développement en informatique
478B	Techniciens de production, d'exploitation en informatique

478C	Techniciens d'installation, de maintenance, support et services aux utilisateurs en informatique
478D	Techniciens des télécommunications et de l'informatique des réseaux
388A	Ingénieurs et cadres d'étude, recherche et développement en informatique
388B	Ingénieurs et cadres d'administration, maintenance, support et services aux utilisateurs en informatique
388C	Chefs de projets informatiques, responsables informatiques
388D	Ingénieurs et cadres technico-commerciaux en informatique et télécommunications
388E	Ingénieurs et cadres spécialistes des télécommunications
463A	Techniciens commerciaux et technico-commerciaux, représentants en informatique

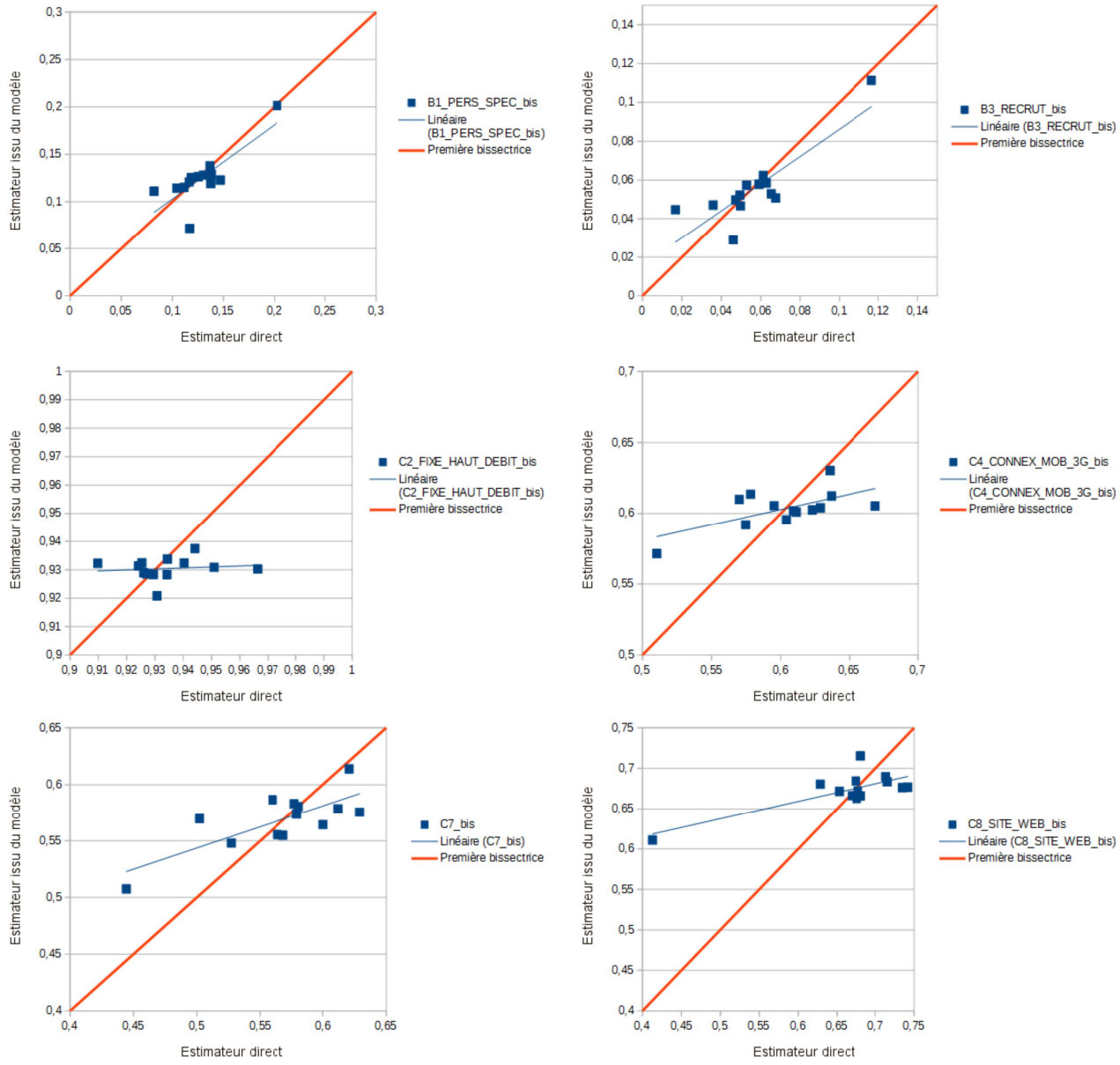
Figure annexe 3 - Coefficients de variation des estimateurs directs avec calage du Pise

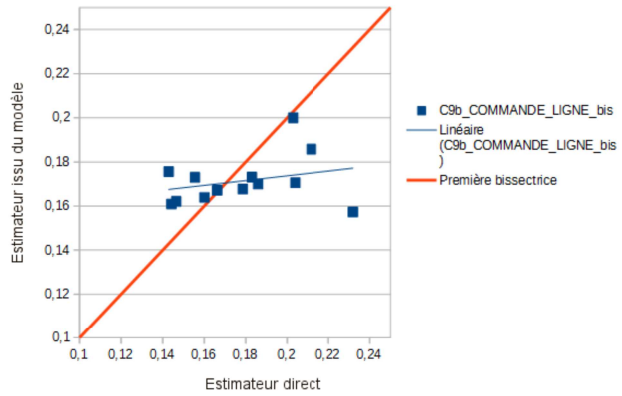
Variable	Ile de France	Centre - Val de Loire	Bourgogne - Franche Comté	Normandie	Hauts de France	Grand Est	Pays de la Loire	Bretagne	Nouvelle Aquitaine	Occitanie	Auvergne - Rhône Alpes	Provence Alpes Côte d'Azur	Corse	**Ensemble**
B1	4,8	16,3	13,7	16,6	10,5	9,8	11,6	12,7	11,5	10,5	7,5	10,9	41,7	2,5
B3	6,6	25,3	24,2	32,2	17,3	15,5	18,5	19,5	18,3	16,3	11,6	16,6	62,8	3,9
C2	0,7	1,5	1,9	1,1	1,4	1,3	1,4	1,6	1,4	1,3	1,1	1,4	4,3	0,4
C4	2,2	5,3	5,1	4,8	4,2	3,5	3,9	4,3	3,8	4,2	2,9	3,8	15,3	1,0
C7	2,2	6,8	5,0	5,0	4,1	3,8	4,3	4,7	4,1	4,6	3,0	4,2	17,2	1,1
C8	2,0	4,8	4,0	4,3	3,8	2,7	3,6	3,7	3,3	3,4	2,2	3,6	18,3	0,9
C9b	5,4	16,4	13,1	13,3	10,0	10,2	11,0	11,3	9,3	9,7	7,4	8,7	27,3	2,5
C10	5,3	16,0	14,1	12,8	11,0	10,2	11,3	12,7	10,5	10,9	7,4	10,1	26,6	2,7
C12	3,2	9,9	8,8	9,1	6,9	6,1	7,1	7,9	6,1	6,3	4,7	6,6	24,5	1,7
C13	2,7	9,4	7,4	8,0	5,8	5,5	6,2	7,6	5,8	5,7	4,0	6,3	27,4	1,4
C15	5,6	13,4	11,9	14,3	10,5	11,1	11,4	13,7	9,0	10,0	7,5	10,6	37,3	2,7
D1	5,0	16,2	15,9	15,7	10,6	11,7	10,5	14,6	11,1	10,0	7,3	10,6	55,2	2,7
E1	7,1	18,9	18,8	16,6	12,8	12,1	14,2	16,4	13,6	14,8	10,6	12,5	38,8	3,6
G1	6,0	16,7	13,3	13,6	11,6	10,6	12,0	12,3	10,5	10,3	7,9	9,5	30,1	2,7
G5	10,5	18,9	17,0	19,5	15,7	12,0	13,7	16,3	12,7	15,6	10,0	20,1	88,6	4,0
G7	2,4	7,5	5,5	5,6	5,0	4,3	5,1	5,5	4,8	4,3	3,2	4,5	17,1	1,2
G8	9,1	19,7	23,0	19,9	15,4	14,0	16,6	17,0	14,7	13,9	10,3	15,5	67,1	4,1

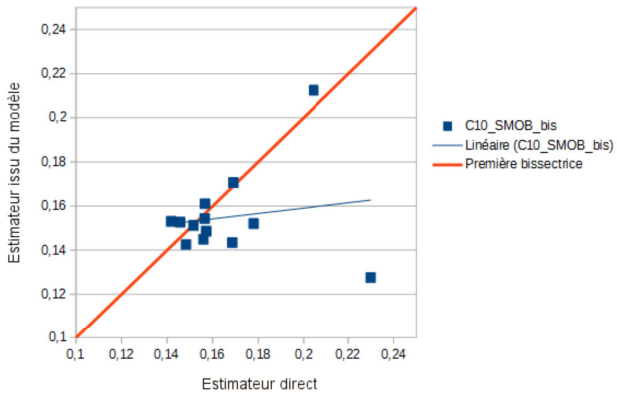
Annexe 4 - Régresseurs significatifs selon la variable d'intérêt, champ complet (une croix signifie que la variable est significative au seuil 5 %)

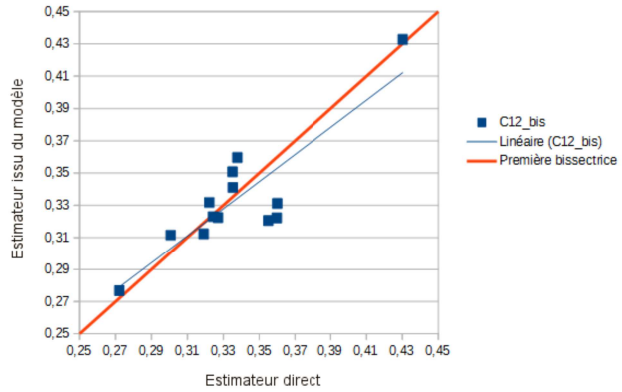
Source		B1_PERS_SPEC	B3_RECRUT	C2_FIXE_HAUT_DEBIT	C4_CONNEX_MOB_3G	C7	C8_SITE_WEB	C9b_COMMANDE_LIGNE	C10_SMOB	C12	C13_ACCES_DIST	C15	E1	D1_CLOUD	G1_VENTES_WEB	G5_VENTES_EDI	G7_ACHATS_WEB	G8_ACHATS_EDI	TOTAL
Sirus – BDS	categorie	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	16
	nbet_a				X	X	X		X	X	X						X	X	8
	sect_act_8																		0
	po	X	X		X	X	X		X	X	X		X	X	X	X		X	13
	strate_a	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	17
Lifi	contour	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	17
Liasses fiscales	lb_norm_imput							X											1
	ld_norm_imput																		0
	ca_lf							X				X							2
	total_bilan_lf	X	X										X						3
Couverture internet	Couv3Gze							X			X				X				3
	Couv4Gze										X			X					2
	Couvmobze																		0
	Eligibles_100M					X		X	X	X	X	X	X	X	X		X		10
	Eligibles_30Mplus										X	X							2
	Eligibles_8Mplus						X								X				2
	Eligibles_3Mplus			X															1
	Eligibles_thd				X														1
	Class_thd		X																1
Fare	taux_endett_14																		0
	taux_export_14	X		X	X	X	X			X	X			X			X		9
	taux_invest_14									X		X				X			3

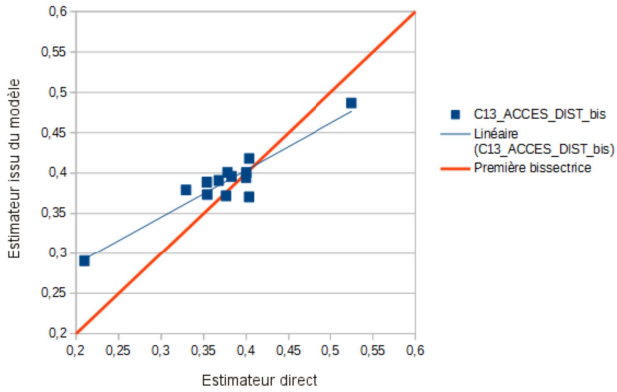
Figure annexe 5 - Comparaison entre les estimateurs directs (calage national) et synthétiques

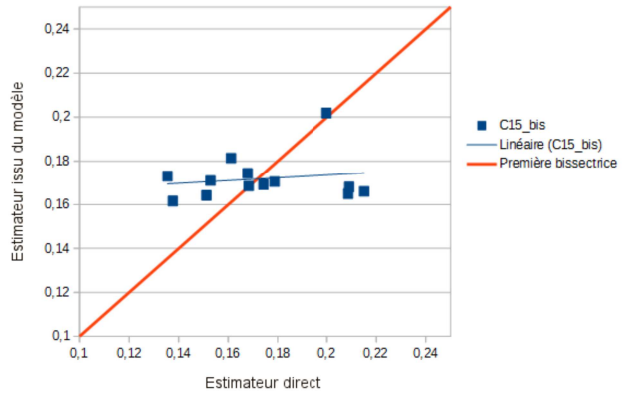


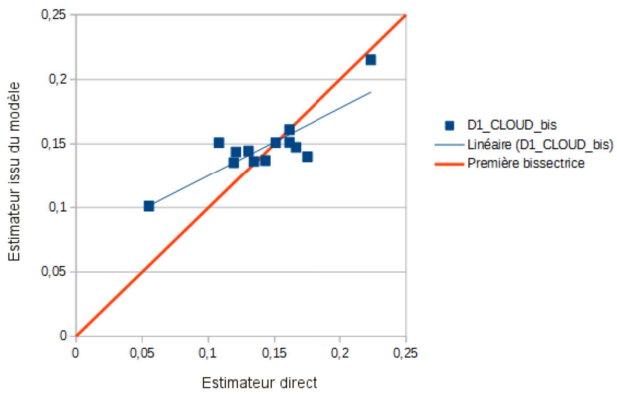

Estimateur issu du modèle
Estimateur direct
C9b_COMMANDE_LIGNE_bis
Linéaire (C9b_COMMANDE_LIGNE_bis)
Première bissectrice

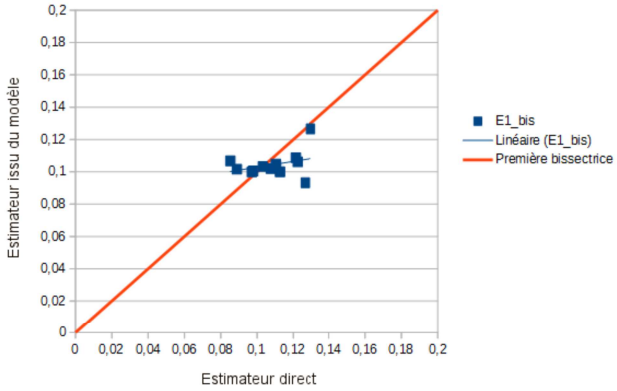

Estimateur issu du modèle
Estimateur direct
C10_SMOB_bis
Linéaire (C10_SMOB_bis)
Première bissectrice

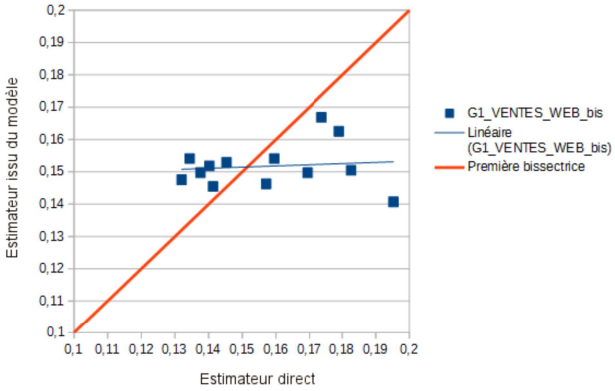

Estimateur issu du modèle
Estimateur direct
C12_bis
Linéaire (C12_bis)
Première bissectrice

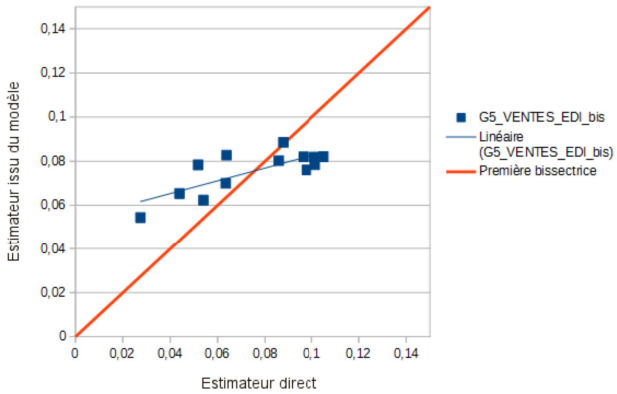

Estimateur issu du modèle
Estimateur direct
C13_ACCES_DIST_bis
Linéaire (C13_ACCES_DIST_bis)
Première bissectrice

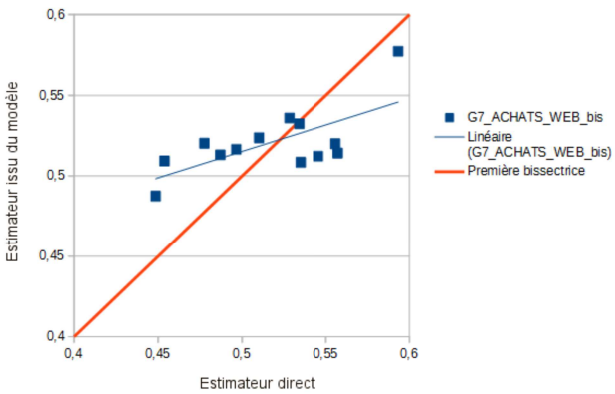

Estimateur issu du modèle
Estimateur direct
C15_bis
Linéaire (C15_bis)
Première bissectrice

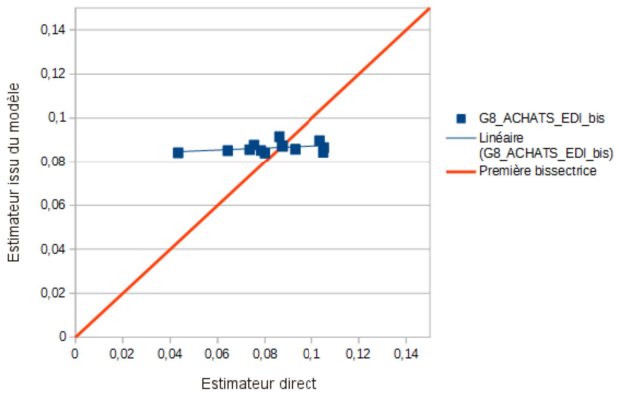

Estimateur issu du modèle
Estimateur direct
D1_CLOUD_bis
Linéaire (D1_CLOUD_bis)
Première bissectrice


Estimateur issu du modèle
Estimateur direct
E1_bis
Linéaire (E1_bis)
Première bissectrice


Estimateur issu du modèle
Estimateur direct
G1_VENTES_WEB_bis
Linéaire (G1_VENTES_WEB_bis)
Première bissectrice


Estimateur issu du modèle
Estimateur direct
G5_VENTES_EDI_bis
Linéaire (G5_VENTES_EDI_bis)
Première bissectrice


Estimateur issu du modèle
Estimateur direct
G7_ACHATS_WEB_bis
Linéaire (G7_ACHATS_WEB_bis)
Première bissectrice


Estimateur issu du modèle
Estimateur direct
G8_ACHATS_EDI_bis
Linéaire (G8_ACHATS_EDI_bis)
Première bissectrice