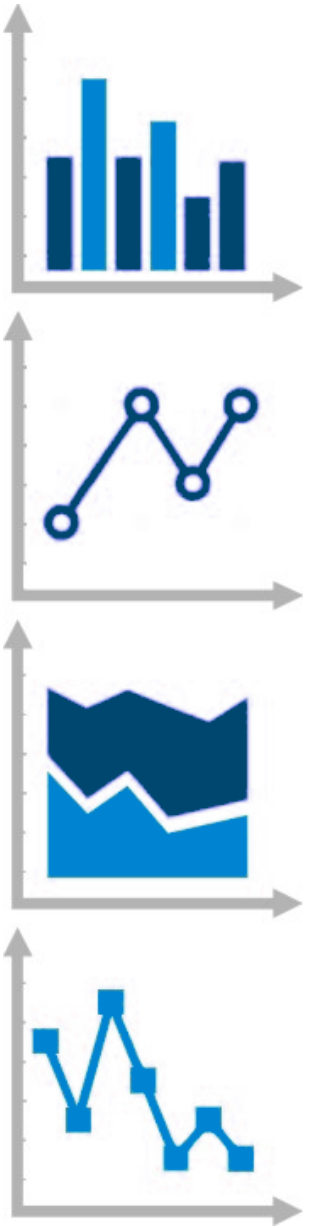
Estimations sur petits domaines de l'enquête TIC entreprises

Nadège Pradines

Division Enquêtes thématiques et études transversales,
Direction des Statistiques d'Entreprises
Insee



Insee
Mesurer pour comprendre

Sommaire

- L'enquête TIC et les régions
- Comment obtenir un résultat assez précis ?
- Estimations localisées
  - Estimations directes
  - Estimations synthétiques
  - Estimations mixtes de Fay-Herriot
- Conclusion

L’enquête TIC en entreprises (1/2)

- Enquête annuelle (dernier millésime disponible : 2016) auprès d’unités légales de 10 personnes ou plus
- Secteurs principalement marchands, hors secteurs financiers et d’assurance et agriculture
- Vise à mieux connaître l’informatisation et la diffusion des technologies de l’information et de la communication (TIC) et le commerce électronique au sein des entreprises

L’enquête TIC en entreprises (2/2)

- Des thématiques/questions récurrentes, sur un rythme annuel ou pluriannuel
- Des questions majoritairement qualitatives
- Thèmes : Emploi de personnel spécialisé, accès à internet en haut début, accès à internet mobile, site web, usage des médias sociaux, accès distant aux outils de l’entreprise, publicité en ligne, analyse de big data, réception de commandes en ligne, achats électroniques, etc.

Le plan de sondage

- Chaque année, 12 500 unités dans l'échantillon, représentatives de 200 000 unités légales en France.
- Echantillon stratifié par chiffre d'affaires, nombre de personnes, secteur d'activité (environ 200 strates non vides, certaines exhaustives)
  - La région n'est pas prise en compte
  - DROM et Corse très peu présents (adresse du siège : 11 en Guyane, 39 en Guadeloupe, 41 en Martinique, 47 en Corse, 94 à la Réunion)
- Renouvelé par moitié chaque année pour stabiliser les variations des indicateurs quantitatifs
  - Pour chaque strate, il existe deux poids (n et n-1) selon l'année à laquelle l'unité a rejoint l'échantillon

Enquêtes thématiques et estimations régionales

- Historiquement, des estimations régionales des enquêtes thématiques entreprises sont produites au moyen d'extensions de l'enquête principale (collecte dédiée)
- CIS 2008 et 2010 : travaux d'estimations régionales par des méthodes de modélisation au PISE de Nantes
- TIC : extension DOM en 2007, Basse-Normandie en 2013.

  → Décision en CD à l'issue de BN2013 de ne plus faire d'extension régionale. La pression sur les ressources conduit à privilégier des alternatives aux extensions régionales
- À l'Insee, la plupart des méthodes d'estimation sur petits domaines ont été mises en œuvre sur des données population-ménages (taux de pauvreté régionaux, taux d'illettrisme régionaux...)

La régionalisation des unités (1/2)

- Une unité légale peut être implantée dans plusieurs régions : lesquelles prendre pour décrire les « entreprises de la région » ?
- Critère du PSAR EER : unités quasi monorégionales, c'est-à-dire dont au moins 80 % de l'activité (effectifs) est localisée dans la même région
- Ce critère est fourni par le FARE localisé 2014
  - … qui ne prend pas en compte les nouvelles régions 2016
- Au final, 8 000 unités participent aux estimations régionales *métropolitaines*, parmi les 10 000 répondantes à l'enquête.

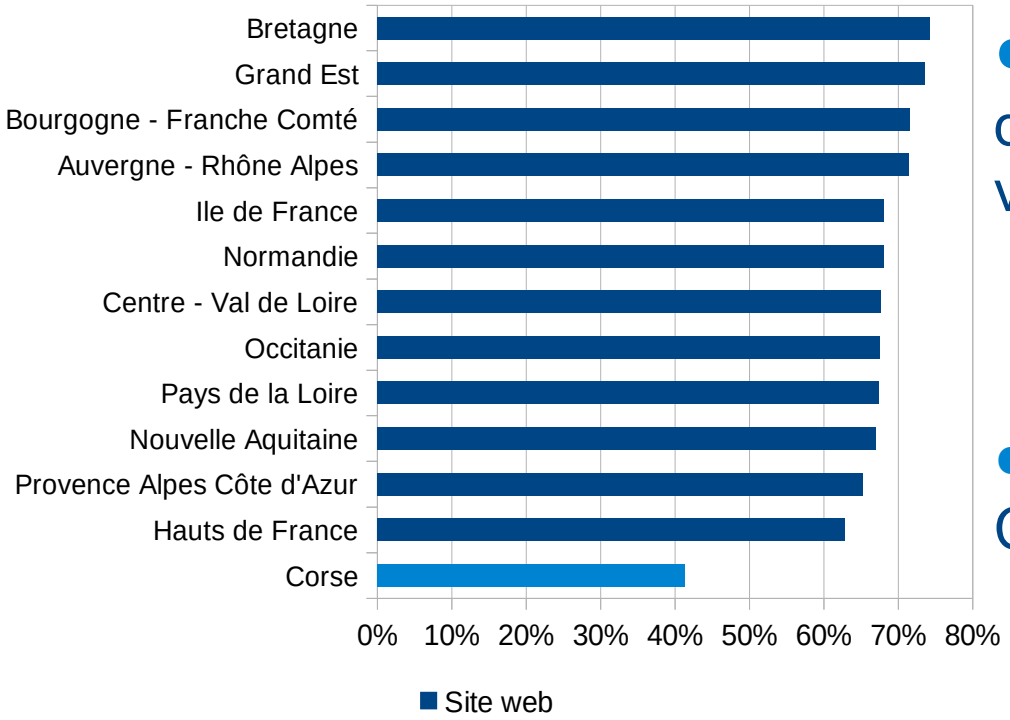
La régionalisation des unités (2/2)

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**

Entreprises non retrouvées dans le Fare : 147

Estimations directes

- Des estimations directes sont produites et leur précision est calculée avec la macro Everest



- Présence d'un site web : la plupart des régions ont un coefficient de variation inférieur à 5 %

C'est plutôt bon !

- Le coefficient de variation pour la Corse est de 18,3 %.

- Pour d'autres paramètres, les estimations directes sont moins bonnes

Objectifs sur la qualité

- Pour ces travaux, on prend pour cible CV < 17,5 %
- On veillera également à limiter le biais autant que possible
- On souhaite idéalement une précision suffisante pour comparer les régions entre elles

Comment améliorer la qualité ? (1/3)

- Le calage sur marges de la base TIC est réalisé en fonction des besoins de diffusion par taille d'entreprise et secteur d'activité à l'échelle nationale

→ Réaliser plutôt un calage par région, sur les seules unités qui y sont localisées

→ Choisir des variables de calage notamment explicatives du niveau régional

Estimateur direct avec calage sur marges régionales

Comment améliorer la qualité ? (2/3)

- Formuler une hypothèse : il n'existe pas d'effet propre à chaque région qu'on ne puisse pas modéliser par des données disponibles sur la région...

  → … c'est-à-dire qu'une fois les caractéristiques de la région prise en compte, une unité normande peut aussi être utilisée pour estimer un résultat occitan.

  → L'outil : un calage sur marges, pour chaque région, de l'ensemble des unités QMR de l'enquête

  → La variance sera faible par construction ; le biais doit en revanche être surveillé.

Estimateur indirect modélisé à partir des données d'enquête

Comment améliorer la qualité ? (3/3)

- Un hybride entre estimateur direct et estimateur modélisé : le modèle de Fay-Herriot (modèle linéaire généralisé appliqué au domaine)

  → Le modèle préfère l'estimateur direct quand il est bon, un estimateur modélisé à partir d'une information auxiliaire sinon

  → L'outil permet de calculer le coefficient de variation en sortie. Si les conditions d'application de la méthode sont respectées, l'estimateur de FH est sans biais.

Estimateur composite

L'information auxiliaire

Pour alimenter calages sur marges et modèle de FH

- Les données des référentiels statistiques (base de sondage)
- Les DADS 2014
- Les liasses fiscales 2015
- Le FARE 2014
- Lifi 2014
- Base communale 2016
- Données de couverture mobile de l'Arcep par commune
- Données de couverture très haut débit de la mission France THD par commune

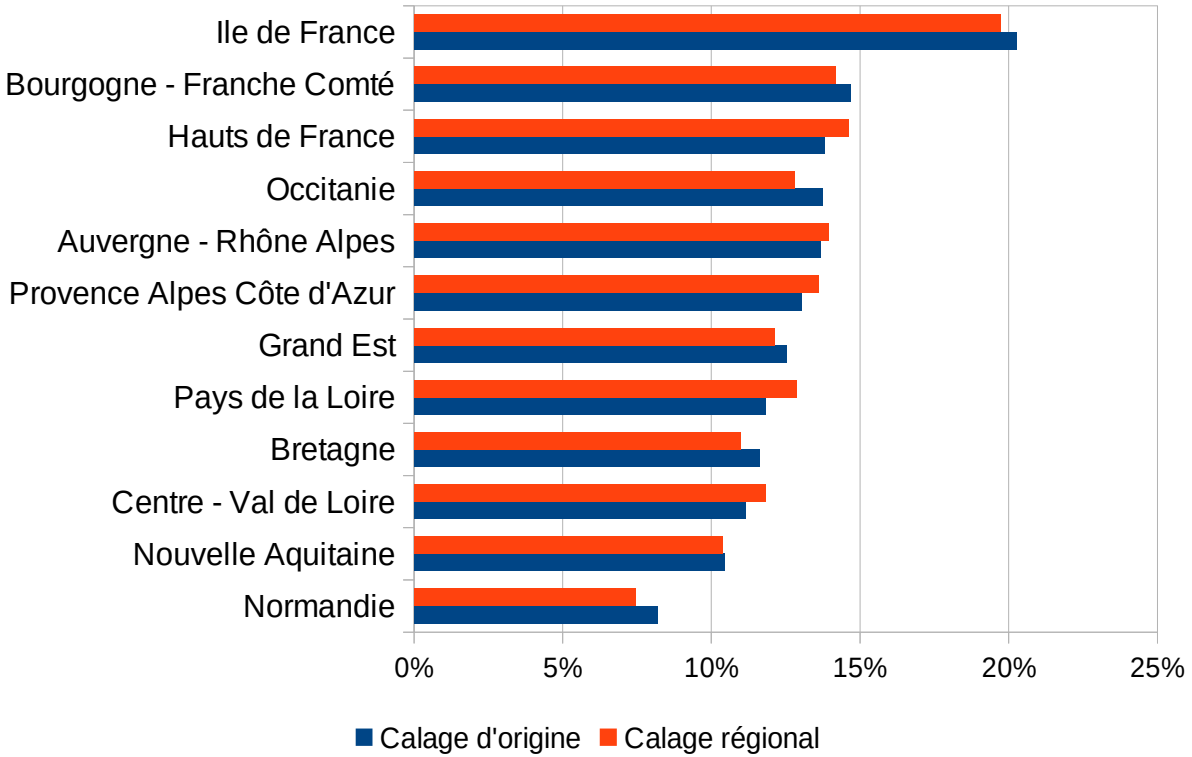
Calage sur marges régionales (1/2)

- Pour améliorer l'estimateur direct, calage sur des marges régionales
- Les marges sont choisies par régression au sein d'informations auxiliaires individuelles préalablement mises en forme et appariées avec la base de sondage
  - On choisit des marges *dans l'ensemble* explicatives, même s'il faudrait, en toute rigueur, faire autant de calages que de paramètres de l'enquête à estimer
- Les calages sont réalisés avec la méthode linéaire tronquée
  - Le calage de la Corse (sur un nombre réduit de marges) augmente la variance (?!) et est mis de côté : possible problème lié aux justifications asymptotiques du calage.

Calage sur marges régionales (2/2)

- Le calage diminue les coefficients de variation de quasiment toutes les estimations. L'estimateur direct est supposé sans biais.

- Exemple : emploi de personnel spécialisé en TIC



Coefficient de variation	calage d'origine	calage régional
Ile de France	4,8 %	4,3 %
Centre - Val de Loire	16,3 %	13,8 %
Bourgogne - Franche Comté	13,7 %	13,0 %
Normandie	16,6 %	15,0 %
Hauts de France	10,5 %	8,6 %
Grand Est	9,8 %	9,3 %
Pays de la Loire	11,6 %	9,5 %
Bretagne	12,7 %	9,6 %
Nouvelle Aquitaine	11,6 %	9,9 %
Occitanie	10,5 %	8,9 %
Auvergne - Rhône Alpes	7,5 %	6,6 %
Provence Alpes Côte d'Azur	10,9 %	9,9 %

→ Il est possible d'améliorer encore la précision

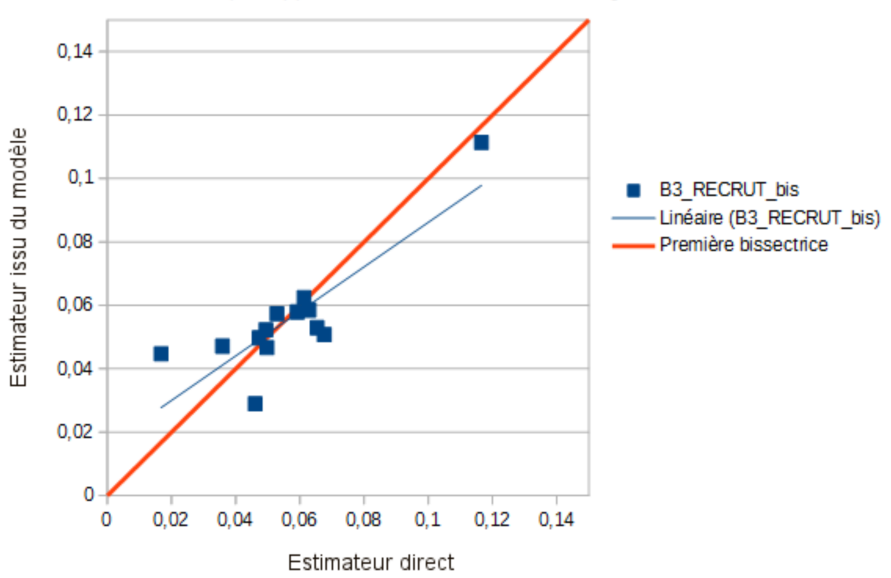
→ Certaines estimations continuent d'avoir un CV au-delà de la cible

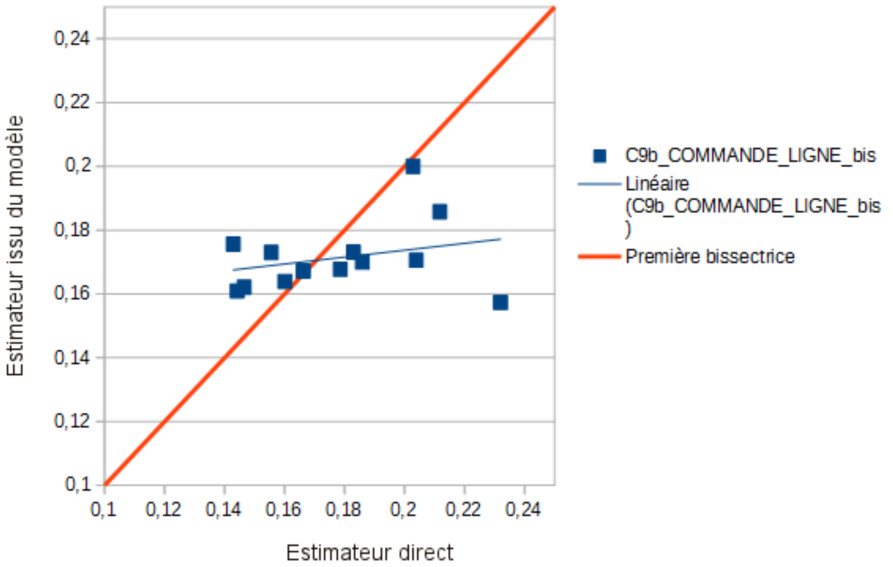
Estimateur synthétique (1/2)

- Modèle : toute la spécificité régionale est contenue dans l'information auxiliaire régionale. Un calage de l'ensemble des unités de la base sur cette information fournit un estimateur synthétique (niveau individu).
- Le calage est mis en œuvre sur les mêmes marges que le calage de l'estimateur direct. Toutes les unités de la base QMR participent au calage.
- Le grand nombre d'unités de la base assure un CV bas. Il y a en contrepartie un fort risque de biais.

Estimateur synthétique (2/2)

- L'étude du biais se fait par analyse graphique par rapport à l'estimateur direct d'origine
- Le biais est parfois très fort (*shrinkage* rédhibitoire)
- Il est possible que ce fort biais vienne d'un mauvais choix des marges (choisies pour être adaptées à *une majorité* de variables)





Estimateur de Fay-Herriot (1/2)

- L'estimateur de Fay-Herriot est un estimateur issu d'un modèle linéaire général. Il est composite et donne la préférence à l'estimateur direct lorsqu'il est bon.
- Chaque estimation est produite avec comme information auxiliaire les meilleurs régresseurs au niveau région
- L'estimateur direct est l'estimateur calé sur des marges régionales, hormis pour la Corse (traitement simplifié).
- Le modèle produit la variance associée aux estimateurs. Plusieurs méthodes sont possibles pour estimer la variance de l'effet local (toutes ne convergent pas).

Erreurs quadratiques moyennes

Erreur quadratique moyenne des estimateurs de Fay-Herriot de B1_RECRUT

	Méthode d'estimation de la variance de l'effet local			
région	**FH**	**ADM**	**REML**	**WF**
11	0,00005	0,00005	0,00005	0,00008
24	0,00014	0,00014	0,00013	0,00016
27	0,00016	0,00016	0,00015	0,00020
28	0,00013	0,00013	0,00013	0,00010
32	0,00011	0,00011	0,00011	0,00013
44	0,00010	0,00010	0,00010	0,00010
52	0,00011	0,00011	0,00011	0,00011
53	0,00013	0,00013	0,00013	0,00009
75	0,00011	0,00010	0,00010	0,00009
76	0,00011	0,00011	0,00011	0,00011
84	0,00007	0,00007	0,00007	0,00008
93	0,00011	0,00011	0,00011	0,00013
94	0,00030	0,00033	0,00029	0,00042

presque identiques

REML meilleur pour reg 24, 27 et 94

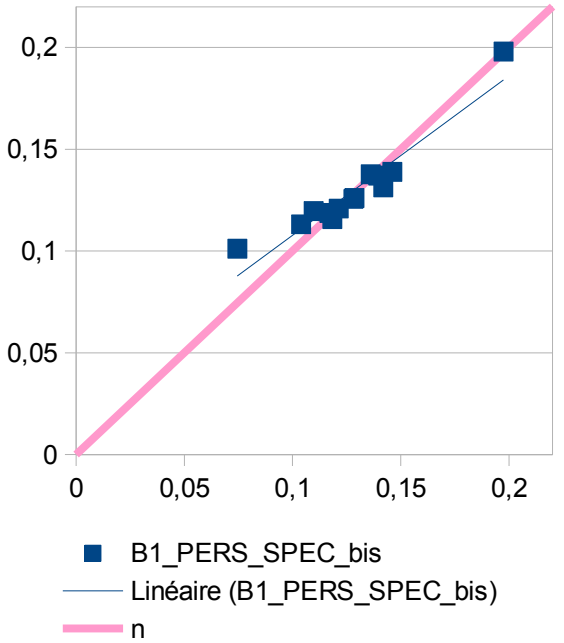
REML meilleur pour reg 24, 27, 75 et 94

variable ; WF bien moins bon pour la Corse

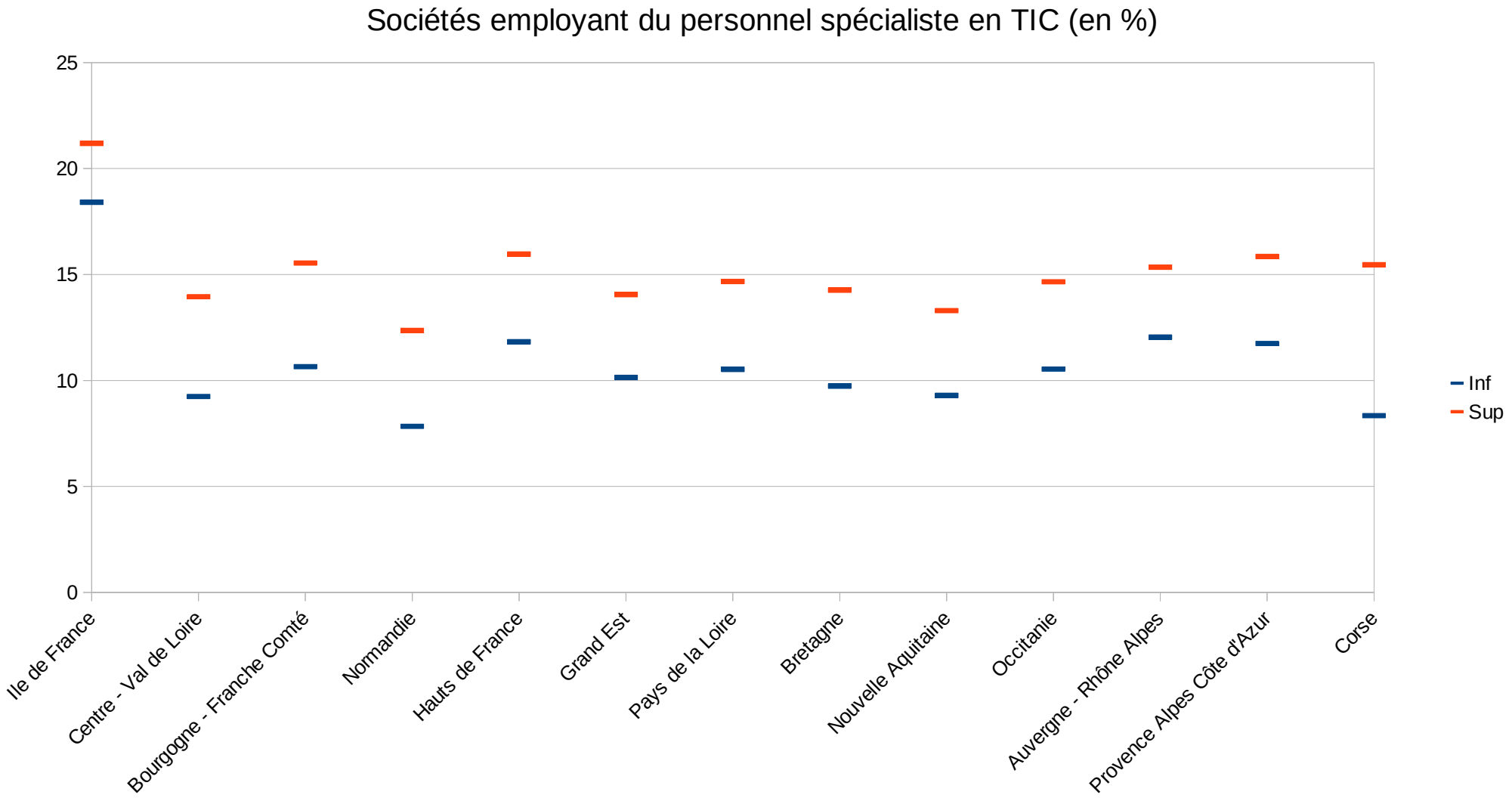
Insee

Estimateur de Fay-Herriot (2/2)

- La méthode de maximum de densité ajusté converge pour toutes les régions.
- La variabilité est améliorée, en particulier pour la Corse (p. ex. pour B1, CV=11,9 %) ; certaines des estimations pour la Corse demeurent cependant au-delà de la cible de précision (G1, G5, C15, D1).
- Le biais apparent est faible :
- Néanmoins, certaines régions n'ont pas gagné autant en précision qu'on l'aurait souhaité.

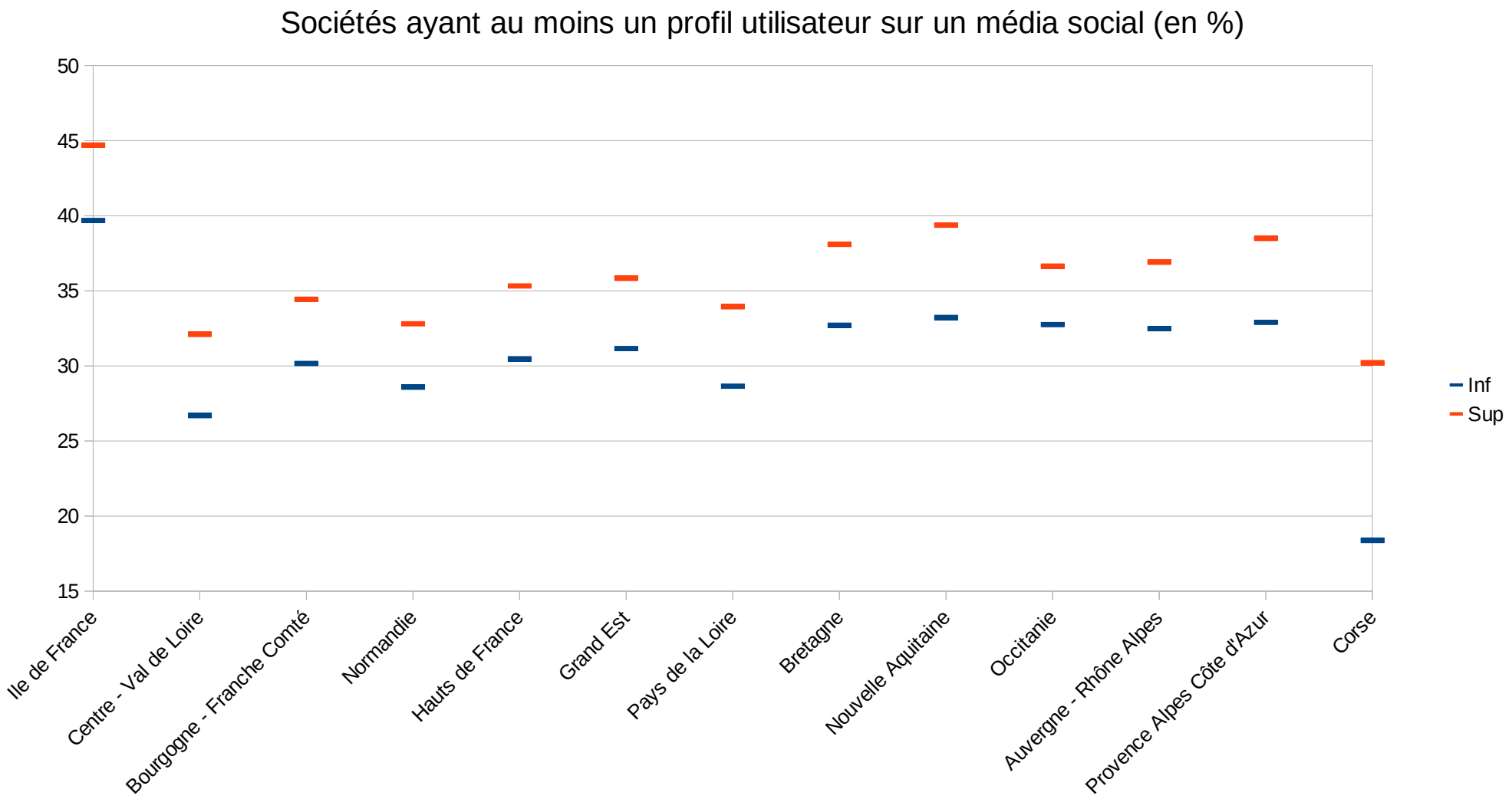


Estimations Fay-Herriot (ADM) (1/5)



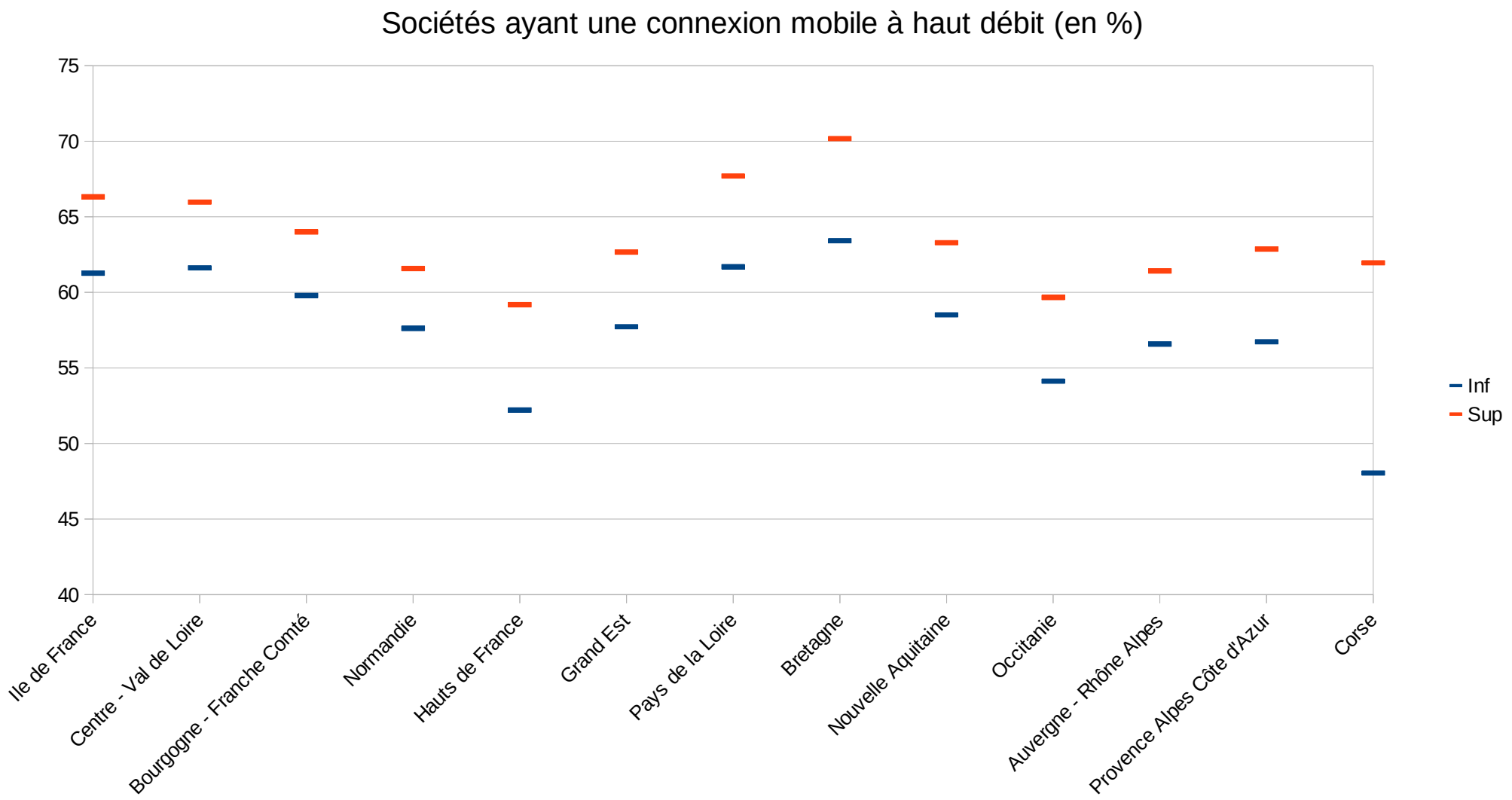
Insee

Estimations Fay-Herriot (ADM) (2/5)



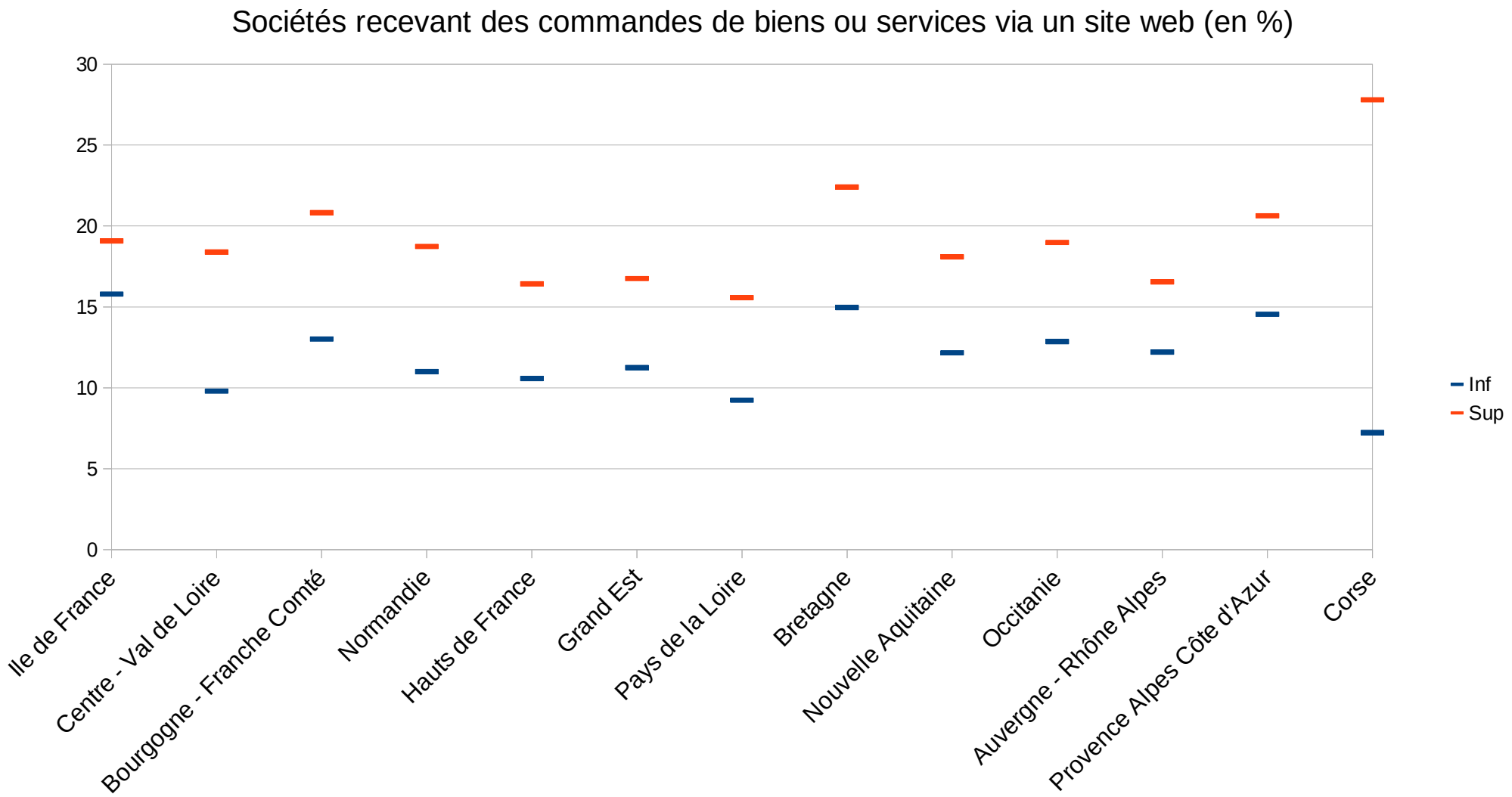
Insee

Estimations Fay-Herriot (ADM) (3/5)



Insee

Estimations Fay-Herriot (ADM) (4/5)



Insee

Enseignements pratiques

- Il est possible d'obtenir des estimateurs régionaux de TIC *raisonnablement* bons
  - Si on n'est pas *trop* ambitieux sur la qualité ;
  - Si on *croit* au modèle (dans l'ensemble bon, mais possiblement faux localement) ;
  - Si on a le *temps*.
- Les différences entre régions ne sont pas très fortes et la qualité ne permet pas d'établir de palmarès des régions.

Pour aller plus loin sur la méthode

- Paradoxalement, un plus grand nombre de domaines fournirait, théoriquement, de meilleures estimations de Fay-Herriot (« *m* grand ») si on croit au modèle.
- Des modélisations au niveau départemental ou sectoriel par région demanderaient davantage de travaux.
- Des modèles de niveau individu pourraient être réalisés sur les variables quantitatives (macro StatsCan) et qualitatives (modèle linéaire mixte généralisé sous SAS)

Estimations régionales de TIC entreprises

Merci pour votre attention !