Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Les médias sociaux comme source de données pour les statistiques officielles; l'Indice de confiance des consommateurs des Pays-Bas

par Jan van den Brakel, Emily Söhler, Piet Daas et Bart Buelens

Date de diffusion : le 21 décembre 2017

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 201-230
Statistique Canada, nº 12-001-X au catalogue

Les médias sociaux comme source de données pour les statistiques officielles; l'Indice de confiance des consommateurs des Pays-Bas

Jan van den Brakel, Emily Söhler, Piet Daas et Bart Buelens[1]

Résumé

L'article aborde la question de savoir comment utiliser des sources de données de rechange, telles que les données administratives et les données des médias sociaux, pour produire les statistiques officielles. Puisque la plupart des enquêtes réalisées par les instituts nationaux de statistique sont répétées au cours du temps, nous proposons une approche de modélisation de séries chronologiques structurelle multivariée en vue de modéliser les séries observées au moyen d'une enquête répétée avec les séries correspondantes obtenues à partir de ces sources de données de rechange. En général, cette approche améliore la précision des estimations directes issues de l'enquête grâce à l'utilisation de données d'enquête observées aux périodes précédentes et de données provenant de séries auxiliaires connexes. Ce modèle permet aussi de profiter de la plus grande fréquence des données des médias sociaux pour produire des estimations plus précises en temps réel pour l'enquête par sondage, au moment où les statistiques pour les médias sociaux deviennent disponibles alors que les données d'enquête ne le sont pas encore. Le recours au concept de cointégration permet d'examiner dans quelle mesure la série de rechange représente les mêmes phénomènes que la série observée au moyen de l'enquête répétée. La méthodologie est appliquée à l'Enquête sur la confiance des consommateurs des Pays-Bas et à un indice de sentiments dérivé des médias sociaux.

Mots-clés : Mégadonnées; inférence sous le plan de sondage; inférence sous le modèle; prédiction immédiate; modélisation de séries chronologiques structurelle; cointégration.

1 Introduction

Habituellement, les instituts nationaux de statistique font appel à l'échantillonnage probabiliste conjugué à l'inférence sous le plan de sondage ou assistée par modèle pour produire les statistiques officielles. Le concept d'échantillonnage probabiliste aléatoire a été élaboré principalement en se fondant sur les travaux de Bowley (1926), Neyman (1934), ainsi que Hansen et Hurwitz (1943). Consulter, par exemple, Cochran (1977) ou Särndal, Swensson et Wretman (1992) pour une introduction détaillée à la théorie de l'échantillonnage. Il s'agit d'une approche généralement reconnue, puisqu'elle repose sur une théorie mathématique solide qui montre comment, moyennant la combinaison appropriée d'un plan d'échantillonnage aléatoire et d'un estimateur, des inférences statistiques valides peuvent être faites au sujet de grandes populations finies en se basant sur des échantillons relativement petits. En outre, le degré d'incertitude découlant de l'utilisation de petits échantillons peut être quantifié au moyen de la variance des estimateurs.

Une pression constante s'exerce sur les instituts nationaux de statistique afin qu'ils réduisent les coûts administratifs et le fardeau de réponse. De surcroît, la baisse des taux de réponse suscite la recherche de sources d'information statistique de rechange. Cela pourrait se faire en utilisant des données administratives, comme celles des registres de l'impôt, ou d'autres grands ensembles de données – ce qu'il est convenu

1. Jan van den Brakel, Statistics Netherlands, Methodology Department, Heerlen, Pays-Bas et Maastricht University School of Business and Economics, Department of Quantitative Economics, Pays-Bas. Courriel : ja.vandenbrakel@cbs.nl; Emily Söhler, étudiante en économétrie, Maastricht University; Piet Daas et Bart Buelens, Statistics Netherlands, Methodology Department, Heerlen, Pays-Bas.

d'appeler les mégadonnées – qui sont générés en tant que sous-produit des processus non reliés directement à la production de statistiques. Les renseignements sur l'heure et le lieu de l'activité de réseau fournis par les compagnies de téléphonie mobile, les messages sur les médias sociaux provenant de Twitter et de Facebook, ainsi que les comportements de recherche sur Internet provenant de Google Trends en sont des exemples. Un problème commun à ces sources de données est que le processus de génération des données est inconnu et vraisemblablement sélectif en ce qui a trait à la population cible recherchée. Par conséquent, l'utilisation de ces données pour la production de statistiques officielles représentatives de la population cible pose un problème difficile. Il n'existe aucun plan d'échantillonnage aléatoire facilitant la généralisation des conclusions et des résultats obtenus avec les données disponibles à une population cible plus grande. Donc, l'extraction d'information statistiquement pertinente à partir de ces sources est une tâche compliquée (Daas et Puts, 2014a).

Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau (2013) étudient le problème de l'utilisation d'échantillons non probabilistes et mentionnent que des procédures d'inférence sous le plan de sondage peuvent être appliquées pour corriger le biais de sélection. Buelens, Burger et van den Brakel (2015) explorent la possibilité d'utiliser des algorithmes statistiques d'apprentissage automatique pour corriger le biais de sélection. Au lieu de remplacer les données d'enquête par des données administratives ou des mégadonnées, on peut se servir de ces sources pour améliorer l'exactitude des données d'enquête dans les procédures d'inférence sous un modèle. Marchetti, Giusti, Pratesi, Salvati, Giannotti, Perdreschi, Rinzivillo, Pappalardo et Gabrielli (2015), ainsi que Blumenstock, Cadamuro et On (2015) ont utilisé des mégadonnées comme source d'information auxiliaire pour des modèles transversaux d'estimation sur petits domaines.

De nombreuses enquêtes réalisées par les instituts nationaux de statistique sont des enquêtes répétées. Dans le présent article, nous appliquons une approche de modélisation de séries chronologiques structurelle multivariée pour combiner les séries obtenues au moyen d'une enquête répétée avec des séries provenant d'autres sources de données. Cet exercice répond à plusieurs objectifs. Premièrement, une procédure d'estimation fondée sur un modèle de séries chronologiques augmente la précision des estimations directes en tirant parti de la corrélation temporelle entre les estimations directes issues des diverses éditions de l'enquête. Le recours à la modélisation de séries chronologiques dans le but d'améliorer la précision des données d'enquête a été envisagé par de nombreux auteurs en remontant jusqu'à Blight et Scott (1973). Deuxièmement, l'extension du modèle de séries chronologiques au moyen d'une série auxiliaire permet de modéliser la corrélation entre les composantes non observées des modèles de séries chronologiques structurels, par exemple, les composantes de tendance et saisonnière. Harvey et Chung (2000) proposent un modèle de séries chronologiques pour l'Enquête sur la population active au Royaume-Uni étendu par une série sur les nombres de bénéficiaires de prestations. Si un tel modèle révèle des corrélations fortement positives entre les composantes, cela pourrait accroître encore davantage la précision des estimations des séries chronologiques de l'enquête. Les indicateurs dérivés des médias sociaux sont généralement disponibles plus fréquemment que la série reliée obtenue au moyen d'enquêtes périodiques. L'approche de modélisation de séries chronologiques susmentionnée peut donc être utilisée pour faire des prédictions précoces en temps réel quant aux résultats de l'enquête, au moment où les données des médias sociaux sont

disponibles, tandis que celles de l'enquête ne le sont pas encore. Dans ce cas, les données des médias sociaux constituent une forme de prédiction immédiate. Troisièmement, on peut appliquer le concept de cointégration dans le contexte des modèles espace-état multivariés pour déterminer dans quelle mesure les deux séries sont identiques. Si les composantes tendance des deux séries observées sont cointégrées, ces séries ont pour moteur une tendance commune sous-jacente. On peut soutenir que si une série auxiliaire est cointégrée à la série de l'enquête, les deux séries représentent le même processus stochastique sous-jacent. Cet argument pourrait servir à motiver qu'une statistique mesurée au moyen d'une source de mégadonnées est représentative d'une population cible. Toutefois, cet argument est plutôt empirique et moins solide que la théorie de l'échantillonnage probabiliste, qui prouve que l'échantillonnage aléatoire combiné à un estimateur (approximativement) sans biais sous le plan produit des statistiques représentatives.

L'Enquête sur la confiance des consommateurs (ECC) des Pays-Bas est une enquête réalisée mensuellement auprès d'environ 1 000 personnes en vue de mesurer les sentiments de la population néerlandaise au sujet du climat économique au moyen de ce que l'on appelle l'Indice de confiance des consommateurs (ICC). Daas et Puts (2014b) ont élaboré, à partir des plateformes de médias sociaux, indépendamment de l'ECC, un indice de sentiments qui s'est avéré très bien reproduire l'ICC. Cet indice est nommé Indice basé sur les médias sociaux (IMS). Dans le présent article, nous appliquons l'approche de modélisation de séries chronologiques structurelle multivariée susmentionnée aux deux séries pour tenter d'améliorer la précision de l'ICC. Nous illustrons aussi comment l'IMS peut être utilisé dans ce modèle de séries chronologiques pour faire des prédictions précoces ou prédictions immédiates de l'ICC.

À la section 2, nous décrivons le plan de sondage de l'ECC et la procédure d'estimation utilisée pour produire l'ICC. L'approche suivie par Daas et Puts (2014b) pour construire un indice de sentiments à partir des plateformes de médias sociaux est également décrite. À la section 3, nous proposons un modèle de séries chronologiques structurel pour la série de l'ICC et la série de l'IMS. À la section 4, nous présentons les résultats obtenus au moyen de ce modèle. Enfin, à la section 5, nous concluons l'article par une discussion.

2 Données

2.1 Enquête sur la confiance des consommateurs des Pays-Bas

L'Indice de confiance des consommateurs (ICC), qui est basé sur une enquête mensuelle appelée Enquête sur la confiance des consommateurs (ECC), mesure l'opinion des ménages résidant aux Pays-Bas au sujet du climat économique en général et de leur situation financière en particulier. L'ECC est une enquête continue. Chaque mois, un échantillon autopondéré d'environ 2 500 ménages est tiré selon un plan d'échantillonnage à deux degrés stratifié d'une base de sondage dérivée du Registre municipal des Pays-Bas. Des intervieweurs prennent contact avec les ménages dont le numéro de téléphone est connu pour remplir le questionnaire par interview téléphonique assistée par ordinateur durant les dix premiers jours ouvrables du mois. En moyenne, on obtient un échantillon net d'environ 1 000 ménages répondants, ce qui représente un taux de réponse d'environ 40 %. Une part importante de la non-réponse correspond aux

ménages pour lesquels aucun numéro de téléphone fixe n'est disponible. Le taux de réponse des ménages dont le numéro de téléphone est connu est de l'ordre de 60 %.

L'ICC est fondé sur cinq questions auxquelles peut être donnée une réponse positive, neutre ou négative. Les questions font référence à la situation économique ou financière au cours des 12 mois précédents ou aux attentes des enquêtés au cours des 12 mois à venir. Soit $P_{1,t}^{q}$, $P_{2,t}^{q}$, et $P_{3,t}^{q}$, les pourcentages de personnes qui ont répondu à la question $q = 1, \ldots, 5,$ durant le mois t de manière positive, neutre ou négative, respectivement. L'ICC est défini comme étant la différence entre les pourcentages de réponses positives et négatives, en prenant la moyenne sur les cinq questions :

$$I_t = \frac{1}{Q} \sum_{q=1}^{Q} \left(P_{1,t}^{q} - P_{3,t}^{q} \right). \tag{2.1}$$

Puisque l'échantillon est autopondéré et qu'aucune information auxiliaire n'est utilisée dans la procédure d'estimation, les pourcentages sont estimés en prenant la moyenne d'échantillon, c'est-à-dire

$$P_{j,t}^{q} = \frac{100}{n_t} \sum_{i=1}^{n} \delta_{i,j,t}^{q}, \tag{2.2}$$

pour la question $q = 1, \ldots, 5,$ et la catégorie de réponse $j = 1, 2, 3.$ Dans (2.2), n_t est la taille d'échantillon nette durant le mois t et $\delta_{i,j,t}^{q}$ est une variable indicatrice binaire qui est égale à un si le répondant i a choisi la catégorie j pour la question q. En supposant que les ménages sont sélectionnés selon un plan d'échantillonnage aléatoire simple sans remise, on peut prouver que la variance de (2.1) peut être estimée par

$$\begin{aligned} \text{Var}\left(I_t\right) &= \frac{1}{Q^2} \sum_{q=1}^{Q} \left[\text{Var}\left(P_{1,t}^{q}\right) + \text{Var}\left(P_{3,t}^{q}\right) \right] - \frac{2}{Q^2} \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \text{Cov}\left(P_{1,t}^{q}, P_{3,t}^{q'}\right) \\ &+ \frac{1}{Q^2} \sum_{q=1}^{Q} \sum_{q' \neq q}^{Q} \left[\text{Cov}\left(P_{1,t}^{q}, P_{1,t}^{q'}\right) + \text{Cov}\left(P_{3,t}^{q}, P_{3,t}^{q'}\right) \right], \end{aligned} \tag{2.3}$$

avec

$$\text{Var}\left(P_{j,t}^{q}\right) = \frac{1}{n_t} P_{j,t}^{q} \left(100 - P_{j,t}^{q}\right), \qquad \text{Cov}\left(P_{j,t}^{q}, P_{j',t}^{q'}\right) = \frac{1}{n_t} \left(P_{jj',t}^{qq'} - P_{j,t}^{q} P_{j',t}^{q'}\right),$$

$$\text{Cov}\left(P_{j,t}^{q}, P_{j',t}^{q'}\right) = \frac{1}{n_t} \left(P_{jj',t}^{qq'} - P_{j,t}^{q} P_{j',t}^{q'}\right), \quad \text{Cov}\left(P_{j,t}^{q}, P_{j',t}^{q}\right) = -\frac{1}{n_t} P_{j,t}^{q} P_{j',t}^{q},$$

$$P_{jj',t}^{qq'} = \frac{100}{n_t} \sum_{i=1}^{n} \delta_{i,j,t}^{q} \, \delta_{i,j',t}^{q'}.$$

La figure 2.1 montre l'ICC avec un intervalle de confiance (IC) à 95 % calculé en utilisant l'approche décrite à la présente section, observé durant la période de décembre 2000 à mars 2015. La publication officielle de l'ICC est manquante pour octobre 2013.

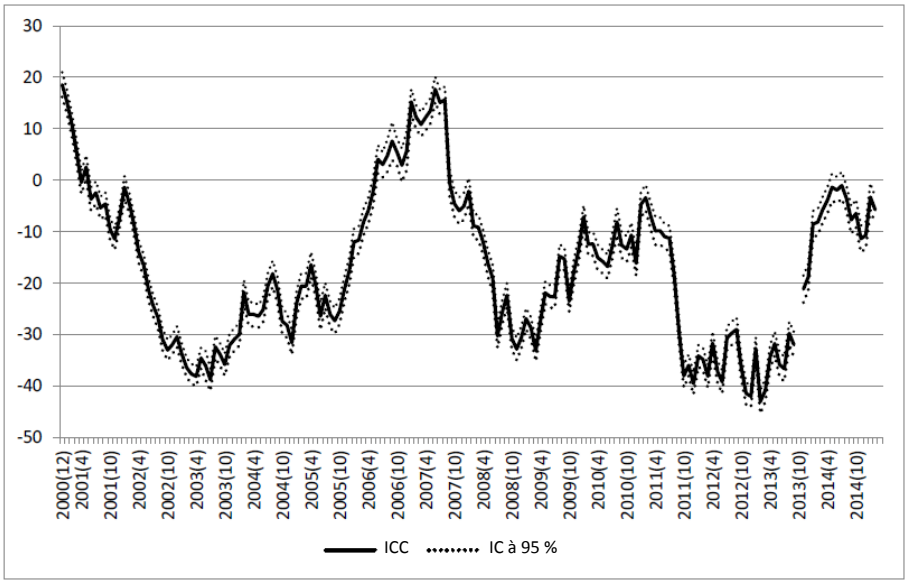

Figure 2.1 Indice de confiance des consommateurs (ICC) avec un intervalle de confiance à 95 %.

2.2 Sentiments sur les médias sociaux

Afin d'essayer de réduire les coûts administratifs et le fardeau de réponse, Daas et Puts (2014b) ont élaboré, en prenant pour sources les médias sociaux, un indice de sentiments pouvant servir d'indicateur de remplacement de l'ICC. Ils sont partis des messages affichés sur les plateformes de médias sociaux les plus populaires aux Pays-Bas et rédigés en néerlandais. Ils les ont classés comme étant des messages positifs, neutres ou négatifs en se servant d'une variante de la classification fondée sur une analyse au niveau de la phrase (Pang et Lee, 2008). L'indice est calculé en prenant la différence entre les pourcentages de messages positifs et négatifs.

Diverses combinaisons de tous les messages affichés sur Facebook et sur Twitter avec et sans application de certains filtres sur les phrases ont été comparées à l'ICC. La combinaison de tous les messages faisant partie du domaine public sur Facebook et de messages sur Twitter filtrés afin qu'ils contiennent des pronoms personnels est celle dont la corrélation avec l'ICC était la plus élevée. Il a fallu filtrer les messages sur Twitter parce qu'un grand nombre de ceux-ci ne sont pas très informatifs. Voir Daas et Puts (2014b) pour des renseignements plus détaillés. Dans le cadre de leur étude, Daas et Puts (2014b) ont également constaté que d'importants changements de comportement du public sur les médias sociaux, tels ceux causés par les grands événements et les variations du nombre de messages affichés sur chaque plateforme, ont un effet perturbateur sur la série. L'indicateur final qu'ils proposent correspond à la moyenne du sentiment dans les messages sur Facebook et sur Twitter durant chaque période.

À la figure 2.2, l'Indice basé sur les médias sociaux (IMS) est comparé à l'ICC pour la période allant de juin 2010 à mars 2015. Les deux séries se situent clairement à des niveaux différents, mais présentent une évolution plus ou moins similaire. Au cours de la période présentée, l'ICC est constamment négatif, tandis que l'IMS est constamment positif. La taille ou amplitude des mouvements est également considérablement

plus grande pour l'ICC que pour l'IMS. De nombreux facteurs sont à l'origine de cette différence, puisque l'ICC est fondé sur une enquête dont la collecte des données est effectuée par téléphone et l'IMS est fondé sur la classification de messages affichés sur Twitter et sur Facebook. La question intéressante est celle de savoir dans quelle mesure l'évolution des deux séries est comparable.

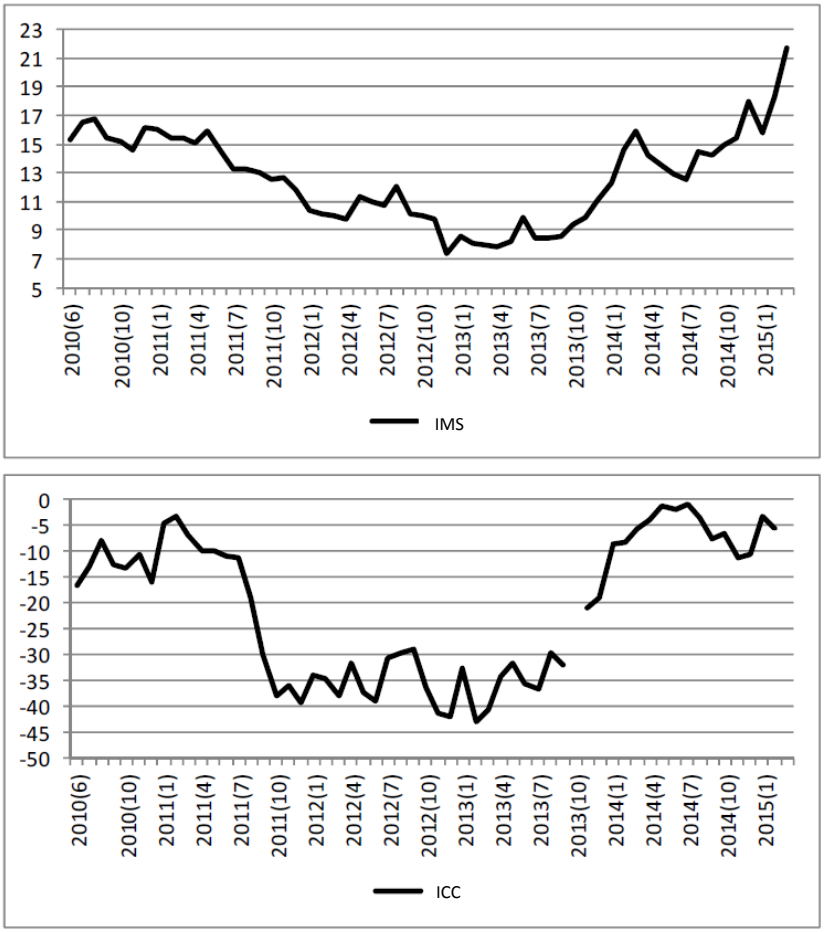

Figure 2.2 Comparaison de l'Indice basé sur les médias sociaux (IMS, graphique supérieur) à l'Indice de confiance des consommateurs (ICC, graphique inférieur).

2.3 Aspects qualitatifs de l'ICC et de l'IMS

L'exactitude d'une statistique est mesurée par sa variance et son biais. Pour simplifier, nous faisons uniquement la distinction entre le biais de sélection et le biais de mesure. La variance d'une statistique issue d'un sondage, comme l'ICC, dépend de la taille de l'échantillon et représente habituellement une part importante de l'incertitude de la statistique d'échantillon. Dans le cas des sources de mégadonnées, le concept de variance d'échantillonnage n'a pas de sens puisque le processus de génération des données n'est pas un échantillonnage probabiliste dans une population cible donnée. En lieu et place, on pourrait se servir des composantes de la variance du modèle utilisé pour décrire le processus supposé de génération des données comme mesure d'exactitude. La variance sous le modèle des statistiques obtenues par application de modèles de séries chronologiques à une série issue de données recueillies sur Internet ou sur les médias sociaux sera systématiquement positive selon la volatilité de la série, laquelle dépend principalement de la

fréquence d'observation de la série et de la dynamique du processus stochastique plutôt que du volume des données.

Le biais de sélection d'une statistique issue d'un sondage est approximativement nul dans des conditions de réponse complète. Toutefois, en pratique, il existe un biais de sélection qui résulte de la non-réponse sélective, de la couverture incomplète de la base de sondage et de la mesure dans laquelle la stratégie de travail sur le terrain permet d'atteindre avec succès la population cible. Dans le cas de l'ICC, un contact est pris uniquement avec les membres de la population dont on connaît le numéro d'appel d'une ligne téléphonique fixe et le taux de réponse de cette sous-population est d'environ 60 %. Dans le cas des sources de mégadonnées, le biais de sélection est généralement inconnu. Dans le présent article, nous appliquons le concept de cointégration pour évaluer le degré auquel l'IMS et l'ICC mesurent le même concept. Notons, cependant, que dans le cas de la cointégration, l'IMS pourrait refléter un biais de sélection relatif à la non-réponse et à la couverture similaire à celui de l'ICC. Baker et coll. (2013) ont fait remarquer qu'il existe des similarités entre le biais de sélection observé pour les échantillons probabilistes et pour l'approche non probabiliste utilisée par les sources de données telles que les médias sociaux.

Le biais de mesure de statistiques issues d'un sondage dépend habituellement de la mesure dans laquelle les variables conceptuelles que l'on veut mesurer sont concrétisées dans le questionnaire, mais aussi du mode de collecte des données et de la qualité des intervieweurs. Dans le cas des enquêtes, le problème de biais de mesure se pose parce que les variables d'intérêt sont mesurées indirectement en ce sens qu'on demande aux participants de faire des déclarations au sujet de leur comportement, ce qui introduit toutes sortes d'erreurs de mesure. Dans le cas de l'ICC, on peut se demander dans quelle mesure les enquêtés sont capables d'exprimer leur confiance à long terme dans l'économie et dans quelle mesure leur réponse est influencée par des émotions de court terme. Ces problèmes ne se posent pas dans le cas des mégadonnées si celles-ci contiennent des mesures directes du comportement des gens. Dans le cas d'un indice basé sur les médias sociaux tel que l'IMS, on peut se demander à quel point il mesure un concept similaire à celui de l'ICC. À la sous-section 2.2, nous avons déjà mentionné que les changements importants de comportement du public sur les médias sociaux perturbent la série. Tout spécialement à la fin de la série, un changement soudain de comportement sur les médias sociaux sera très difficile à distinguer d'un vrai point de retournement. Par exemple, une série issue de Google Trend sur les recherches liées aux emplois vacants pourrait suivre la courbe d'une série officielle sur le chômage. Cette série mesure le chômage, cependant, le comportement de recherche avant le début de la crise financière de 2009 pourrait être entièrement différent de celui de la période suivant directement la crise financière, ce qui invalide le concept que l'on veut mesurer.

3 Modélisation de séries chronologiques structurelle de l'ICC et de l'IMS

La présente section décrit l'élaboration de modèles de séries chronologiques structurels univarié et bivarié pour l'ICC et l'IMS. Dans un modèle de séries chronologiques structurel, la série est décomposée en une composante tendance, une composante saisonnière, d'autres composantes cycliques, une composante de régression et une composante irrégulière. Chaque composante est supposée suivre un modèle stochastique, ce qui permet que les composantes tendance, saisonnière et cyclique, mais aussi les

coefficients de régression dépendent du temps. Au besoin, des composantes autorégressives-moyennes mobiles (ARMA) peuvent être ajoutées pour tenir compte de l'autocorrélation dans la série au-delà de ces composantes structurelles. Consulter Harvey (1989) ou Durbin et Koopman (2012) pour des précisions au sujet de la modélisation de séries chronologiques structurelle.

La question abordée dans le présent article est celle de savoir dans quelle mesure l'IMS suit une courbe semblable à l'ICC, de sorte que l'IMS puisse être utilisé dans la procédure d'estimation de l'ICC, voire, dans le cas le plus extrême, le remplacer. Pour traiter cette question, nous élaborons un modèle de séries chronologiques structurel bivarié pour l'ICC et pour l'IMS, et modélisons la corrélation entre les termes de perturbation des différentes composantes du modèle structurel pour les deux séries. Nous appliquons le concept de cointégration pour déterminer si les composantes non observées des deux séries sont sous-tendues par des facteurs communs. Si, par exemple, les tendances des deux séries sont sous-tendues par une tendance commune, on pourrait soutenir que l'IMS représente une évolution des sentiments comparable à l'ICC. L'IMS pourrait aussi être utilisé comme une série auxiliaire dans une procédure d'estimation de l'ICC fondée sur un modèle ou dans une procédure de prédiction immédiate pour obtenir des estimations en temps réel plus précises.

3.1 Modèle univarié de la série de l'ICC

En guise de première étape, nous proposons un modèle de séries chronologiques univarié pour la série de l'ICC. Selon l'approche fondée sur le plan de sondage décrite à la section 2.1, l'information observée chaque mois sur l'échantillon sert à calculer une estimation de l'ICC pour le mois en question. Un inconvénient de cette approche est que l'information observée lors des périodes précédentes n'est pas utilisée pour obtenir des estimations plus précises de l'ICC. Dans le domaine des techniques d'enquête, il est fréquent d'appliquer des modèles de séries chronologiques pour obtenir des estimations pour des enquêtes périodiques. Blight et Scott (1973) et Scott et Smith (1974) ont proposé de considérer les paramètres de population inconnus comme une réalisation d'un processus stochastique qui peut être décrit au moyen d'un modèle de séries chronologiques. Cela introduit des relations entre les paramètres de population estimés à différents points dans le temps dans le cas d'échantillons non chevauchants ainsi que chevauchants. La modélisation explicite de cette relation entre les estimations issues de l'enquête au moyen d'un modèle de séries chronologiques peut servir à combiner l'information observée sur l'échantillon dans le passé pour améliorer la précision des estimations obtenues au moyen d'enquêtes périodiques. Parmi les références clés à des auteurs qui ont appliqué l'approche des séries chronologiques aux données d'enquêtes répétées pour améliorer l'efficacité des estimations par sondage, nous mentionnerons Scott, Smith et Jones (1977), Tam (1987), Binder et Dick (1989, 1990), Bell et Hillmer (1990), Tiller (1992), Rao et Yu (1994), Pfeffermann et Burck (1990), Pfeffermann (1991), Pfeffermann et Rubin-Bleuer (1993), Pfeffermann, Feder et Signorelli (1998), Pfeffermann et Tiller (2006), Harvey et Chung (2000), Feder (2001), Lind (2005) et van den Brakel et Krieg (2009, 2015).

L'élaboration d'un modèle de séries chronologiques pour les estimations par sondage observées au moyen d'une enquête périodique débute par un modèle énonçant que l'estimation par sondage peut être décomposée en la valeur de la variable dans la population et une erreur d'échantillonnage :

$$I_t = \theta_t + e_t, \tag{3.1}$$

où θ_t désigne l'ICC réel au mois t sous un dénombrement complet de la population cible et e_t, l'erreur d'échantillonnage.

L'ICC est observé mensuellement. Par conséquent, à titre de première étape, la série du paramètre de population finie peut être décomposée en une tendance stochastique, une composante saisonnière pour modéliser les écarts systématiques par rapport à la tendance durant une année, et une composante de bruit blanc pour les variations restantes, inexpliquées. Ces considérations mènent au modèle qui suit pour la série du paramètre de population finie :

$$\theta_t = L_t + S_t + \xi_t, \tag{3.2}$$

où L_t désigne une tendance stochastique, S_t, une composante saisonnière stochastique et ξ_t, la variation inexpliquée du paramètre de population finie. L'insertion de (3.2) dans le modèle de mesure (3.1) donne

$$I_t = L_t + S_t + \xi_t + e_t. \tag{3.3}$$

Dans une enquête transversale, il est difficile de séparer l'erreur d'échantillonnage du bruit blanc du paramètre de population. Donc, les deux composantes sont combinées en un terme de perturbation

$$\upsilon_t = \xi_t + e_t. \tag{3.4}$$

Nous supposons que $E(\upsilon_t) = 0$ et $\text{Var}(\upsilon_t) = \sigma_\upsilon^2$. Pour tenir compte de la variance non homogène dans les erreurs d'échantillonnage, Binder et Dick (1990) ont proposé une erreur de mesure où les termes de perturbation υ_t sont proportionnels aux erreurs-types de I_t, c'est-à-dire

$$\upsilon_t = \sqrt{\text{Var}(I_t)}\ \tilde{\upsilon}_t, \tag{3.5}$$

avec $E(\tilde{\upsilon}_t) = 0$, $\text{Var}(\tilde{\upsilon}_t) = \sigma_\upsilon^2$ et où $\text{Var}(I_t)$ est définie par (2.3) et est utilisée comme une information a priori dans le modèle de séries chronologiques. Un tel modèle serait utile si l'erreur d'échantillonnage est plus importante que le bruit blanc dans le paramètre de population. Pour la présente application, les premières analyses indiquent que la variance du bruit blanc de la population est importante, ce qui invalide (3.5). En outre, toujours dans cette application, la variance de l'erreur d'échantillonnage est constante au cours du temps. Nous avons donc décidé de combiner l'erreur d'échantillonnage et le bruit blanc de la population, et avons supposé que la variance était constante au cours du temps. Savoir comment tenir compte de la variance d'échantillonnage est une question qui se pose également dans le cas des variances de désaisonnalisation (Pfeffermann et Sverchkov, 2014). Bell (2005) a étudié la contribution de la variance d'échantillonnage à la variance de l'erreur d'estimation des séries désaisonnalisées et à la composante non saisonnière. Dans le cas de panels (rotatifs), l'erreur d'échantillonnage peut être isolée du bruit blanc de la population. Dans le cas des enquêtes transversales répétées, il est difficile d'identifier les composantes distinctes et les deux termes sont donc combinés en un terme de perturbation qui inclut à la fois la variance d'échantillonnage et la variation inexpliquée du paramètre de population.

Un exercice approfondi de sélection du modèle a montré qu'un modèle de tendance lissé est le plus approprié pour représenter la tendance et le cycle économique dans la série de l'ICC. Le modèle de tendance lissé est défini comme étant (Durbin et Koopman, 2012) :

$$L_t = L_{t-1} + R_t,$$

$$R_t = R_{t-1} + \eta_t, \quad \mathrm{E}(\eta_t) = 0, \tag{3.6}$$
$$\mathrm{Cov}(\eta_t, \eta_{t'}) = \begin{cases} \sigma_\eta^2 & \text{si} \quad t = t' \\ 0 & \text{si} \quad t \neq t' \end{cases}.$$

L'ajout d'une composante aléatoire pour le niveau dans (3.6) améliore la log-vraisemblance de cinq unités, mais aboutit à un surajustement des données en ce sens que le signal lissé suit presque exactement la série observée, avec une très petite variance de l'erreur de mesure. Un modèle au niveau local (niveau aléatoire sans une pente) améliore la log-vraisemblance de trois unités, mais a également tendance à surajuster les données.

La composante saisonnière est modélisée par un modèle trigonométrique, qui est défini comme étant (Durbin et Koopman, 2012) :

$$S_t = \sum_{j=1}^{6} \gamma_{jt}, \tag{3.7}$$

avec

$$\begin{aligned} \gamma_{jt} &= \gamma_{jt-1} \cos(\lambda_j) + \tilde{\gamma}_{jt-1} \sin(\lambda_j) + \omega_{jt}, \\ \tilde{\gamma}_{jt} &= -\gamma_{jt-1} \sin(\lambda_j) + \tilde{\gamma}_{jt-1} \cos(\lambda_j) + \tilde{\omega}_{jt}. \end{aligned}$$

Ici, λ_j désigne la fréquence des différents cycles exprimée en radians et définie comme étant

$$\lambda_j = \frac{2\pi j}{12}, \qquad \text{pour} \qquad j = 1, \ldots, 6.$$

Pour les termes de perturbation, il est supposé que

$$E(\omega_{jt}) = 0, \qquad E(\tilde{\omega}_{jt}) = 0,$$
$$\mathrm{Cov}(\omega_t, \omega_{t'}) = \begin{cases} \sigma_\omega^2 & \text{si} \quad t = t' \\ 0 & \text{si} \quad t \neq t' \end{cases}.$$

Par souci de parcimonie, nous supposons que la structure de variance est la même avec le même hyperparamètre pour $\tilde{\omega}_{jt}$. Qui plus est, nous supposons que ω_t et $\tilde{\omega}_t$ ne sont pas corrélés.

Après l'introduction de la composante de tendance stochastique (3.6) et de la composante saisonnière (3.7), aucune autre composante cyclique n'est nécessaire. La procédure de sélection du modèle a montré que deux changements de niveau sont nécessaires pour modéliser des sauts soudains dans la série. Le premier est dû à la crise financière de septembre 2008, et le second, au ralentissement économique de septembre 2011. Enfin, une valeur aberrante ponctuelle est nécessaire pour septembre 2007. L'ajout de ces trois composantes accroît la log-vraisemblance de 15 unités. Nous arrivons ainsi au modèle qui suit pour la série observée de l'ICC

$$I_t = L_t + S_t + \beta^{07}\delta_t^{07} + \beta^{08}\delta_t^{08} + \beta^{11}\delta_t^{11} + \upsilon_t, \tag{3.8}$$

avec

$$\delta_t^{07} = \begin{cases} 1 & \text{si} \quad t = 2007(9) \\ 0 & \text{si} \quad t \neq 2007(9) \end{cases}, \qquad \delta_t^{08} = \begin{cases} 1 & \text{si} \quad t \geq 2008(9) \\ 0 & \text{si} \quad t < 2008(9) \end{cases}, \qquad \delta_t^{11} = \begin{cases} 1 & \text{si} \quad t \geq 2011(9) \\ 0 & \text{si} \quad t < 2011(9) \end{cases},$$

et β^x représente les coefficients de régression correspondants.

Enfin, des composantes autorégressives (AR) et de moyennes mobiles (MA pour *moving average*) peuvent être ajoutées au modèle de séries chronologiques structurel (3.8). Dans la présente application, rien n'indique que de telles composantes soient nécessaires, puisqu'il n'y a aucun signe évident d'une corrélation sériale résiduelle entre les innovations standardisées. L'ajout d'un processus AR(1) ou MA(1) à (3.8) augmente la log-vraisemblance de 5 et de 4,5 unités, respectivement. L'ajout de modèles AR ou MA d'ordre deux n'améliore pas davantage la log-vraisemblance. L'ajout d'un processus ARMA(1,1) n'accroît pas non plus davantage la log-vraisemblance. Un processus AR(1) ou MA(1) améliore légèrement le corrélogramme, mais augmente aussi l'erreur-type des signaux lissés filtrés. Donc, nous avons finalement choisi le modèle (3.8) pour la série de l'ICC.

Les modèles espace-état supposent que les termes de perturbation suivent des lois normales indépendantes. Ces hypothèses se traduisent en l'hypothèse que les innovations suivent des lois normales indépendantes. Le tableau A.1 en annexe donne un aperçu des statistiques de qualité de l'ajustement appliquées aux innovations standardisées. Les valeurs obtenues pour l'asymétrie, l'aplatissement et le test de Bowman-Shenton ne révèlent pas d'écarts par rapport à la loi normale pour les innovations standardisées. Les valeurs pour le test de Ljung-Box et le test de Durban-Watson n'indiquent aucune corrélation sériale entre les innovations standardisées. Ces observations sont également confirmées par un corrélogramme (non présenté). En conclusion, selon ces diagnostics, le modèle (3.8) est raisonnablement bien ajusté à la série de l'ICC.

3.2 Modèle bivarié des séries de l'ICC et de l'IMS

L'étape suivante consiste à combiner le modèle univarié pour l'ICC avec la série pour l'IMS. Avant de combiner l'ICC et l'IMS dans un modèle bivarié, nous élaborons un modèle univarié pour l'IMS afin de mieux comprendre le comportement de cette série. Une procédure de sélection de modèle similaire à celle effectuée pour la série de l'ICC à la sous-section 3.1 indique que la série observée pour l'IMS peut être modélisée avec un modèle de tendance lisse et une composante de bruit blanc pour la variation inexpliquée. Aucune présence significative d'une composante saisonnière ou d'un cycle économique n'est établie. Il n'existe aucun signe de valeur aberrante ni de changements de niveau. Nous n'avons pas inclus de composante AR(1) et MA(1) puisqu'il n'existe aucune corrélation sériale entre les innovations standardisées. Ces observations ont mené à un modèle bivarié pour l'ICC et l'IMS dans lequel l'ICC comprend une tendance et une composante saisonnière, tandis que l'IMS comprend une composante tendance.

Les tableaux A.2 et A.3 en annexe donnent un aperçu des statistiques de qualité de l'ajustement pour les innovations standardisées de l'ICC et de l'IMS, respectivement. Rien n'indique que les innovations standardisées s'écartent d'une loi normale dans l'une ou l'autre des deux séries. L'hypothèse nulle d'absence de corrélation sériale entre les innovations standardisées n'a pas pu être rejetée. Le corrélogramme des innovations pour l'IMS montre toutefois un patron saisonnier non significatif (données non présentées). Les innovations de l'IMS présentent aussi une hétéroscédasticité.

Les termes de perturbation des tendances des deux séries sont corrélés. Puisque la série pour l'IMS est disponible à partir de juin 2010, le modèle pour l'ICC contient aussi le dernier changement de niveau en septembre 2011, mais non la valeur aberrante ponctuelle en septembre 2007 et le changement de niveau en septembre 2008. Par conséquent, nous obtenons le modèle bivarié suivant :

$$\begin{pmatrix} I_t \\ X_t \end{pmatrix} = \begin{pmatrix} L_t^I \\ L_t^X \end{pmatrix} + \begin{pmatrix} S_t^I \\ 0 \end{pmatrix} + \begin{pmatrix} \beta^{11}\delta_t^{11} \\ 0 \end{pmatrix} + \begin{pmatrix} \upsilon_t^I \\ \upsilon_t^X \end{pmatrix}, \tag{3.9}$$

dans lequel L_t^I et L_t^X désignent le modèle de tendance lissé défini en (3.6) avec la structure de covariance

$$\begin{aligned}
\text{Cov}\left(\eta_t^I, \eta_{t'}^I\right) &= \begin{cases} \sigma_{\eta I}^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}, \\
\text{Cov}\left(\eta_t^X, \eta_{t'}^X\right) &= \begin{cases} \sigma_{\eta X}^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}, \\
\text{Cov}\left(\eta_t^I, \eta_{t'}^X\right) &= \begin{cases} \sigma_{\eta I}\sigma_{\eta X}\rho_\eta & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}.
\end{aligned}$$

Dans la dernière expression, ρ_η désigne la corrélation entre les perturbations de la pente de l'ICC et de l'IMS. De surcroît, S_t^I représente l'effet saisonnier défini par (3.7) et δ_t^{11}, le changement de niveau en septembre 2011 avec le coefficient de régression correspondant β^{11}. Enfin, υ_t^I et υ_t^X sont les termes de perturbation pour les séries de l'ICC et de l'IMS, qui sont définis comme il suit :

$$\begin{aligned}
E(\upsilon_t^I) &= E(\upsilon_t^X) = 0, \\
\text{Cov}\left(\upsilon_t^I, \upsilon_{t'}^I\right) &= \begin{cases} \sigma_{\upsilon I}^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}, \\
\text{Cov}\left(\upsilon_t^X, \upsilon_{t'}^X\right) &= \begin{cases} \sigma_{\upsilon X}^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}, \\
\text{Cov}\left(\upsilon_t^I, \upsilon_{t'}^X\right) &= 0 \text{ pour tout } t \text{ et } t'.
\end{aligned}$$

Si le modèle détecte une forte corrélation entre les tendances de l'ICC et de l'IMS, alors les tendances des deux séries se développeront dans la même direction plus ou moins simultanément. Dans ces conditions, l'information supplémentaire provenant de la série de l'IMS aboutira à une plus grande précision des estimations des chiffres de l'ICC. Dans le cas d'une forte corrélation entre les perturbations des tendances, c'est-à-dire si $\rho_\eta \to 1$, les tendances sont dites cointégrées. Dans ces conditions, il existe une tendance commune sous-jacente qui dicte l'évolution des tendances des deux séries observées. Pour le voir, nous notons que la matrice de covariance des perturbations de la pente est obtenue sous forme d'une décomposition en valeurs singulières :

$$\text{cov}\begin{pmatrix} \eta_t^I \\ \eta_t^X \end{pmatrix} = \begin{pmatrix} \sigma_{\eta I}^2 & \sigma_{\eta I}\sigma_{\eta X}\rho_\eta \\ \sigma_{\eta I}\sigma_{\eta X}\rho_\eta & \sigma_{\eta X}^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}. \tag{3.10}$$

Au lieu de $\sigma_{\eta I}^2$, $\sigma_{\eta X}^2$ et ρ_η, ce sont les paramètres d_1, d_2 et a qui sont estimés. Si $d_2 \to 0$, il s'ensuit que $\rho_\eta \to 1$. Dans ces conditions, la matrice de covariance des perturbations de la pente est de rang réduit et les deux tendances sont sous-tendues par une tendance commune. Cela implique que les perturbations des pentes des deux séries montent ou descendent simultanément et que les perturbations de la pente de l'IMS peuvent être prédites parfaitement à partir des perturbations de la pente de l'ICC au moyen de $\eta_t^X = a\eta_t^I$. En outre, la pente de la série de l'IMS peut s'exprimer sous forme d'une combinaison linéaire de la pente de la série de l'ICC par l'expression $R_t^X = aR_t^I + \bar{R}$. De même, la tendance de la série de l'IMS peut être exprimée sous forme d'une combinaison linéaire de la tendance pour la série de l'ICC par l'expression $L_t^X = aL_t^I + \bar{L} + \bar{R}t$. Notons que $\bar{R}$ et $\bar{L}$ sont des constantes qui sont calculées à partir des états estimés aux deux dernières périodes de la série.

La cointégration accroît la précision des estimations de la tendance et du signal de la série de l'ICC, permet de formuler des modèles plus parcimonieux, mais pourrait aussi être considérée comme un argument en vue de remplacer la série de l'ICC par celle de l'IMS, puisque les deux séries sont sous-tendues par une même tendance commune et représentent toutes deux cette tendance. Pour une discussion plus détaillée de la cointégration dans le contexte de la modélisation espace-état, consulter Koopman, Harvey, Shephard et Doornik (2009, sections 6.4 et 9.1).

3.3 Estimation des modèles de séries chronologiques structurels

Le moyen général d'analyser un modèle de séries chronologiques structurel consiste à l'exprimer dans la représentation dite espace-état et à appliquer le filtre de Kalman pour obtenir des estimations optimales pour les variables d'état (voir par exemple, Durbin et Koopman (2012)). Le logiciel pour l'analyse et l'estimation des modèles de séries chronologiques est développé en Ox en combinaison avec les sous-routines de SsfPack 3.0; voir Doornik (2009) et Koopman, Shephard et Doornik (2008).

Toutes les variables d'état sont non stationnaires et initialisées au moyen d'un prior diffus, c'est-à-dire que les espérances des états initiaux sont nulles et que la matrice de covariance initiale des états est diagonale avec de grands éléments diagonaux. Dans Ssfpack 3.0, une fonction de log-vraisemblance diffuse exacte s'obtient à l'aide de la procédure proposée par Koopman (1997). Les estimations du maximum de vraisemblance (MV) pour les hyperparamètres, c'est-à-dire les composantes de variance des processus stochastiques pour les variables d'état, sont obtenues avec une procédure d'optimisation numérique (algorithme de Broyden-Fletcher-Goldfarb-Shanno (BFGS), Doornik, 2009). Pour éviter d'obtenir des estimations de variance négatives, on estime les variances log-transformées. Le lecteur trouvera d'autres renseignements techniques sur l'analyse des modèles espace-état dans Harvey (1989) ou dans Durbin et Koopman (2012).

Sous l'hypothèse que les termes de perturbation suivent une loi normale, on peut appliquer le filtre de Kalman pour obtenir des estimations optimales des variables d'état, voir par exemple, Durbin et Koopman (2012). Le filtre de Kalman suppose que les termes de variance et de covariance sont connus d'avance et ces termes sont souvent appelés hyperparamètres. En pratique, ces hyperparamètres sont inconnus et, par conséquent, remplacés par l'estimation de leur MV. Les estimations pour les variables d'état pour la période t fondées sur l'information disponible jusqu'à la période t inclusivement sont appelées *estimations filtrées*.

Elles s'obtiennent à l'aide du filtre de Kalman où les estimations du MV des hyperparamètres sont fondées sur la série chronologique complète. Les estimations filtrées des vecteurs d'état antérieurs peuvent être mises à jour si de nouvelles données deviennent disponibles. Cette procédure, appelée lissage, donne des *estimations lissées* qui sont fondées sur la série chronologique complète.

Les erreurs-types des estimations obtenues avec le filtre de Kalman ne reflètent pas l'incertitude supplémentaire due à l'utilisation des estimations du MV pour les hyperparamètres inconnus. Donc, les estimations des erreurs-types sont trop optimistes.

4 Résultats

4.1 Modèle univarié de la série de l'ICC

L'analyse univariée est fondée sur le modèle (3.8) décrit à la section 3.1 et appliqué à la série de l'ICC obtenue de décembre 2000 à mars 2015. Le tableau 4.1 donne les estimations du MV pour les hyperparamètres du modèle.

Tableau 4.1
Estimations du maximum de vraisemblance des hyperparamètres du modèle univarié de l'ICC

Écart-type	Estimation du MV
Tendance (σ_η)	1,18
Composante saisonnière (σ_ω)	0,0025
Équation de mesure (σ_υ)	2,46

La moyenne des erreurs-types des estimations directes pour l'ICC est égale à 1,21. Il découle du tableau 4.1 que l'écart-type des termes de perturbation de l'équation de mesure est égale à 2,46. Cela illustre le fait que le bruit blanc de la population est plus grand que la variance des termes de perturbation de la mesure, tel qu'il est indiqué par le choix de la structure de variance pour (3.4) à la section 3.1.

Dans le graphique supérieur de la figure 4.1, la tendance lissée plus les valeurs aberrantes sont comparées aux estimations directes pour l'ICC. Dans le graphique inférieur de la figure 4.1, le signal lissé, défini comme étant la tendance plus les valeurs aberrantes plus la composante saisonnière, est comparé aux estimations directes pour l'ICC. Dans la série contenant la tendance lissée et les valeurs aberrantes, l'effet saisonnier, le bruit blanc du paramètre de population et l'erreur d'échantillonnage sont supprimés de la série originale. Il découle de la figure 4.1 que le modèle de séries chronologiques permet d'obtenir une estimation plus stable de l'ICC. La série de la tendance filtrée plus les valeurs aberrantes est comparée aux estimations lissées à la figure 4.2. Cette série filtrée est une approximation de ce que l'on obtiendrait dans la production des statistiques officielles si aucune révision n'était publiée. Il s'ensuit que, même dans ce cas, il est possible d'éliminer une part considérable de la variation à haute fréquence et des fluctuations saisonnières. Les deux figures montrent que le filtre de Kalman fournit des imputations lissées, mais aussi filtrées, plausibles pour l'observation manquante en octobre 2013.

La figure 4.3 donne la courbe saisonnière lissée de la série de l'ICC. Puisque les effets saisonniers sont presque invariants dans le temps, les effets sont présentés pour les 12 mois d'une année seulement. Il existe clairement des effets négatifs significatifs en octobre, novembre et décembre, et des effets positifs en janvier et en août. L'objectif de l'ICC est de mesurer la confiance à long terme des enquêtés, puisque toutes les questions font référence à leur situation financière et économique au cours des 12 derniers mois ou à leurs attentes pour les 12 mois à venir. Cependant, le patron saisonnier clair et significatif indique que les réponses des enquêtés sont manifestement dictées par des émotions à beaucoup plus court terme, qui sont, entre autres, sujettes à des fluctuations saisonnières.

À la figure 4.4, l'erreur-type des estimations directes pour l'ICC est comparée aux erreurs-types de la série de la tendance filtrée et lissée plus les valeurs aberrantes. Les pics de la courbe de l'erreur-type des estimations filtrées et lissées sont le résultat des variables de valeurs aberrantes et de l'observation manquante en 2013. Si une variable de valeur aberrante est activée à un certain point dans le temps, il faut estimer un nouveau coefficient de régression. Cela produit une incertitude supplémentaire dans les estimations du modèle, ce qui se manifeste par un pic soudain de l'erreur-type de la tendance filtrée et lissée. En 2013, une observation est manquante, ce qui donne aussi lieu à une incertitude supplémentaire, puisque le modèle espace-état produit une prédiction pour la valeur manquante.

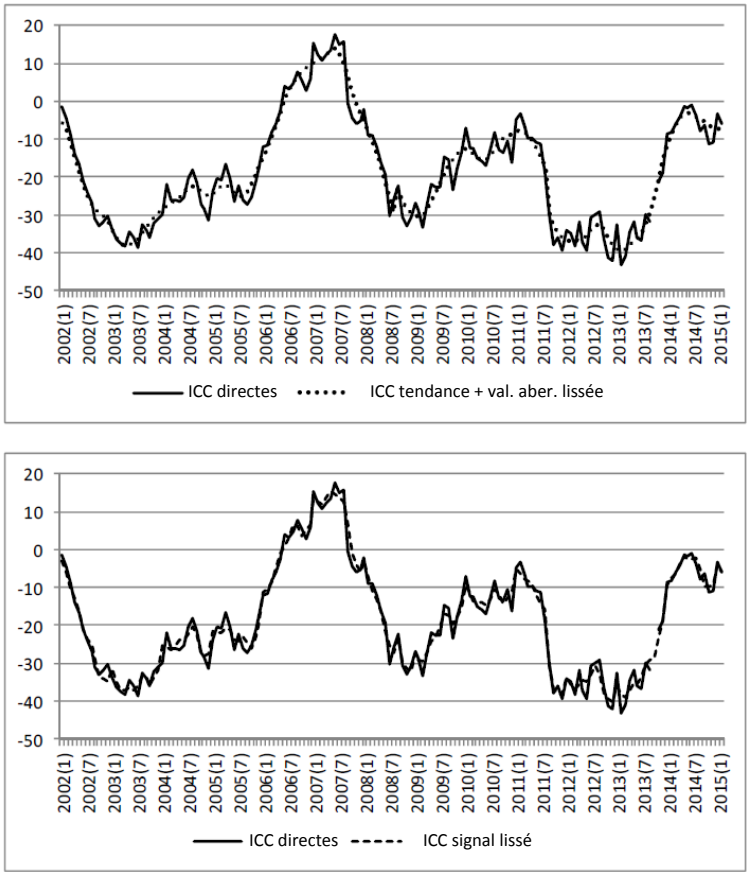

Figure 4.1 Tendance plus valeurs aberrantes lissée comparées aux estimations directes de l'ICC (graphique supérieur) et signal lissé (tendance plus valeurs aberrantes plus saisonnière) comparé aux estimations directes de l'ICC (graphique inférieur).

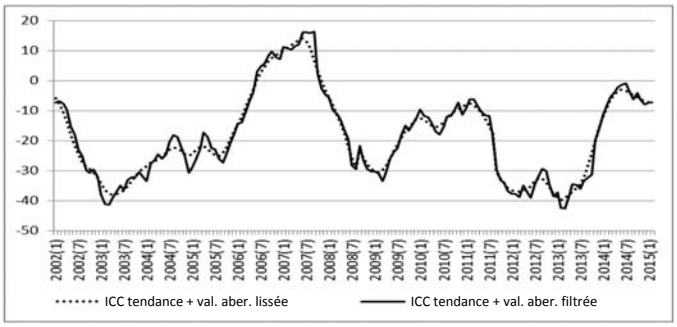

Figure 4.2 Tendance plus valeurs aberrantes filtrée comparées à la tendance plus valeurs aberrantes lissée pour l'ICC.

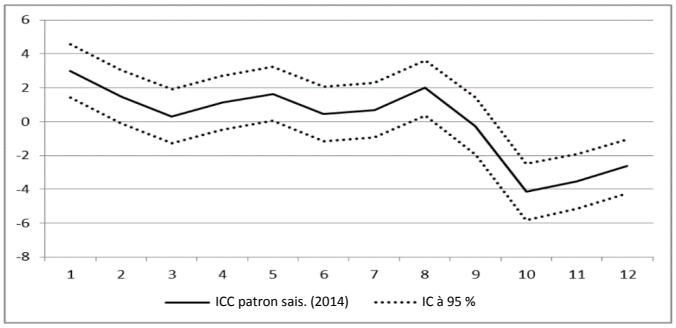

Figure 4.3 Patron saisonnier lissé de l'ICC pour 2014.

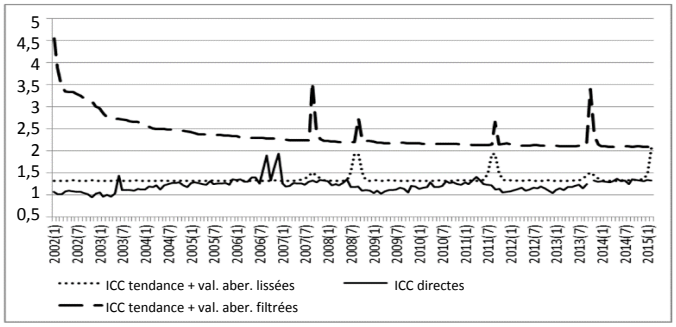

Figure 4.4 Erreur-type des tendance plus valeurs aberrantes lissées et filtrées comparée aux estimations directes de l'ICC.

Les erreurs-types des estimations lissées sont un peu plus grandes que celles des estimations directes. Les erreurs-types des estimations filtrées sont considérablement plus grandes que celles des estimations directes. Ce résultat est remarquable. Les estimations filtrées et lissées fournies par le modèle de séries chronologiques s'appuient sur un beaucoup plus grand ensemble de données puisque les données d'échantillon des périodes précédentes (dans le cas des estimations filtrées) ou la série de données complète (dans le cas des estimations lissées) sont utilisées pour obtenir une estimation optimale pour l'ICC mensuel. Par contre, les estimations directes sont fondées sur l'échantillon observé durant le mois en question seulement. La plupart des applications dans lesquelles les modèles de séries chronologiques structurels sont utilisés comme une forme d'estimation sur petits domaines donnent lieu à d'importantes réductions de

l'erreur-type comparativement aux estimations directes; voir, par exemple, van den Brakel et Krieg (2009, 2015) et Bollineni-Balabay, van den Brakel et Palm (2015, 2017).

La raison pour laquelle, dans la présente application, une approche de modélisation de séries chronologiques produit des erreurs-types pour les estimations filtrées et lissées issues du modèle de séries chronologiques plus grandes que les erreurs-types des estimations directes tient à une grande composante de bruit blanc dans la valeur de population réelle de l'ICC. Rappelons, comme il est mentionné à la section 3.1, que le terme de perturbation de (3.8) contient deux composantes, à savoir l'erreur d'échantillonnage et la variation de haute fréquence inexpliquée de la valeur de population réelle, exprimées par (3.4). Rappelons du tableau 4.1 que σ_{υ} est égale à 2,46 et est deux fois plus grande que la valeur moyenne des erreurs-types des estimations directes. Il s'agit d'un fort indice que la variance de la composante de bruit blanc dans la variable de population réelle est du même ordre que celle de l'erreur d'échantillonnage. L'estimateur direct de l'ICC obtenu à la section 2 traite l'ICC pour chaque mois particulier comme une variable fixe, mais inconnue. La variance de l'estimateur direct mesure uniquement l'incertitude, puisqu'un petit échantillon est observé au lieu de la population complète pour estimer l'ICC. Elle ne mesure pas la variation de haute fréquence de la valeur de population au cours du temps. Cela explique pourquoi l'approche de modélisation de séries chronologiques n'aboutit pas à une réduction de l'erreur-type de l'ICC estimé.

Même si le gain de précision des estimations de niveau obtenues au moyen du modèle de séries chronologiques est limité, les estimations de la tendance sont plus stables, comme l'illustrent les figures 4.1 et 4.2. Un modèle de séries chronologiques demeurera donc utile pour isoler une tendance à long terme plus stable de la variation de haute fréquence dans le paramètre de population et de l'erreur d'échantillonnage. Comme les variables d'état de la composante tendance des périodes subséquentes présenteront une forte corrélation positive, on peut s'attendre à ce que l'approche de modélisation de séries chronologiques offre davantage de gain si l'on se concentre sur les mouvements mois-à-mois; consulter, par exemple, Harvey et Chung (2000). Les estimations filtrées du mouvement mois-à-mois de l'ICC sont définies comme étant

$$\Delta_{t|t} = L_{t|t} - L_{t-1|t} + \beta^{08}_{t|t}\,\delta^{08}_{t} - \beta^{08}_{t-1|t}\,\delta^{08}_{t-1} + \beta^{11}_{t|t}\,\delta^{11}_{t} - \beta^{11}_{t-1|t}\,\delta^{11}_{t-1}, \tag{4.1}$$

où la notation $\Theta_{t|t'}$ désigne l'estimation de la variable d'état Θ pour la période t sachant les données observées jusqu'à la période t'. La valeur aberrante ponctuelle en 2007(9) est, évidemment, supprimée du signal. Qui plus est, les coefficients de régression sont invariants dans le temps. Par conséquent, $\beta^{x}_{t|t} = \beta^{x}_{t-1|t}$ pour $x =$ 08 et 11. Puisque $t =$ 2008(9) et $t =$ 2011(9) sont les mois où δ^{08}_{t} et δ^{11}_{t} changent de valeur, l'expression (4.1) peut être simplifiée en

$$\Delta_{t|t} = L_{t|t} - L_{t-1|t} + \beta^{08}_{t|t}\,\tilde{\delta}^{08}_{t} + \beta^{11}_{t|t}\,\tilde{\delta}^{11}_{t}, \tag{4.2}$$

avec $\tilde{\delta}^{08}_{t} = 1$ si $t =$ 2008(9) et $\tilde{\delta}^{08}_{t} = 0$ pour toutes les autres périodes et $\tilde{\delta}^{11}_{t} = 1$ si $t =$ 2011(9) et $\tilde{\delta}^{11}_{t} = 0$ pour toutes les autres périodes. Les estimations lissées pour le mouvement mois-à-mois de l'ICC sont définies comme étant

$$\Delta_{t|T} = L_{t|T} - L_{t-1|T} + \beta^{08}_{t|T}\,\tilde{\delta}^{08}_{t} + \beta^{11}_{t|T}\,\tilde{\delta}^{11}_{t}. \tag{4.3}$$

Pour comparer les mouvements mois-à-mois fondés sur (4.2) et (4.3) aux estimations directes, les effets saisonniers lissés en (3.8) sont soustraits des estimations directes. Les erreurs-types des estimations directes ne sont pas corrigées pour cet ajustement.

La figure 4.5 compare les estimations directes du mouvement mois-à-mois aux estimations lissées (graphique supérieur) et aux estimations filtrées (graphique du milieu) obtenues au moyen du modèle de séries chronologiques. Le graphique inférieur compare les erreurs-types des estimations lissées, filtrées et directes. Les estimations filtrées, et en particulier les estimations lissées du mouvement mois-à-mois, ont un patron plus stable que les estimations directes, ce que reflètent également les erreurs-types. Les fortes corrélations positives des états de la composante tendance entre périodes subséquentes produisent des erreurs-types pour les estimations filtrées et lissées du mouvement mois-à-mois qui sont clairement plus petites que pour l'estimateur direct. Font exception les deux périodes où un changement de niveau est nécessaire. L'introduction d'un changement de niveau accroît le niveau d'incertitude pendant une courte période.

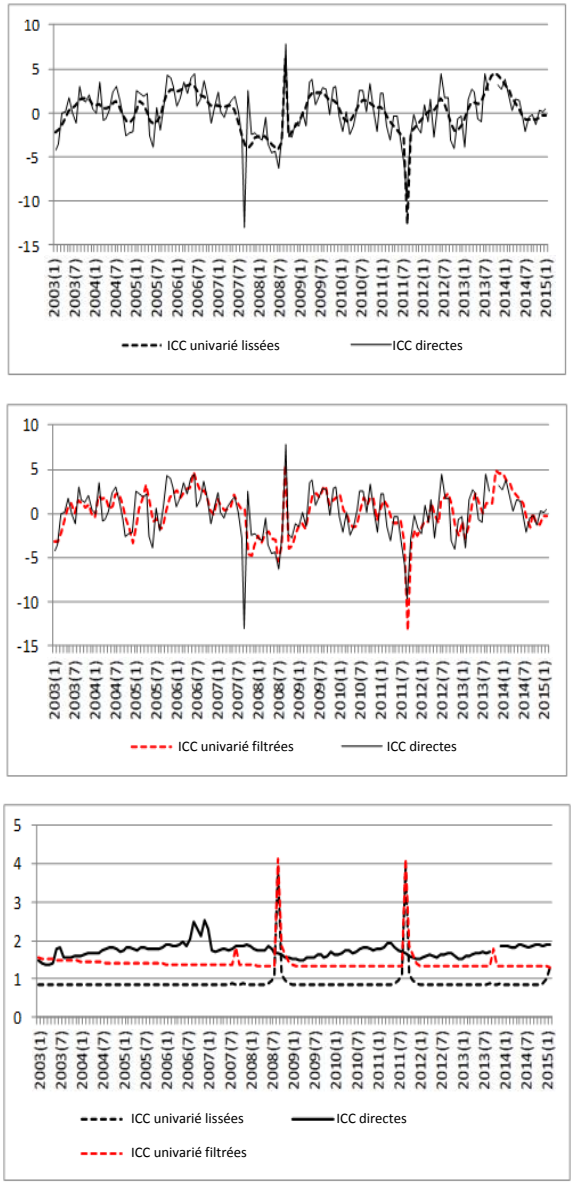

Figure 4.5 Comparaison des estimations du modèle univarié et des estimations directes du mouvement mois-à-mois. Graphique supérieur : estimations lissées, graphique du milieu : estimations filtrées, graphique inférieur : erreurs-types.

La réduction de l'erreur-type est mesurée par l'écart moyen relatif des erreurs-types (EMRET) et, pour les estimations filtrées, est définie comme étant $\text{EMRET} = 100/(T-t')*\sum_{t=t'}^{T}\left[\text{et}\left(\hat{\Delta}_t\right)-\text{et}\left(\Delta_{t|t}\right)\right]\Big/\text{et}\left(\hat{\Delta}_t\right)$, avec et $\left(\hat{\Delta}_t\right)$ l'erreur-type pour l'estimation directe du mouvement mois-à-mois. Pour les estimations lissées, EMRET s'obtient en remplaçant et $\left(\Delta_{t|t}\right)$ par et $\left(\Delta_{t|T}\right)$. Durant la période observée à partir de 2003(1), l'EMRET pour les estimations lissées égale 47 % et pour les estimations filtrées, 17 %.

4.2 Modèle bivarié pour les séries de l'ICC et de l'IMS

À la présente section, nous appliquons le modèle bivarié (3.9) proposé à la section 3.2 aux séries de l'ICC et de l'IMS, qui sont disponibles pour la période allant de juin 2010 à mars 2015. Notons que les composantes de série chronologique pour l'ICC sont réestimées en utilisant la série plus courte. Le tableau 4.2 donne les estimations du maximum de vraisemblance des hyperparamètres. Le modèle décèle une forte corrélation positive de l'ordre de 0,92 entre les perturbations de pente de l'ICC et de l'IMS. Cependant, rien n'indique que les deux tendances sont cointégrées et partagent une tendance commune. Un test du rapport de vraisemblance est appliqué pour déterminer le degré de signification de la corrélation entre les perturbations de pente dans le modèle bivarié. Quand le paramètre de corrélation est fixé à zéro, la log-vraisemblance diminue, passant de -229,9 à -233,9. La valeur p du test du rapport de vraisemblance correspondant égale 0,0047, ce qui indique que la corrélation entre les tendances des deux séries diffère de manière nettement significative de zéro et ne doit pas être supprimée du modèle bivarié. Quand le paramètre de corrélation est fixé à un (en choisissant d_2 dans (3.10) égal à zéro), la log-vraisemblance passe de -229,9 à -242,1. La valeur p du test du rapport de vraisemblance correspondant avec un degré de liberté égale zéro, ce qui indique que les tendances ne sont pas cointégrées.

Tableau 4.2
Estimations du maximum de vraisemblance des hyperparamètres du modèle bivarié de l'ICC et de l'IMS

Écart-type	Estimation du MV
Tendance de l'ICC $(\sigma_{\eta I})$	1,25
Composante saisonnière de l'ICC (σ_{ω})	7,5E-6
Tendance de l'IMS $(\sigma_{\eta X})$	0,25
Équation de mesure de l'ICC $(\sigma_{\upsilon I})$	2,68
Équation de mesure de l'IMS $(\sigma_{\upsilon X})$	0,84
Corrélation des tendances de l'ICC et de l'IMS (ρ_{η})	0,92

À la figure 4.6, nous comparons les estimations lissées de la pente de l'ICC (axe des x) et de l'IMS (axe des y) sous le modèle sans corrélation, le modèle avec une estimation du MV pour la corrélation $\left(\rho_{\eta}=0{,}92\right)$ et le modèle à tendance commune avec $\rho_{\eta}=1{,}0$. Le modèle avec pentes non corrélées révèle une corrélation nettement positive entre les pentes si les deux séries sont estimées indépendamment (graphique de gauche de la figure 4.6). Cette corrélation apparaît dans le modèle qui tient compte de la

corrélation (graphique du milieu de la figure 4.6). Cependant, il existe manifestement un écart entre les pentes des deux séries, ce que l'on peut constater en comparant le graphique du modèle avec une corrélation estimée par le MV (graphique du milieu de la figure 4.6) au graphique du modèle avec un facteur commun (graphique de droite de la figure 4.6).

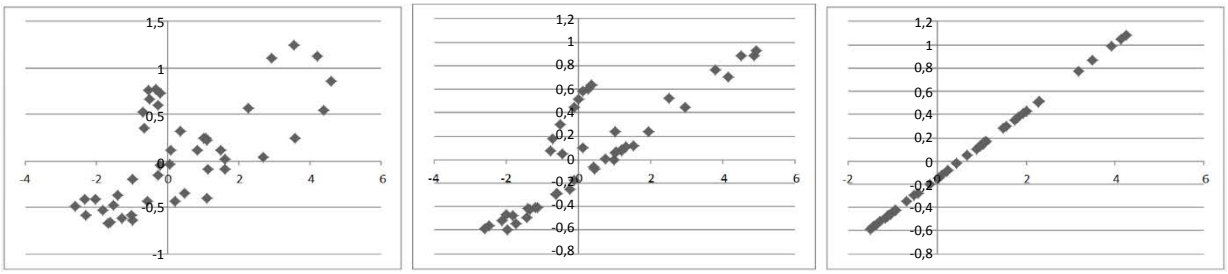

Figure 4.6 Nuages de points des pentes lissées de l'ICC (axe des x) et de l'IMS (axe des y) pour un modèle sans corrélation (graphique de gauche), avec corrélation estimée par le MV (graphique du milieu) et corrélation fixée à un (graphique de droite).

La figure 4.7 donne une comparaison de la série observée de l'IMS avec la tendance lissée obtenue sous le modèle bivarié. La figure 4.8 donne une comparaison des estimations directes de la série de l'ICC avec la tendance lissée plus la valeur aberrante sous le modèle univarié et le modèle bivarié. Comme le montre la figure 4.8, le niveau et l'évolution des estimations lissées de la série de l'ICC sont presque identiques sous les modèles univarié et bivarié.

À la figure 4.9, nous comparons les erreurs-types des estimations directes de la série de l'ICC avec la tendance lissée plus la valeur aberrante sous le modèle univarié et le modèle bivarié. Pour que la comparaison soit juste, les résultats pour le modèle univarié et le modèle bivarié sont fondés sur des séries de même longueur. Pour cela, le modèle univarié est réestimé en se basant sur la série allant de juin 2010 à mars 2015. La figure 4.9 révèle que l'erreur-type sous le modèle bivarié est légèrement plus petite que l'erreur-type sous le modèle univarié si les deux modèles sont appliqués à des séries de longueur égale, comme prévu étant donné la forte corrélation positive significative entre les termes de perturbation des tendances des deux séries. En revanche, si le modèle univarié est appliqué à la série disponible à partir de décembre 2000, les erreurs-types pour les estimations lissées sous le modèle univarié sont légèrement plus faibles que sous le modèle bivarié comme le montre la figure 4.10.

Pour résumer, il apparaît que le modèle bivarié décèle une forte corrélation entre les séries de l'ICC et de l'IMS. L'utilisation de la série de l'IMS comme série auxiliaire améliore légèrement la précision des estimations fondées sur un modèle pour l'ICC. Puisque la série de l'ICC est neuf années plus longue que celle de l'IMS, dans le modèle univarié, la plus grande précision obtenue avec la série auxiliaire est compensée par l'information supplémentaire disponible avant 2010 dans la série de l'ICC.

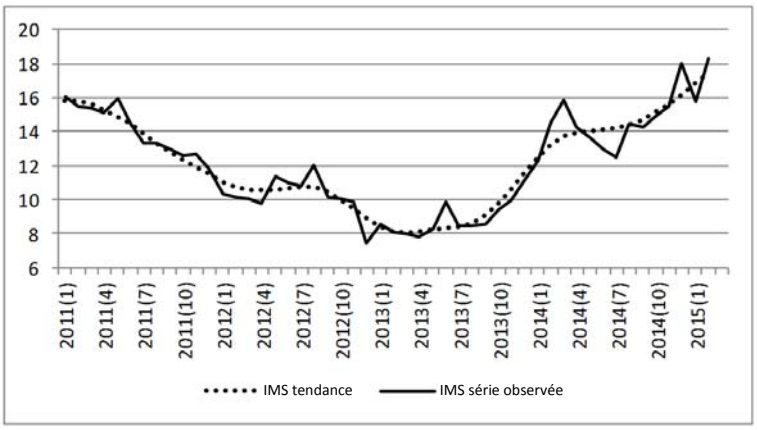

Figure 4.7 Série observée et tendance lissée de l'IMS.

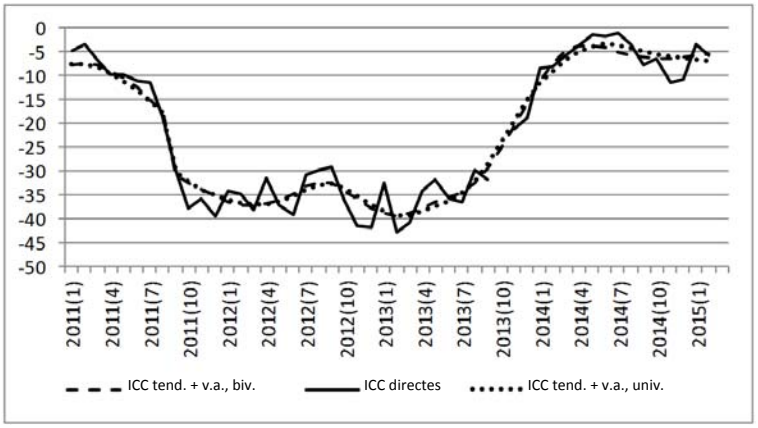

Figure 4.8 Comparaison des estimations directes et de la tendance lissée plus la valeur aberrante pour l'ICC sous les modèles bivarié et univarié de l'ICC.

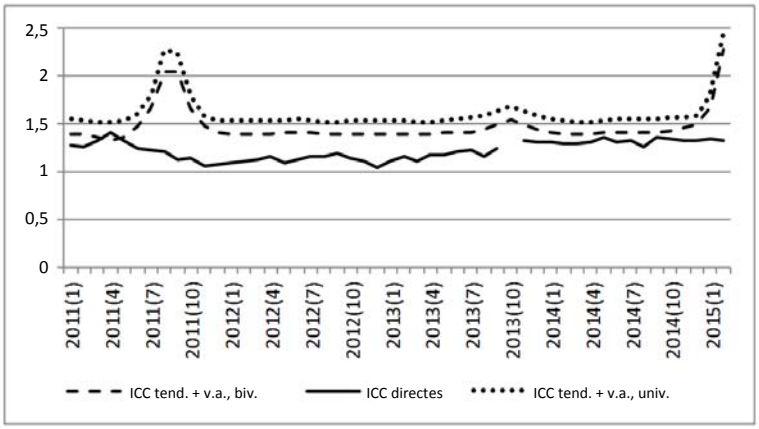

Figure 4.9 Comparaison des erreurs-types des estimations directes de l'ICC et de la tendance lissée plus valeur aberrante sous les modèles bivarié et univarié pour l'ICC si les deux modèles sont appliqués à des séries de même longueur (juin 2010 à mars 2015).

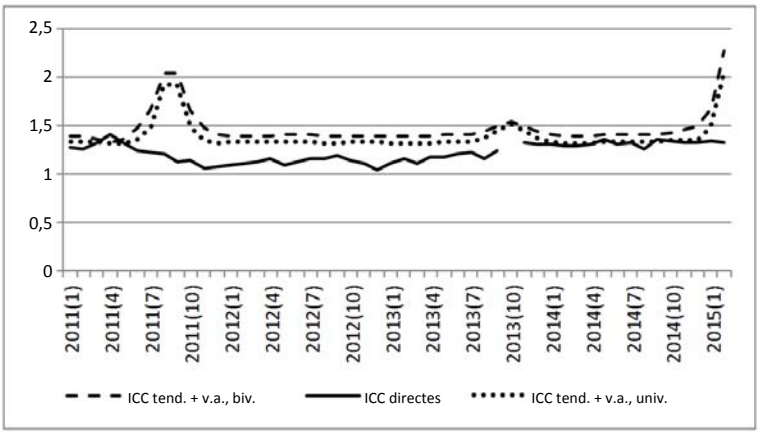

Figure 4.10 Comparaison des erreurs-types des estimations directes de l'ICC et de la tendance lissée plus valeur aberrante sous les modèles bivarié et univarié pour l'ICC si le modèle univarié est appliqué à la série complète de l'ICC (décembre 2000).

Le graphique supérieur de la figure 4.11 compare les estimations directes du mouvement mois-à-mois aux estimations lissées obtenues au moyen des modèles univarié et bivarié de séries chronologiques (tous deux fondés sur la série observée à partir de juin 2010). Le graphique inférieur donne une comparaison des erreurs-types de ces estimations. Durant la période observée à partir de 2011(1), l'EMRET pour les estimations lissées sous le modèle univarié vaut 39 % et sous le modèle bivarié, 43 %. L'EMRET pour les estimations filtrées sous le modèle univarié vaut 7 % et sous le modèle bivarié, 14 %. Comme dans le cas du modèle univarié, l'approche de modélisation de séries chronologiques produit des estimations plus stables et plus précises du mouvement mois-à-mois. L'utilisation de la série de l'IMS améliore légèrement la précision des mouvements mois-à-mois comparativement au modèle univarié.

Une fois que l'estimation directe de l'ICC pour le mois t devient disponible, la valeur ajoutée de la série de l'IMS est limitée à améliorer l'estimation de la série chronologique pour l'ICC pour le mois t. Toutefois, un inconvénient des sondages est que les données sont généralement moins à jour que celles provenant des médias sociaux. La valeur ajoutée de l'IMS devient plus évidente lorsque l'on profite de la plus grande fréquence de cette série pour produire des prédictions précoces ou prédictions immédiates pour l'ICC au moyen du modèle espace-état bivarié. Si une première prédiction immédiate pour l'ICC est nécessaire durant le mois t ou directement à la fin de ce mois, le modèle univarié produit seulement une prévision une étape à l'avance. Par contre, dès que les résultats pour la série de l'IMS sont disponibles durant le mois t ou à la fin du mois t, le modèle bivarié exploite la forte corrélation entre les séries pour faire une prédiction plus précise pour l'ICC, déjà avant que l'estimation directe pour le mois t devienne disponible.

Afin d'illustrer la valeur ajoutée qu'offre l'IMS dans une procédure de prédiction immédiate pour l'ICC, au graphique supérieur de la figure 4.12, nous comparons les prévisions une étape à l'avance pour la tendance plus valeur aberrante de la série de l'ICC obtenue avec le modèle univarié à l'estimation obtenue avec le modèle bivarié si l'IMS pour le mois t est disponible, mais que l'estimation directe de l'ICC manque encore. Les estimations lissées pour la tendance plus valeur aberrante de l'ICC obtenues au moyen du modèle univarié sont incluses comme référence. Dans le graphique inférieur de la figure, nous comparons les erreurs-types de ces trois estimations.

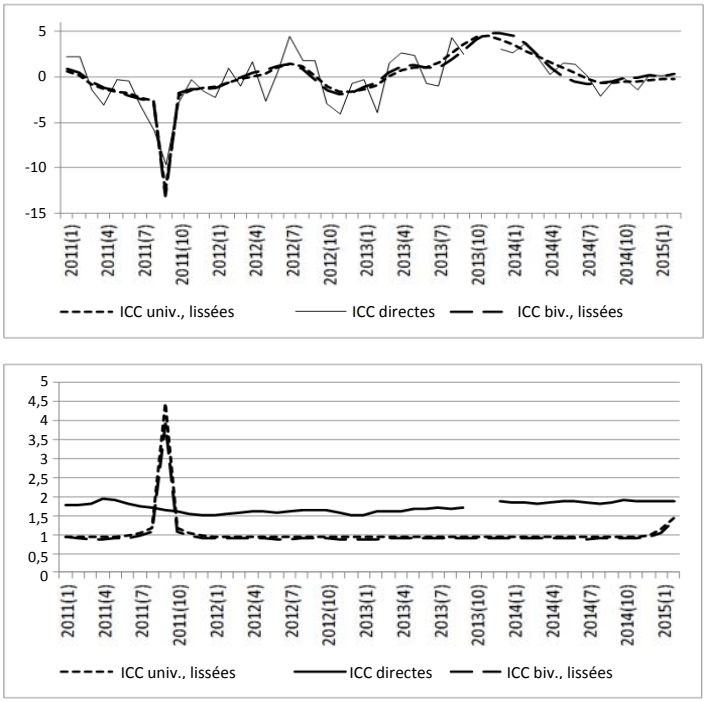

Figure 4.11 Comparaison des estimations sous le modèle bivarié, sous le modèle univarié et directes des mouvements mois-à-mois. Graphique supérieur : estimations lissées, graphique inférieur : erreurs-types.

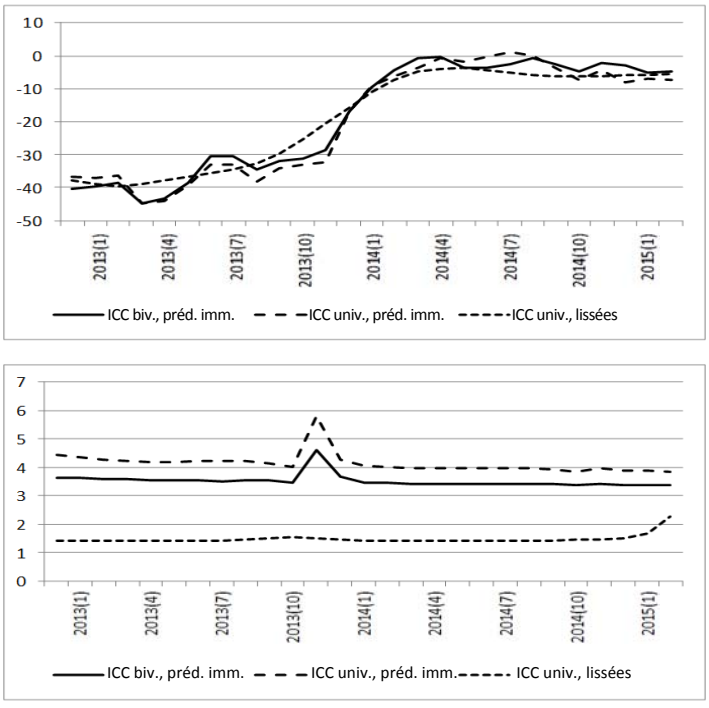

Figure 4.12 Comparaison des estimations pour la tendance plus valeur aberrante de la série de l'ICC; modèle univarié avec prévisions une étape à l'avance (ICC univ., prédiction immédiate), modèle bivarié si l'IMS pour le mois *t* est disponible, mais que l'estimation directe de l'ICC manque (ICC biv., prédiction immédiate) et estimations lissées avec le modèle univarié (ICC univ., lissées). Graphique supérieur : comparaison des estimations ponctuelles. Graphique inférieur : comparaison des erreurs-types.

Lorsque les estimations lissées issues du modèle univarié sont prises comme référence, nous utilisons comme mesure de la grandeur de la révision l'écart moyen absolu relatif (EMAR) entre les prédictions immédiates et les estimations lissées, lequel est défini par $\text{EMAR} = 100/(T-t') * \sum_{t=t'}^{T} \left|\theta_{t|T} - \theta_{t|t-1}\right| \Big/ \left|\theta_{t|T}\right|$, où $\theta_t = L_t + \beta^{11}\delta_t^{11}$ désigne la tendance plus valeurs aberrantes de la série de l'ICC. En se basant sur les mois observés à partir de $t =$ 2013(1), l'EMAR pour les prédictions immédiates obtenues au moyen du modèle univarié égale 35 % et au moyen du modèle bivarié, 31 %. Cela montre que les révisions sont un peu plus petites, et donc plus stables si l'on utilise les prédictions immédiates pour l'ICC obtenues au moyen du modèle bivarié. La différence de précision entre les prédictions immédiates obtenues avec le modèle univarié et le modèle bivarié est mesurée par l'EMRET qui est défini ici comme étant $\text{EMRET} = 100/(T-t') * \sum_{t=t'}^{T} \left[\text{et}\left(\theta_{t|t-1}^{\text{uni}}\right) - \text{et}\left(\theta_{t|t-1}^{\text{biv}}\right)\right] \Big/ \text{et}\left(\theta_{t|t-1}^{\text{biv}}\right)$. En se basant sur les mois observés à partir de $t =$ 2013(1), la différence de précision entre les deux prédictions immédiates selon cet EMRET est égale à 17 %. La figure 4.12, ainsi que l'EMAR et l'EMRET illustrent que l'IMS améliore la stabilité et la précision des prédictions immédiates pour l'ICC.

5 Discussion

Pendant des décennies, les instituts nationaux de statistique se sont appuyés sur l'échantillonnage probabiliste pour produire les statistiques officielles. Cette approche est fondée sur une théorie éprouvée en vue de faire des inférences statistiques valides pour de grandes populations cibles finies en partant d'échantillons aléatoires relativement petits. Au cours des dernières décennies, de plus en plus de sources de données de rechange, comme les données administratives et les mégadonnées, sont devenues disponibles et la question qui se pose est celle de savoir comment utiliser ces sources de données pour produire les statistiques officielles. Un important problème concerne la généralisation des résultats obtenus au moyen de ces sources à une population cible finie. Puisque le processus de génération des données est habituellement inconnu, la façon de faire des inférences valides au moyen de ces sources de données n'est pas évidente.

Le présent article traite de la façon d'utiliser les sources de données administratives et de mégadonnées pour produire des statistiques officielles. L'approche la plus extrême consiste à remplacer les données d'enquête par des données provenant de sources de rechange reliées, en courant le risque d'introduire, par exemple, un biais de sélection. Puisque la plupart des enquêtes sont réalisées de manière répétée, nous proposons une approche de modélisation de séries chronologiques pour déterminer dans quelle mesure les sources de données de rechange reliées reflètent une même évolution que la série obtenue au moyen d'une enquête répétée. Un modèle espace-état multivarié permet de modéliser la corrélation entre les composantes sous-jacentes non observées des deux séries. Le cas où les composantes du modèle de séries chronologiques sont cointégrées constitue un fort indice que les deux sources de données ont pour moteur le même facteur sous-jacent. Le cas échéant, cela permettrait d'arguer qu'une source de rechange peut remplacer les enquêtes existantes, puisqu'elles reflètent la même évolution d'un processus, généralement à un niveau différent.

La théorie qui sous-tend l'échantillonnage probabiliste pour l'inférence sur des populations finies est plus solide que l'application du concept de cointégration. Les séries obtenues à partir des médias sociaux ou de Google Trends sont choisies de manière à maximiser la corrélation avec la série issue de l'enquête par sondage et ne mesurent pas nécessairement le même concept que l'enquête. Rien ne garantit que cette corrélation est basée sur une causalité réelle ni que la corrélation persistera dans l'avenir. En revanche, la théorie de l'échantillonnage repose sur une théorie mathématique rigoureuse montrant qu'une stratégie

d'échantillonnage correcte, c'est-à-dire la combinaison appropriée d'un échantillon probabiliste et d'un estimateur approximativement sans biais sous le plan, aboutit à des inférences statistiques valides pour les populations cibles choisies.

Même dans le cas de séries cointégrées, une évaluation approfondie du modèle, par exemple une forme de validation croisée, sera nécessaire pour avoir la certitude que la source de données de rechange est un remplacement valide. Consulter aussi, dans ce contexte, Eichler (2013) pour une discussion de l'utilisation de la causalité de Granger pour l'inférence causale dans le cas de séries chronologiques multiples. Au lieu de remplacer une enquête périodique par des sources de données reliées, ces dernières peuvent être utilisées comme série auxiliaire dans une approche de modélisation de séries chronologiques multivariée en vue d'améliorer la précision des estimations directes ou des estimations du mouvement période-à-période des estimations directes obtenues au moyen d'une enquête périodique. Un autre avantage important des sources de mégadonnées consiste à profiter de la plus grande fréquence de ces sources de données pour faire des prédictions précoces ou des prédictions immédiates plus précises quand, en temps réel, les estimations issues de l'enquête ne sont pas encore disponibles, mais que la covariable l'est déjà. Le modèle de séries chronologiques appliqué dans le présent article, proposé pour la première fois par Harvey et Chung (2000), représente une approche générique pour une procédure d'estimation fondée sur un modèle pour des enquêtes périodiques. Naturellement, les enquêtes par sondage posent aussi des problèmes. À titre d'exemple, les taux de réponse continuellement à la baisse et les modes de collecte des données ne permettant pas d'atteindre la population cible produisent aussi un biais de sélection. Dans cette situation, la cointégration avec une série reliée dérivée des médias sociaux pourrait être un signe qu'il existe des similarités entre le biais de sélection dans les sources de mégadonnées non probabilistes et le biais de couverture et de sélection de la non-réponse dans un sondage, comme l'ont souligné Baker et coll. (2013).

Dans l'application à l'ICC, l'approche de modélisation de séries chronologiques ne réduit pas la variance de l'estimateur direct quand elle est utilisée pour produire des estimations de niveau. Cela tient au fait que l'erreur-type du modèle de séries chronologiques reflète l'erreur d'échantillonnage et le bruit blanc du paramètre de population. L'erreur-type de l'estimateur direct reflète uniquement l'erreur d'échantillonnage. Dans le cas de l'ICC, la composante de variance du bruit blanc du paramètre de population est aussi grande que la variance de l'erreur d'échantillonnage. L'approche du modèle espace-état est néanmoins utile pour produire les chiffres officiels de l'ICC, puisqu'elle filtre une tendance plus stable de l'opinion des personnes interrogées au sujet du climat économique à partir de la série observée d'estimations directes. Cependant, la situation change si le modèle de séries chronologiques est utilisé pour estimer les mouvements mois-à-mois. Les estimations de la tendance stable résultent d'une forte corrélation positive entre les estimations de périodes subséquentes. Par conséquent, les erreurs-types du mouvement mois-à-mois obtenues au moyen du modèle de séries chronologiques sont nettement plus petites que celles des estimations directes. Les erreurs-types des mouvements mois-à-mois lissés sont environ 47 % plus petites que celles des estimations directes. Les erreurs-types des estimations filtrées sont environ 17 % plus petites que celles des estimations directes.

L'utilisation de l'IMS comme série auxiliaire dans un modèle espace-état bivarié réduit légèrement l'erreur-type des estimations modélisées de l'ICC. Cependant, puisque la série de l'IMS disponible est relativement courte, la réduction obtenue au moyen de cette série auxiliaire ne surpasse pas la perte de l'information de la série de l'ICC qui a été observée durant la période avant que l'IMS soit disponible. Toutefois, puisque les deux séries présentent une même évolution et que les données des médias sociaux

sont disponibles rapidement, l'IMS s'est avéré utile comme série auxiliaire dans le modèle bivarié pour produire des prédictions immédiates plus fiables pour l'ICC en temps réel, au moment où l'IMS devient disponible tandis que l'ICC ne l'est pas encore. Dans cette application, l'IMS réduit d'environ 17 % les erreurs-types de l'ICC dans une procédure de prédiction immédiate.

D'aucuns se demanderont si, dans sa mise en œuvre actuelle, l'IMS mesure le même concept que celui qu'essaie de mesurer l'ICC et comment on pourrait tirer parti de toutes les possibilités qu'offrent les données des médias sociaux et d'autres sources de mégadonnées pour obtenir de meilleures mesures de la confiance des consommateurs que celles produites par les ICC et IMS actuels. Au lieu de construire un indice basé sur les médias sociaux en prenant la différence entre les messages jugés positifs et négatifs, on pourrait construire un IMS en examinant les concepts qui sous-tendent les questions utilisées pour l'ICC. Par exemple, si l'on mesure la confiance des consommateurs par le nombre d'achats de biens coûteux au cours des 12 derniers mois ou par la tendance des ménages à acheter des biens coûteux, les indices fondés sur les médias sociaux devraient mesurer la recherche de ce genre de biens sur Internet (automobiles, maisons, produits blancs, etc.), ainsi que les achats réels de ces biens durant les mois précédents. Le grand avantage de cette approche est qu'elle permet de mesurer directement le comportement actuel des ménages, alors qu'une enquête le mesure indirectement, ce qui induit une plus grande erreur de mesure. Cela pourrait aboutir à des séries cointégrées qui mesurent des concepts similaires et améliorent davantage ou même remplacent l'ICC.

Remerciements

Les auteurs sont reconnaissants au rédacteur associé et aux examinateurs pour leur lecture attentive d'une version antérieure de cet article et pour leurs commentaires constructifs ayant contribué à améliorer significativement le contenu de l'article. Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas forcément les politiques de *Statistics Netherlands*.

Annexe A

Diagnostics du modèle

Tableau A.1
Modèle univarié (3.8) pour l'ICC 172-24 observations

Diagnostic	Valeur	Valeur *p*	IC à 95 %	
			Inf.	Sup.
Log-vraisemblance	-464			
Moyenne des innovations standardisées	0,0152			
Variance des innovations standardisées	1,0851			
Asymétrie des innovations standardisées	0,0276			
Aplatissement des innovations standardisées	2,8901			
Test de Bowman-Shenton[1] sur la normalité des innovations standardisées	0,0926	0,955		
Test de Ljung-Box[2] sur la corrélation sériale des innovations standardisées	24,108	0,287		
Test de Durban-Watson[3] sur la corrélation sériale des innovations standardisées $(T = 148)$	2,082		1,68	2,32
Test F[4] sur l'hétéroscédasticité des innovations standardisées $(df_{\text{num}} = df_{\text{denom}} = 60)$	0,913		0,60	1,67

1) Statistique de Bowman-Shenton : loi du χ^2_2.
2) Statistique du test de Ljung-Box pour la corrélation sériale dans les 24 premiers décalages temporels : loi du χ^2_{21}.
3) Statistique du test de Durban-Watson approximée par $N(2, 4/T)$.
4) Statistique F : loi de $F^{df_{\text{num}}}_{df_{\text{denom}}}$.

Tableau A.2
Modèle bivarié (3.9) pour l'ICC 57-24 observations

Diagnostic	Valeur	Valeur *p*	IC à 95 %	
			Inf.	Sup.
Log-vraisemblance	-230			
Moyenne des innovations standardisées	-0,0872			
Variance des innovations standardisées	0,9777			
Asymétrie des innovations standardisées	0,0982			
Aplatissement des innovations standardisées	2,545			
Test de Bowman-Shenton[1] sur la normalité des innovations standardisées	0,3382	0,844		
Test de Ljung-Box[2] sur la corrélation sériale des innovations standardisées	18,060	0,645		
Test de Durban-Watson[3] sur la corrélation sériale des innovations standardisées ($T = 33$)	2,133		1,32	2,68
Test F[4] sur l'hétéroscédasticité des innovations standardisées ($df_{\text{num}} = df_{\text{denom}} = 15$)	0,783		0,35	2,86

Tableau A.3
Modèle bivarié (3.9) pour l'IMS 57-12 observations

Diagnostic	Valeur	Valeur *p*	IC à 95 %	
			Inf.	Sup.
Log-vraisemblance	-230			
Moyenne des innovations standardisées	0,0954			
Variance des innovations standardisées	1,0437			
Asymétrie des innovations standardisées	-0,1311			
Aplatissement des innovations standardisées	2,5331			
Test de Bowman-Shenton[1] sur la normalité des innovations standardisées	0,5377	0,764		
Test de Ljung-Box[2] sur la corrélation sériale des innovations standardisées	24,208	0,283		
Test de Durban-Watson[3] sur la corrélation sériale des innovations standardisées ($T = 45$)	2,028		1,42	2,58
Test F[4] sur l'hétéroscédasticité des innovations standardisées ($df_{\text{num}} = df_{\text{denom}} = 20$)	0,329		0,41	2,46

1) Statistique de Bowman-Shenton : loi du χ^2_2.
2) Statistique du test de Ljung-Box pour la corrélation sériale dans les 24 premiers décalages temporels : loi du χ^2_{21}.
3) Statistique du test de Durban-Watson approximée par $N(2, 4/T)$.
4) Statistique F : loi de $F^{df_{\text{num}}}_{df_{\text{denom}}}$.

Bibliographie

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. et Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143, publié pour la première fois en ligne le 26 septembre 2013, doi:10.1093/jssam/smt008.

Bell, W.R. (2005). Some considerations of seasonal adjustment variances. Census Bureau. Article accessible à l'adresse https://www.census.gov/ts/papers/jsm2005wrb.pdf.

Bell, W.R., et Hillmer, S.C. (1990). Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques. *Techniques d'enquête*, 16, 2, 205-227. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14535-fra.pdf.

Binder, D.A., et Dick, J.P. (1989). Enquêtes répétées – Modélisation et estimation. *Techniques d'enquête*, 15, 1, 31-48. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1989001/article/14579-fra.pdf.

Binder, D.A., et Dick, J.P. (1990). Méthode pour l'analyse des modèles ARMMI. *Techniques d'enquête*, 16, 2, 251-265. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14533-fra.pdf.

Blight, B.J.N., et Scott, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 35, 61-66.

Blumenstock, J., Cadamuro, G. et On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350, 1073-1076.

Bollineni-Balabay, O., van den Brakel, J.A. et Palm, F. (2015). Multivariate state-space approach to variance reduction in series with level and variance breaks due to sampling redesigns. Accepté pour publication dans *Journal of the Royal Statistical Society, Series A*.

Bollineni-Balabay, O., van den Brakel, J.A. et Palm, F. (2017). La modélisation espace-état appliquée aux séries chronologiques de l'Enquête sur la population active des Pays-Bas : sélection de modèles et estimation de l'erreur quadratique moyenne. *Techniques d'enquête*, 43, 1, 47-75. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14819-fra.pdf.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l'Institut International de Statistique*, 22, Supplément au livre 1, 6-62.

Buelens, B., Burger, J. et van den Brakel, J.A. (2015). Predictive inference for non-probability samples: A simulation study. Document de discussion 2015-13, Statistics Netherlands, Heerlen.

Cochran, W. (1977). *Sampling Theory*. New York: John Wiley & Sons, Inc.

Daas, P., et Puts, M. (2014a). Big data as a source of statistical information. *The Survey Statistician*, 69, 22-31.

Daas, P., et Puts, M. (2014b). Social media sentiment and consumer confidence. European Central Bank Statistics paper series No. 5, Frankfurt Allemagne.

Doornik, J.A. (2009). *An Object-oriented Matrix Programming Language Ox 6*. Londres: Timberlake Consultants Press.

Durbin, J., et Koopman, S.J. (2012). *Time Series Analysis by State Space Methods, Second Edition*. Oxford: Oxford University Press.

Eichler, M. (2013). Causal inference with multiple time series: Principles and problems. *Philosophical transactions of the Royal Statistical Society A*, 371, édition 1997.

Feder, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Statistica Neerlandica*, 55, 182-199.

Hansen, M.H., et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14, 333-362.

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A.C., et Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A*, 163, 303-339.

Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, 1630-1638.

Koopman, S.J., Harvey, A., Shephard, N. et Doornik, J.A. (2009). *STAMP 8.2*, Londres: Timberlake Consultants Press.

Koopman, S.J., Shephard, N. et Doornik, J.A. (2008). *SsfPack 3.0: Statistical Algorithms for Models in State Space Form*, Londres: Timberlake Consultants Press.

Lind, J.T. (2005). Repeated surveys and the Kalman filter. *Econometrics Journal*, 8, 418-427.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Perdreschi, D., Rinzivillo, S., Pappalardo, L. et Gabrielli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31, 263-281.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

Pang, B., et Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 1-135.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, 163-175.

Pfeffermann, D., et Burck, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 2, 229-249. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14534-fra.pdf.

Pfeffermann, D., et Rubin-Bleuer, S. (1993). Modélisation conjointe robuste de séries de données sur l'activité pour de petites régions. *Techniques d'enquête*, 19, 2, 159-174. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1993002/article/14458-fra.pdf.

Pfeffermann, D., et Sverchkov, M. (2014). Estimation of mean squared error of X-11-ARIMA and other estimators of time series components. *Journal of Official Statistics*, 30, 811-838.

Pfeffermann, D., et Tiller, R. (2006). Small area estimation with state space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

Pfeffermann, D., Feder, M. et Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics*, 16, 339-348.

Rao, J.N.K., et Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511-528.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Scott, A.J., et Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of* the *American Statistical Association*, 69, 674-678.

Scott, A.J., Smith, T.M.F. et Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review/Revue Internationale de Statistique*, 45, 13-28.

Tam, S.-M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review/Revue Internationale de Statistique*, 55, 1, 63-73.

Tiller, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

van den Brakel, J.A., et Krieg, S. (2009). Estimation du taux de chômage mensuel par modélisation structurelle de séries chronologiques dans un plan de sondage avec renouvellement de panel. *Techniques d'enquête*, 35, 2, 193-207. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2009002/article/11040-fra.pdf.

van den Brakel, J.A., et Krieg, S. (2015). Remédier aux petites tailles d'échantillon, au biais de groupe de renouvellement et aux discontinuités dans les plans de sondage avec renouvellement de panel. *Techniques d'enquête*, 41, 2, 281-312. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14231-fra.pdf.