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

Techniques d'enquête

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

par Oksana Bollineni-Balabay, Jan van den Brakel et Franz Palm

Date de diffusion : le 22 juin 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, juin 2017
Vol. 43, nº 1, pp. 47-75
Statistique Canada, nº 12-001-X au catalogue

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

Oksana Bollineni-Balabay, Jan van den Brakel et Franz Palm[1]

Résumé

La modélisation de séries chronologiques structurelle est une puissante technique de réduction des variances pour les estimations sur petits domaines (EPD) reposant sur des enquêtes répétées. Le bureau central de la statistique des Pays-Bas utilise un modèle de séries chronologiques structurel pour la production des chiffres mensuels de l'Enquête sur la population active (EPA) des Pays-Bas. Cependant, ce type de modèle renferme des hyperparamètres inconnus qui doivent être estimés avant que le filtre de Kalman ne puisse être appliqué pour estimer les variables d'état du modèle. Le présent article décrit une simulation visant à étudier les propriétés des estimateurs des hyperparamètres de tels modèles. La simulation des distributions de ces estimateurs selon différentes spécifications de modèle viennent compléter les diagnostics types pour les modèles espace-état. Une autre grande question est celle de l'incertitude entourant les hyperparamètres du modèle. Pour tenir compte de cette incertitude dans les estimations d'erreurs quadratiques moyennes (EQM) de l'EPA, différents modes d'estimation sont pris en compte dans une simulation. En plus de comparer les biais EQM, cet article examine les variances et les EQM des estimateurs EQM envisagés.

Mots-clés : Bootstrap; hyperparamètre; modèles espace-état; EQM réelle; chômage.

1 Introduction

Les chiffres de la population active que produisent les organismes nationaux de statistique (ONS) sont généralement tirés d'enquêtes sur la population active. On constate un intérêt grandissant pour la production de ces indicateurs à intervalles mensuels (EUROSTAT 2015). Toutefois, la taille des échantillons est généralement trop faible, même à l'échelon national, pour pouvoir se fier aux estimateurs fondés sur le plan des théories classiques de l'échantillonnage, pour produire des chiffres mensuels suffisamment précis de la population active (Särndal, Swensson et Wretman 1992; Cochran 1977). Dans ces situations, il est cependant possible d'utiliser les techniques d'estimation sur petits domaines (EPD) pour améliorer la taille utile des échantillons des domaines en question, en empruntant les renseignements de périodes antérieurs ou d'autres domaines (voir Rao et Molina 2015; Pfeffermann 2013). Les enquêtes répétées, notamment, se prêtent à l'amélioration dans le cadre des modèles de séries chronologiques structurels (SCS) ou multiniveau.

Les modèles SCS, tout comme les modèles multiniveau, comportent normalement des hyperparamètres inconnus qui doivent être estimés. Si l'incertitude qui les accompagne (c'est ce que nous appellerons l'incertitude des hyperparamètres) n'est pas prise en compte, les erreurs quadratiques moyennes (EQM) estimées des variables explicatives de domaine seront entachées d'un biais négatif. Dans le cadre de la modélisation multiniveau, la prise en compte de cette incertitude est une pratique à la fois nécessaire et

1. Oksana Bollineni-Balabay, Statistics Netherlands, Division de la méthodologie et de la qualité, C.P. 4481, 6401CZ Heerlen, Pays-Bas. Courriel : oksana-bl@yandex.ru, obay@cbs.nl; Jan van den Brakel, Statistics Netherlands et School of Business and Economics de l'Université de Maastricht, C.P. 616, 6200 MD Maastricht, Pays-Bas; Franz Palm, Université de Maastricht, Pays-Bas.

courante; elle se fait habituellement grâce à l'utilisation de la méthode du meilleur prédicteur linéaire sans biais empirique (MPLSBE) ou d'un modèle bayésien hiérarchique (voir Rao et Molina 2015, chapitres 6, 7 et 10). Les modèles SCS ne sont pas utilisés aussi courrамment que les modèles multiniveau dans les estimations sur petits domaines. Le filtre de Kalman, habituellement appliqué en ajustement aux modèles SCS, ne tient pas compte de l'incertitude des hyperparamètres et produit donc des estimations EQM à biais négatif. Les applications qui démontrent les avantages considérables des modèles SCS par rapport au modèles types fondés sur le plan traitent les hyperparamètres estimés d'un modèle comme étant connus (voir, par exemple, Bollineni-Balabay, van den Brakel et Palm 2016a; Krieg et van den Brakel 2012; Pfeffermann et Rubin-Bleuer 1993; Tiller 1992).

Au bureau central de la statistique des Pays-Bas (Statistics Netherlands), un modèle SCS à plusieurs variables proposé par Pfeffermann (1991) est utilisé pour produire les chiffres mensuels officiels de population active aux fins de l'Enquête sur la population active (EPA) des Pays-Bas. Comme dans bien d'autres pays, l'EPA est fondé sur un plan de sondage avec renouvellement de panel et ses échantillons sont trop petits pour produire ces chiffres mensuels. Le modèle SCS appliqué aux estimations fondées sur le plan de sondage utilise les données d'échantillonnage de périodes antérieures et tient compte du biais de renouvellement de l'échantillon (BRE) et de l'autocorrélation des erreurs d'enquête. C'est ainsi qu'on obtient des estimations mensuelles suffisamment précises de la population active en chômage (voir van den Brakel et Krieg 2015). Les modèles SCS sont également utilisés pour la production des statistiques officielles du *US Bureau of Labor Statistics* (Tiller 1992). Plusieurs ONS dans le monde commencent à manifester de l'intérêt à l'égard de cette technique, notamment en Australie (Zhang et Honchar 2016) en Israël et au Royaume-Uni (ONS 2015).

Nous présentons ici une étude élargie par simulation de Monte-Carlo, où le modèle de l'EPA sert de processus de génération de données. Cette simulation nous éclaire sur le processus de sélection de modèle, avant la mise en œuvre, aux fins de la production des statistiques officielles. D'abord, l'évaluation des distributions des estimateurs des hyperparamètres selon différentes spécifications de modèles fait ressortir l'importance de conserver certains hyperparamètres dans le modèle. Les diagnostics types pour les modèles espace-état ne fournissent que des renseignements limités sur des hyperparamètres non pertinents. S'il y a surspécification, non seulement la distribution des estimations des hyperparamètres redondants risque de grandement s'éloigner de la normalité, mais les estimations des autres hyperparamètres pourraient aussi s'en trouver perturbées. Disons donc que, même si le diagnostic est satisfaisant, il serait encore avisé de simuler le modèle et d'examiner la distribution de l'estimateur de maximum de vraisemblance (MV) de ses hyperparamètres.

Un autre but de la simulation est d'évaluer dans quelle mesure l'incertitude entourant les estimations des hyperparamètres influe sur l'estimation des EQM dans les modèles SCS. L'absence de prise en compte de cette incertitude dans l'estimation EQM est acceptable seulement si les séries chronologiques disponibles sont suffisamment longues. Ce qu'on appréciera comme période suffisamment longue variera selon les applications. Le plus souvent, les séries chronologiques ininterrompues dont disposent les ONS sont relativement courtes, surtout à cause du remaniement des enquêtes. Les études spécialisées proposent plusieurs moyens de tenir compte de l'incertitude des hyperparamètres dans un modèle SCS, qu'il s'agisse de l'approximation asymptotique, du bootstrap ou d'un traitement bayésien complet (pour ce dernier cas,

voir Durbin et Koopman 2012, chapitre 13). Nous considérerons notamment dans notre exposé l'approximation asymptotique conçue par Hamilton (1986) et le bootstrap, paramétrique ou non, conçu par Pfeffermann et Tiller (2005) et Rodriguez et Ruiz (2012). Appliquées au modèle de l'EPA, ces méthodes visent à dégager la meilleure méthode d'estimation EQM dans cette application de la vie réelle. Nous montrerons aussi comment le problème de l'incertitude des hyperparamètres s'atténue au gré d'une progression de 48 à 200 mois des séries chronologiques de l'EPA.

Notre contribution sera quadruple. Premièrement, nous démontrerons comment la simulation de Monte-Carlo peut servir à contrôler la surspécification d'un modèle (hyperparamètres redondants). Deuxièmement, nous ferons voir le meilleur des modes proposés d'estimation EQM dans l'EPA et livrerons une évaluation plus réaliste de la réduction des variances dans le modèle SCS par opposition à la modélisation type fondée sur le plan. Troisièmement, notre étude de Monte-Carlo viendra infirmer ce que disent Rodriguez et Ruiz (2012) de la supériorité de leur méthode sur la méthode bootstrap de Pfeffermann et Tiller (2005) dans un modèle plus complexe. Quatrièmement, nous jetterons un éclairage, en dehors de la comparaison des biais EQM, sur la variance et les EQM des estimateurs EQM. Autant que nous sachions, la variabilité des méthodes bootstrap mentionnées n'a pas encore été étudiée.

Voici comment se structure notre propos. À la section 2, nous décrirons l'EPA et le modèle actuellement utilisé par Statistics Netherlands. À la section 3, nous passerons en revue les modes énumérés d'estimation EQM. À la section 4, nous détaillerons le cadre de simulation propre à l'EPA. Enfin, nous décrirons nos résultats à la section 5 et livrerons des observations en conclusion à la section 6.

2 Enquête sur la population active des Pays-Bas

2.1 Plan de sondage

L'Enquête sur la population active (EPA) des Pays-Bas repose sur un plan de sondage avec renouvellement de panel depuis octobre 1999. Chaque mois, on prélève un échantillon d'adresses selon un plan d'échantillonnage stratifié à deux degrés. Les strates correspondent géographiquement à des régions. Les municipalités sont les unités primaires d'échantillonnage et les adresses, les unités secondaires. Tous les ménages résidant à une adresse sont compris dans l'échantillon. Nous considérerons ici les données d'observation de l'EPA entre janvier 2001 et juin 2010, période où on a recueilli les données de la première vague par des interviews sur place assistées par ordinateur (IPAO) et par les soins d'intervieweurs visitant à domicile les ménages échantillonnés. Après un maximum de six tentatives, l'intervieweur dépose une lettre pour le répondant, lui demandant d'appeler pour prendre rendez-vous. Quand un membre d'un ménage ne peut être contacté, on permet une interview par substitution auprès des membres du même ménage. Les répondants sont interviewés à nouveau à quatre reprises, à des intervalles trimestriels. Au cours de ces quatre vagues subséquentes, les données sont recueillies par interview téléphonique assistée par ordinateur (ITAO) et les personnes répondent à un questionnaire condensé permettant d'établir tout changement de leur situation sur le marché du travail. Les interviews par substitution sont permises. Les numéros de téléphone cellulaire et les numéros confidentiels de lignes terrestres sont recueillis dès la première vague pour prévenir

toute érosion du panel. Au début de l'application du plan de sondage avec renouvellement de panel pour l'EPA, la taille brute d'échantillon était d'environ 6 200 adresses par mois en moyenne et, dans environ 65 % des cas, il s'agissait de ménages qui répondaient entièrement. Les taux de réponse des vagues qui suivent sont d'environ 90 % du taux de la vague qui précède.

L'estimateur par la régression généralisée (ERG) (Särndal et coll. 1992) est appliqué pour estimer la population active en chômage totale. Cet estimateur tient compte de la complexité du plan d'échantillonnage et exploite l'information auxiliaire disponible dans les registres pour corriger, du moins en partie, toute non-réponse sélective. Soit Y_t^j l'ERG du nombre total de chômeurs dans le mois t pour la j^{e} vague de répondants. On obtient cinq estimations semblables par mois, chacune étant directement fondée sur l'échantillon ayant accédé à l'enquête dans le mois $t-l,\ l=\{0,3,6,9,12\}$. L'estimateur ERG de ce total de population se définit ainsi :

$$Y_t^j = \sum_{k\in s} w_{k,t}\left(\sum_{i=1}^{n_{k,t}} y_{i,k,t}\right), \tag{2.1}$$

où $y_{i,k,t}$ représente les observations de l'échantillon avec 1 si la i^{e} personne dans le k^{e} ménage est en chômage et avec zéro dans les autres cas; $n_{k,t}$ est le nombre de personnes de 15 ans et plus dans le k^{e} ménage; enfin, les $w_{k,t}$ sont les poids de régression du ménage k au moment t. La méthode de Lemaître et Dufour (1987) sert à l'obtention de poids égaux pour toutes les personnes appartenant à un même ménage :

$$w_{k,t} = \frac{1}{\pi_{k,t}}\left[1+\left(\mathbf{X}_t - \sum_{k\in s}\frac{\mathbf{x}_{k,t}}{\pi_{k,t}}\right)\left(\sum_{k\in s}\frac{\mathbf{x}_{k,t}\mathbf{x}_{k,t}'}{\pi_{k,t}g_{k,t}}\right)^{-1}\frac{\mathbf{x}_{k,t}}{g_{k,t}}\right], \tag{2.2}$$

où $\pi_{k,t}$ est la probabilité d'inclusion du ménage k au moment t, $g_{k,t}$ la taille du ménage k au moment t et $\mathbf{x}_{k,t} = \sum_{i=1}^{n_{k,t}} \mathbf{x}_{i,k,t}$, $\mathbf{x}_{i,k,t}$ étant un vecteur de J dimensions avec l'information auxiliaire de modèle de pondération sur la i^{e} personne dans le k^{e} ménage au moment t. Le vecteur $\mathbf{X}_t$ contient les totaux de population des variables auxiliaires. Le modèle de pondération est défini par les variables suivantes (le nombre de catégories figure entre parenthèses) : âge(5)sexe + région(44) + sexe(2) × âge(21) + âge(5) × état matrimonial(2) + ethnicité(8), où × désigne l'interaction des variables et où âge(5)sexe est une variable en huit classes avec l'âge en cinq catégories, dont les deuxième, troisième et quatrième se détaillent pour les deux sexes.

La variance de l'estimateur ERG Y_t^j est ainsi approchée :

$$\widehat{\text{Var}}\left(Y_t^j\right) = \sum_{h=1}^{H}\frac{n_{h,t}}{n_{h,t}-1}\left(\sum_{k=1}^{n_{h,t}}\left(w_{k,t}\hat{e}_{k,t}\right)^2 - \frac{1}{n_{h,t}}\left(\sum_{k=1}^{n_{h,t}} w_{k,t}\hat{e}_{k,t}\right)^2\right), \quad j=\{1,2,3,4,5\}, \tag{2.3}$$

où les résidus ERG sont $\hat{e}_{k,t} = \sum_{i=1}^{n_{k,t}}\left(y_{i,k,t} - \mathbf{x}_{i,k,t}'\hat{\boldsymbol{\beta}}_t\right)$; $n_{h,t}$ est le nombre de ménages dans la strate h (H étant le nombre total de strates). Le vecteur $\hat{\boldsymbol{\beta}}_t$ est un estimateur du type Horvitz-Thompson du coefficient de régression qui vient de la régression de la variable cible sur les variables auxiliaires de l'échantillon.

2.2 Le modèle SCS de l'EPA

Il y a deux raisons pour lesquelles Statistics Netherlands a décidé de passer à un modèle de production fondé sur les séries chronologiques, en juin 2010. La première était que l'échantillon de l'EPA était de trop petite taille pour produire des estimations mensuelles. Puisque l'échantillon dans la première vague était constitué d'environ 4 000 ménages, en moyenne, les estimations ERG de la population active en chômage présentaient un coefficient de variation d'environ 4 % à l'échelon national, ce qui était jugé trop instable pour la publication des statistiques officielles. Il faut aussi dire que les chiffres mensuels du chômage doivent être diffusés pour six domaines en fonction d'une classification sexe-âge. Les estimations fondées sur le plan pour ces domaines présentent des coefficients de variation bien plus élevés. Une autre difficulté avec l'EPA est ce que l'on appelle le biais de renouvellement de l'échantillon (BRE), c'est-à-dire les différences systématiques entre les estimations issues des différentes vagues (voir, par exemple, Bailar 1975, ou Pfeffermann 1991). Parmi les explications courantes de ce biais figurent l'érosion de l'échantillon, les effets d'échantillon longitudinal et les différences entre les questionnaires et les modes propres aux diverses vagues successives. Dans le cas de l'EPA, on présume que les estimations de la première vague sont les plus fiables et que celles des vagues subséquentes sous-estiment systématiquement les effectifs de chômeurs. Pour un examen plus détaillé, voir van den Brakel et Krieg (2009).

Les deux problèmes sont résolus avec le modèle SCS qui utilise en entrée cinq séries d'estimations ERG pour les cinq vagues considérées. Dans cette modélisation, on décompose une série observée en plusieurs composantes inobservées (tendance et composante saisonnière, par exemple). On peut employer le filtre de Kalman, en combinaison facultative avec un algorithme de lissage, pour extraire ces composantes de la série chronologique observée. C'est ainsi qu'on sépare les estimations des composantes définissant le signal du chômage de la variance inexpliquée du paramètre de population, ainsi que de la variance d'échantillonnage. Cela donne généralement des estimations ponctuelles moins instables et des erreurs-types bien moindres que celles qui caractérisent les estimations ERG. En modélisant les différences systématiques entre les cinq séries en entrée, le modèle tient aussi compte du biais de renouvellement du panel.

Dans chaque mois t, un vecteur à cinq dimensions $\mathbf{Y}_t = \left(Y_t^1\ Y_t^2\ Y_t^3\ Y_t^4\ Y_t^5\right)'$ est observé. Il contient les estimations ERG de nombre total de chômeurs pour les cinq vagues considérées. En se fondant sur Pfeffermann (1991), van den Brakel et Krieg (2009) ont conçu le modèle suivant pour les estimations ERG $\mathbf{Y}_t$:

$$\mathbf{Y}_t = \mathbf{1}_5 \xi_t + \boldsymbol{\lambda}_t + \mathbf{e}_t, \tag{2.4}$$

où $\mathbf{1}_5$ est un vecteur colonne à cinq dimensions de uns, où ξ_t est le paramètre réel de population (scalaire) qui est inconnu, où $\boldsymbol{\lambda}_t$ est un vecteur contenant des variables d'état pour le BRE et enfin où $\mathbf{e}_t$ est un vecteur des erreurs d'enquête en corrélation avec les erreurs correspondantes des vagues antérieures (nous présentons cette structure plus loin). Pour le paramètre réel de population, nous posons que $\xi_t = L_t + \gamma_t + \varepsilon_t$, soit la somme d'une tendance stochastique L_t, d'une composante saisonnière stochastique γ_t, et d'une composante irrégulière $\varepsilon_t \overset{\text{iid}}{\sim} N\left(0, \sigma_\varepsilon^2\right)$.

Dans le cas de la tendance stochastique L_t, nous posons ce qu'on appelle le modèle de lissage de la tendance :

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1}, \\ R_t &= R_{t-1} + \eta_{R,t}, \end{aligned}$$

où L_t et R_t correspondent au niveau et à la pente du paramètre réel de population; le terme de perturbation de la pente présente la distribution suivante : $\eta_{R,t} \overset{\text{iid}}{\sim} N\left(0, \sigma_R^2\right)$.

Pour la composante saisonnière γ_t, nous posons le modèle trigonométrique :

$$\gamma_t = \sum_{l=1}^{6} \gamma_{t,l},$$

où chacune de ces six harmoniques suit le processus suivant :

$$\begin{aligned} \gamma_{t,l} &= \cos\left(h_l\right)\gamma_{t-1,l} + \sin\left(h_l\right)\gamma^*_{t-1,l} + \omega_{t,l}, \\ \gamma^*_{t,l} &= -\sin\left(h_l\right)\gamma_{t-1,l} + \cos\left(h_l\right)\gamma^*_{t-1,l} + \omega^*_{t,l}, \end{aligned}$$

$h_l = \frac{\pi l}{6}$ étant la l^{e} fréquence saisonnière, $l = \{1, \ldots 6\}$. Nous posons que les termes stochastiques $\omega_{t,l}$ et $\omega^*_{t,l}$ à espérance nulle sont normalement et indépendamment distribués et présentent la même variance dans et entre tous les harmoniques :

$$\begin{aligned} &\text{Cov}\left(\omega_{t,l}, \omega_{t',l'}\right) = \text{Cov}\left(\omega^*_{t,l}, \omega^*_{t',l'}\right) = \begin{cases} \sigma^2_\omega & \text{si } l = l' \text{ et } t = t', \\ 0 & \text{si } l \neq l' \text{ ou } t \neq t', \end{cases} \\ &\text{Cov}\left(\omega_{t,l}, \omega^*_{t,l}\right) = 0 \text{ pour tous les } l \text{ et } t. \end{aligned}$$

La deuxième composante en (2.4) est le biais de renouvellement (BRE). Nous posons que la première vague est sans biais, ainsi que l'expliquent van den Brakel et Krieg (2009). Les BRE des vagues qui suivent sont fonction du temps et se modélisent comme des processus à marche aléatoire. On justifie le tout en disant que les procédures de terrain subissent de fréquents changements et que, par ailleurs, les taux de réponse évoluent progressivement dans le temps, ce qui rend le BRE tributaire du temps, comme l'illustrent van den Brakel et Krieg (2015) (voir la figure 4.3). Le vecteur BRE des cinq vagues peut s'écrire ainsi : $\boldsymbol{\lambda}_t = \left(0 \ \lambda_t^2 \ \lambda_t^3 \ \lambda_t^4 \ \lambda_t^5\right)'$, avec :

$$\lambda_t^j = \lambda_{t-1}^j + \eta_{\lambda,t}^j, \quad j = \{2, 3, 4, 5\}.$$

Nous posons que les perturbations BRE ne sont pas corrélées entre les vagues et que leur distribution est normale, c'est-à-dire $\eta_{\lambda,t}^j \overset{\text{iid}}{\sim} \left(0, \sigma_\lambda^2\right)$, avec égalité des variances dans les quatre vagues.

La dernière composante en (2.4) est celle des erreurs d'enquête pour les cinq estimations ERG, c'est-à-dire $\mathbf{e}_t = \left(e_t^1 \ e_t^2 \ e_t^3 \ e_t^4 \ e_t^5\right)'$. Pour tenir compte de l'hétérogénéité des erreurs d'échantillonnage causée par les variations temporelles de taille d'échantillon, nous modélisons ces erreurs en proportion des erreurs-types fondées sur le plan, d'après le modèle d'erreur de mesure proposé par Binder et Dick (1990), c'est-à-dire $e_t^j = \tilde{e}_t^j z_t^j$, où $z_t^j = \sqrt{\widehat{\text{Var}}\left(Y_t^j\right)}$ et $\tilde{e}_t^j$ sont des erreurs d'échantillonnage réduites ou normalisées en

fonction d'un processus stationnaire que nous définirons plus loin. Ici, les $\widehat{\mathrm{Var}}(Y_t^j)$ sont les estimations des variances, fondées sur le plan, qui sont tirées des microdonnées en (2.3). Ils sont traités comme variances d'échantillonnage connues a priori dans le modèle SCS.

Comme l'échantillon de la première vague n'est pas en chevauchement avec les échantillons observés par le passé, les $\tilde{e}_t^t$ peuvent se modéliser comme du bruit blanc avec $E(\tilde{e}_t^1) = 0$ et $\mathrm{Var}(\tilde{e}_t^1) = \sigma_{v_1}^2$. La variance des erreurs d'échantillonnage e_t^t sera égale à la variance des estimations ERG si l'estimation de maximum de vraisemblance des $\sigma_{v_1}^2$ est à peu près égale à l'unité.

Dans les vagues qui suivent, les erreurs d'enquête sont en corrélation avec les erreurs d'enquête des vagues antérieures. Nous estimons le coefficient d'autocorrélation à partir des données d'enquête par la méthode que proposent Pfeffermann, Feder et Signorelli (1998). La structure d'autocorrélation est mise en modélisation autorégressive AR(1) et le coefficient d'autocorrélation s'obtient par les équations de Yule-Walker (van den Brakel et Krieg 2009):

$$\tilde{e}_t^j = \rho \tilde{e}_{t-3}^{j-1} + \nu_t^j,\ \nu_t^j \overset{\text{iid}}{\sim} N\left(0, \sigma_{v_j}^2\right), \quad j = \{2, 3, 4, 5\}.$$

Nous posons que le coefficient d'autocorrélation du premier ordre est commun aux quatre vagues. Son estimation fait fonction d'information a priori dans le modèle. Comme $\tilde{e}_t^j$ représente un processus AR(1), $\mathrm{Var}(\tilde{e}_t^j) = \sigma_{v_j}^2 \big/ (1-\rho^2)$. La variance de l'erreur d'échantillonnage e_t^j correspond approximativement à $\widehat{\mathrm{Var}}(Y_t^j)$ si l'estimation de maximum de vraisemblance des $\sigma_{v_j}^2$ est à peu près égale à $(1-\rho^2)$. Nous posons cinq hyperparamètres différents $\sigma_{v_j}^2$, $j = \{1, 2, 3, 4, 5\}$, pour les erreurs d'échantillonnage comme composantes des cinq vagues.

Nous regroupons les variances de perturbation avec le paramètre d'autocorrélation ρ dans un vecteur d'hyperparamètres appelé $\boldsymbol{\theta} = \left(\sigma_R^2\ \sigma_\omega^2\ \sigma_\varepsilon^2\ \sigma_\lambda^2\ \sigma_{v_1}^2\ \sigma_{v_2}^2\ \sigma_{v_3}^2\ \sigma_{v_4}^2\ \sigma_{v_5}^2\ \rho\right)'$, le vecteur contenant seulement les variances de perturbation est $\boldsymbol{\theta}_\sigma = \left(\sigma_R^2\ \sigma_\omega^2\ \sigma_\varepsilon^2\ \sigma_\lambda^2\ \sigma_{v_1}^2\ \sigma_{v_2}^2\ \sigma_{v_3}^2\ \sigma_{v_4}^2\ \sigma_{v_5}^2\right)'$. Pour éviter les estimations négatives, nous estimons à l'échelle logarithmique les hyperparamètres des variances de perturbation dans $\boldsymbol{\theta}_\sigma$. Nous employons la méthode du quasi-maximum de vraisemblance (voir, par exemple, Harvey 1989) où on traite les estimations $\hat{\rho}$ comme étant connues. Dans cette étude, l'analyse numérique se fait avec OxMetrics 5 (Doornik 2007) en combinaison avec le progiciel *SsfPack 3.0* (Koopman, Shephard et Doornik 2008).

3 Modes d'estimation EQM

D'ordinaire, on ajuste les modèles structurels linéaires de séries chronologiques ayant des composantes inobservées en appliquant le filtre de Kalman à l'espace-état une fois formé à partir de ces composantes. On peut voir dans Bollineni-Balabay, van den Brakel et Palm (2016b) quelle est la représentation en espace-état du modèle SCS pour l'EPA. Le vecteur d'état $\boldsymbol{\alpha}_t$ contient les variables d'état définies à la section précédente, c'est-à-dire la tendance, la pente, les harmoniques saisonnières, le BRE, le bruit blanc de population et les erreurs d'enquête. Nous initialisons toutes les variables d'état non stationnaires en prenant une distribution antérieure diffuse (à moyenne nulle et à très grande variance). Les cinq composantes des

erreurs d'enquête $\tilde{e}_t^j$, $j = \{1, 2, 3, 4, 5\}$ et le bruit blanc de population ε_t sont des variables d'état stationnaires initialisées avec des zéros. Nous tenons la variance initiale des erreurs d'échantillonnage de la première vague pour égale à l'unité et nous considérons que la variance des autres vagues correspond à $(1-\rho^2)$. On pourrait même prendre une petite valeur pour la variance initiale de ε_t.

On extrait habituellement des estimations filtrées du vecteur d'état $\boldsymbol{\alpha}_t$ et de sa matrice des covariances $\mathbf{P}_{t|t}$ à l'aide du filtre de Kalman (voir Harvey 1989). Ainsi, $\mathbf{P}_{t|t}$ contient les EQM extraites par le filtre conditionnellement à l'information obtenue jusqu'au moment t inclusivement :

$$\mathbf{P}_{t|t} = \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)'\right], \tag{3.1}$$

où nous posons que $\boldsymbol{\theta}$ est la valeur réelle des hyperparamètres et où l'espérance se prend sur la codistribution du vecteur d'état et des valeurs Y au moment t. Dans la pratique, le vecteur réel des hyperparamètres est remplacé par son estimation $\hat{\boldsymbol{\theta}}$ dans les récursions par filtre de Kalman. Dans ce cas, l'EQM en (3.1) n'est plus l'EQM réelle. On la qualifie de « naïve », puisqu'elle ne tient pas compte de l'incertitude autour des estimations $\hat{\boldsymbol{\theta}}$. L'EQM réelle devient ainsi :

$$\mathbf{EQM}_{t|t} = \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t\right)'\right],$$

ce qui représente une valeur supérieure à la valeur EQM en (3.1) et peut se décomposer comme la somme de l'incertitude du filtre et de l'incertitude des paramètres dans une condition de normalité des termes d'erreur :

$$\mathbf{EQM}_{t|t} = \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)'\right] + \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta})\right)'\right]. \tag{3.2}$$

Le premier terme, l'incertitude du filtre, est ce qui est estimé par les estimations EQM $\mathbf{P}_{t|t}$ par le filtre de Kalman. Il faut aller plus loin pour estimer le deuxième terme, l'incertitude des paramètres. Les études spécialisées consacrées à l'estimation EQM proposent deux grandes méthodes, à savoir l'approximation asymptotique et le bootstrap. Le bootstrap peut être paramétrique ou non paramétrique. Quelques observations s'imposent ici au sujet de ces méthodes dans le contexte du modèle SCS appliqué à l'EPA.

Dans le cas du bootstrap paramétrique, les perturbations d'état, $\boldsymbol{\eta}_t$, disons, sont tirées de coestimations de densité normale conditionnelle à plusieurs variables $\boldsymbol{\eta}_t \overset{\text{iid}}{\sim} \mathrm{MN}\left(\mathbf{0}, \hat{\boldsymbol{\Omega}}\right)$, $\hat{\boldsymbol{\Omega}}$ étant évalué à l'estimation $\hat{\boldsymbol{\theta}}$ des hyperparamètres. Ces perturbations servent dans les récursions d'état par filtre de Kalman à produire les variables d'état. Par ailleurs, le bootstrap non paramétrique a pour avantage de ne dépendre d'aucune hypothèse particulière au sujet de cette codistribution. Si dans le bootstrap paramétrique les perturbations d'état viennent de l'estimation de leur distribution, dans le bootstrap non paramétrique il y a rééchantillonnage avec remise dans un nouvel ensemble normalisé en fonction des estimations initiales des hyperparamètres. Les nouveaux ensembles normalisés qui sont rééchantillonnés servent en outre à produire des séries bootstrap $\left\{\mathbf{Y}_1^b, \ldots, \mathbf{Y}_T^b\right\}$ par ce qu'on appelle la forme d'innovation du filtre de Kalman (voir les détails dans Harvey 1989, ou Bollineni-Balabay et coll. 2016b). Dans le modèle de l'EPA, les 13 premiers

points temporels d'un nouvel ensemble normalisé ne font pas l'objet d'un rééchantillonnage et ils constituent ce qu'on appelle l'échantillon diffus (c'est le temps dont on a besoin pour construire une distribution appropriée pour les variables d'état non stationnaires; voir dans Koopman (1997) l'initialisation de telles variables).

Si un modèle SCS compte des composantes non stationnaires comme dans le modèle de l'EPA, les séries produites divergeront probablement de l'ensemble de données au départ de l'application du bootstrap, c'est-à-dire de $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$. Il nous faut donc recourir à une procédure spéciale pour que les échantillons bootstrap soient mis en correspondance avec la configuration de l'ensemble de données initial, ce qu'on peut faire à l'aide d'un algorithme de lissage par simulation qui a été conçu par Durbin et Koopman (2002). On trouvera les détails techniques sur cette application dans Koopman et coll. (2008), chapitre 8.4.2. On n'a pas à prévoir de corrections pour les erreurs d'enquête issues comme nous l'avons décrit des récursions inconditionnelles d'état par le bootstrap paramétrique ou non paramétrique, puisqu'il s'agit d'un bruit (en autocorrélation).

Dans les sections qui suivent, nous présenterons brièvement la méthode asymptotique, ainsi que les applications bootstrap récentes de Rodriguez et Ruiz (2012) (bootstrap RR) et de Pfeffermann et Tiller (2005) (bootstrap PT).

3.1 Application bootstrap de Rodriguez et Ruiz

Rodriguez et Ruiz (2012) ont conçu leur méthode bootstrap d'estimation EQM conditionnelle aux données, ce qui veut dire qu'on applique en plus les hyperparamètres bootstrap à l'ensemble de données initial pour obtenir des estimations bootstrap des variables d'état. Il peut s'agir d'un bootstrap paramétrique ou non avec les étapes suivantes :

1. On estime le modèle et obtient les estimations $\hat{\boldsymbol{\theta}}$ des hyperparamètres.
2. On produit un échantillon bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ à l'aide de $\hat{\boldsymbol{\theta}}$ par bootstrap paramétrique ou non (voir l'introduction de cette section). Si le modèle est non stationnaire, on se doit de corriger l'échantillon bootstrap par simulation de lissage.
3. On se sert de l'ensemble bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ pour obtenir tant les estimations paramétriques d'autocorrélation des erreurs d'enquête $\hat{\rho}^b$ que les estimations bootstrap de maximum de vraisemblance $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^b$. On applique ensuite le filtre de Kalman à la série initiale $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$ et aux $\hat{\boldsymbol{\theta}}^b$, nouvellement estimés, ce qui donne $\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$ et $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$.
4. On reprend B fois les étapes 2 et 3, puis procède à l'estimation EQM de la manière suivante :

$$\widehat{\mathbf{EQM}}_{t|t}^{\mathrm{RR}} = \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) + \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\overline{\boldsymbol{\alpha}}_{t|t}\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\overline{\boldsymbol{\alpha}}_{t|t}\right]', \tag{3.3}$$

où $\overline{\boldsymbol{\alpha}}_{t|t} = \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$.

L'équation (3.3) est applicable aux estimations EQM par bootstrap paramétrique et non paramétrique (nous emploierons dans ce cas les abréviations $\mathrm{EQM}^{\mathrm{RR1}}$ et $\mathrm{EQM}^{\mathrm{RR2}}$ dans la suite du texte).

3.2 Application bootstrap de Pfeffermann et Tiller

La méthode bootstrap conçue par Pfeffermann et Tiller (2005) est un bootstrap inconditionnel, c'est-à-dire que variables d'état bootstrap sont dérivées de l'ensemble de données bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$, et non de l'ensemble de données initial $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$ comme dans Rodriguez et Ruiz (2012). Pfeffermann et Tiller (2005) ont démontré que leur méthode approche l'EQM réelle jusqu'à un ordre de $O(1/T^2)$ (Pfeffermann et Tiller (2005), annexe C) :

$$\widehat{\mathbf{EQM}}_{t|t}^{\mathrm{PT}} = 2\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) + \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]'. \qquad (3.4)$$

L'équation (3.4) est applicable aux estimateurs EQM par bootstrap paramétrique ou non (nous emploierons dans ce cas les abréviations $\mathrm{EQM}^{\mathrm{PT1}}$ et $\mathrm{EQM}^{\mathrm{PT2}}$ dans la suite du texte). Le calcul EQM en (3.4) exige deux exécutions du filtre de Kalman pour chaque série bootstrap. À la première exécution, $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right)$ est estimé à partir de l'ensemble bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ et des paramètres bootstrap $\hat{\boldsymbol{\theta}}^b$. Dans cette exécution, on peut aussi obtenir $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$ par $\hat{\boldsymbol{\theta}}^b$, puisque la matrice $\mathbf{P}_{t|t}$ ne dépend pas des données. Il faut appliquer le filtre de Kalman une deuxième fois pour produire les estimations d'état $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)$ en fonction de $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ et des estimations $\hat{\boldsymbol{\theta}}$ tirées de l'ensemble initial. La procédure se résume ainsi :

1. Estimer le modèle à l'aide de l'ensemble de données initial et obtenir les estimations $\hat{\boldsymbol{\theta}}$ du vecteur des hyperparamètres. Garder les estimations EQM « naïves » $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ pour une utilisation future en (3.4).
2. Utiliser le bootstrap paramétrique ou non pour produire un échantillon bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$. Apporter la correction par simulation de lissage si le modèle est non stationnaire.
3. Établir les estimations bootstrap $\hat{\boldsymbol{\theta}}^b$ des hyperparamètres à partir de l'ensemble bootstrap nouvellement produit. Appliquer le filtre de Kalman une première fois pour obtenir $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right)$ et $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$, et une autre fois pour dégager $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)$, comme décrit en (3.4).
4. Répéter B fois les étapes 2 et 3, puis procéder à l'estimation EQM en (3.4).

Pfeffermann et Tiller (2005) signalent que, dans le cas du bootstrap paramétrique, il est possible d'éviter le deuxième filtre de Kalman, parce que le vecteur d'état réel est produit (et donc connu) pour chaque série bootstrap. On peut donc remplacer les estimations d'état $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)$ en (3.4) par le vecteur réel $\boldsymbol{\alpha}_t^b$ pour obtenir l'estimateur EQM suivant :

$$\widehat{\mathbf{EQM}}_{t|t}^{\mathrm{PT1}} = \mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) + \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]'. \qquad (3.5)$$

Il y a un seul $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ du côté droit de (3.5), puisque le nouveau terme $\mathrm{E}_B\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]'$, qui correspond au dernier terme du côté droit de (3.5), peut lui-même se décomposer comme en (3.2) en une mesure de l'incertitude des paramètres $\mathrm{E}_B\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]'$ et de l'incertitude du filtre $\mathbf{P}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) = \mathrm{E}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]'$, $\hat{\boldsymbol{\theta}}$ étant le vecteur réel des paramètres par lequel on produit les variables d'état bootstrap $\boldsymbol{\alpha}_t^b$. Toutefois, on aura peut-être à prévoir beaucoup plus d'itérations bootstrap pour le terme moyen bootstrap $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]'$ remplaçant $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ si on

veut qu'il y ait convergence. Ajoutons que cette méthode simplifiée peut créer plus de biais si l'hypothèse de normalité n'est pas respectée au sujet des termes d'erreur du modèle. Dans ce cas, la décomposition du terme $\mathrm{E}_B\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\boldsymbol{\alpha}^b_t\right]'$ comme en (3.2) laissera aussi un terme croisé non nul : $\mathrm{E}\left\{\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)\right]\right\}$. Dans cette application, les moyennes bootstrap à terme croisé non nul se sont révélées négligeables, mais la moyenne bootstrap $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]'$ s'éloignait largement (dans les deux sens) du terme qu'elle était censée remplacer, ce qu'expliquerait le fait que l'EQM réelle par filtre de Kalman en (3.1) puisse être tirée de séries en simulation si, dans sa distribution, le vecteur d'état est suffisamment dispersé. Quand on met des modèles non stationnaires en bootstrap, les séries bootstrap suivent forcément la configuration de la série initiale sous-jacente, comme nous l'avons mentionné dans la description de l'algorithme de lissage par simulation. Il se peut donc que le terme $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]'$, qui remplace $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ en (3.5), n'en soit pas suffisamment proche. C'est pourquoi le bootstrap paramétrique (PT1) ou non (PT2) dans cette application dépend de l'estimateur en (3.4).

Disons quelques mots du rôle de la simulation de lissage de Durbin et Koopman (2002) dont nous avons fait mention à la fin de l'introduction à la présente section. Nous avons proposé de l'employer à l'étape de la production des séries boostrap, sans quoi la distribution bootstrap des hyperparamètres tirée de séries non corrigées pour un modèle non stationnaire pourrait être fort différente de ce qu'elle devrait être pour une réalisation particulière des données dont nous disposons. Dans le cas de l'EPA du moins, les distributions bootstrap des hyperparamètres étaient bien plus diffuses sans la simulation de lissage qu'avec celle-ci. De plus, les distributions bootstrap des hyperparamètres qui viennent de séries non corrigées dans l'EPA sont centrées sur des valeurs bien supérieures aux valeurs des hyperparamètres qui ont servi à produire les séries. Le résultat est une moyenne bootstrap extrêmement élevée $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$ (par rapport à $\mathbf{P}_{t|t}(\hat{\boldsymbol{\theta}})$) et, par la suite, des estimations EQM même inférieures aux estimations naïves. Il faut aussi dire que le terme $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)\right]'$ devient très instable dans le temps et prend des proportions excessives quand il n'y a pas de simulation de lissage, ce qui ne compense pas le biais négatif en (3.4) sans la simulation de lissage.

3.3 Approximation asymptotique

Hamilton (1986) a conçu une approximation asymptotique (AA) de l'EQM réelle à l'équation (3.2). Cette approximation peut s'exprimer comme une espérance sur la codistribution asymptotique des hyperparamètres $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)$, celle-ci étant conditionnelle à l'ensemble de données initial $\mathbf{Y}\equiv\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$. Dans la présente application, la partie du vecteur des hyperparamètres qui est estimée par la méthode du maximum de vraisemblance $\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}\right)$ dépend de la valeur estimée du paramètre autorégressif $\hat{\rho}$. Ainsi, la codistribution asymptotique de l'estimateur des hyperparamètres est de la forme suivante : $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)=\pi\left(\hat{\rho}\middle|\mathbf{Y}\right)\pi\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}\middle|\hat{\rho},\mathbf{Y}\right)$. L'EQM est ainsi approchée :

$$\mathbf{EQM}_{t|t}=\mathrm{E}_{\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)}\left[\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right)\right]+\mathrm{E}_{\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)}\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right)-\hat{\boldsymbol{\alpha}}_{t|t}\left(\mathbf{Y}\right)\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right)-\hat{\boldsymbol{\alpha}}_{t|t}\left(\mathbf{Y}\right)\right)'\right], \qquad (3.6)$$

où $\mathrm{E}_{\pi(\hat{\boldsymbol{\theta}}|\mathbf{Y})}$ est une espérance prise sur la codistribution asymptotique de l'estimateur des hyperparamètres $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)$, et où les $\hat{\boldsymbol{\alpha}}_{t|t}(\mathbf{Y})$ sont les estimations du vecteur d'état quand les hyperparamètres ne sont pas connus $\mathrm{E}_{\pi(\hat{\boldsymbol{\theta}}|\mathbf{Y})}\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right)\right]$.

Dans ce cas, nous choisissons la distribution $N(\hat{\rho},\mathrm{Var}(\hat{\rho}))$ comme la distribution asymptotique $\pi(\hat{\rho}|\mathbf{Y})$ des $\hat{\rho}$, d'où sont tirées les réalisations aléatoires $\hat{\rho}$. En général, la distribution d'échantillonnage du coefficient de corrélation revêt une forme complexe, mais elle peut fort bien être approchée par une distribution normale; tel était le cas dans cette application (la distribution normale était un très bon ajustement de la distribution en simulation et de la distribution bootstrap de $\hat{\rho}$). Si on prend l'équation (3) dans Bartlett (1946) et qu'on considère que le coefficient autorégressif dans un processus AR(1) est égal à la corrélation pour le décalage 1, l'estimateur de variance de $\hat{\rho}$ devient $\mathrm{Var}(\hat{\rho}) \approx (1-\hat{\rho}^2)/T$. Dans le cas de l'EPA où $\hat{\rho} = 0{,}208$, cela veut dire que $\widehat{\mathrm{Var}}(\hat{\rho}) \approx 0{,}96(1/T)$. Comme l'erreur-type des $\hat{\rho}$ sert à tirer des réalisations de la distribution asymptotique et que l'extraction de la racine carrée est une fonction concave, l'écart-type de l'échantillon serait une sous-estimation. En tirant donc $\hat{\rho}$ réalisations au moyen de $1/\sqrt{T}$ comme écart-type de la distribution asymptotique, on ferait un choix raisonnable.

On obtient de la manière suivante un échantillon de B réalisations de la distribution asymptotique des hyperparamètres. Après avoir tiré une valeur, $\hat{\rho}^a$ disons, de $\pi(\hat{\rho}|\mathbf{Y})$, nous réestimons les autres hyperparamètres de l'ensemble de données initial pour obtenir $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}}\left|\hat{\rho}^a,\mathbf{Y}\right.$ et la matrice d'information $\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}}\left|\hat{\rho}^a,\mathbf{Y}\right.\right)$. Finalement, nous tirons une réalisation $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^a$ de la distribution $\mathrm{MN}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}},\hat{\mathbf{I}}^{-1}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}}\left|\hat{\rho}^a,\mathbf{Y}\right.\right)\right)$. Nous appliquons à nouveau le filtre de Kalman avec les réalisations $\hat{\rho}^a$ et $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^a$ pour obtenir les estimations d'état $\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a,\mathbf{Y}\right)$ et leurs EQM $\hat{\mathbf{P}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a\right)$. La procédure se répète jusqu'à ce que B itérations $\hat{\boldsymbol{\theta}}^a$ aient été effectuées, après quoi nous dégageons (3.6) en prenant la moyenne des quantités nécessaires sur B itérations. Si tous les hyperparamètres du modèle sont estimés par la méthode du maximum de vraisemblance, B itérations peuvent se faire directement à partir de $\mathrm{MN}\left(\hat{\boldsymbol{\theta}}^{\mathrm{MV}},\hat{\mathbf{I}}^{-1}\left(\hat{\boldsymbol{\theta}}^{\mathrm{MV}}\right)\right)$.

On peut approcher le premier terme en (3.6) par la valeur moyenne de la variance $\mathbf{P}_{t|t}$ par filtre de Kalman sur B réalisations du vecteur des hyperparamètres. Le deuxième terme peut être approché par la variance des estimations du vecteur d'état sur ces mêmes B itérations. Une approximation asymptotique des EQM pourrait se dégager de la manière suivante :

$$\widehat{\mathbf{EQM}}_{t|t}^{\mathrm{AA}} = \frac{1}{\mathrm{B}}\sum_{a=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^a\right) + \frac{1}{\mathrm{B}}\sum_{a=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a,\mathbf{Y}\right)-\overline{\hat{\boldsymbol{\alpha}}}_{t|t}\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a,\mathbf{Y}\right)-\overline{\hat{\boldsymbol{\alpha}}}_{t|t}\right]', \tag{3.7}$$

où $\hat{\boldsymbol{\theta}}^a$ est le résultat du a^{e} tirage à partir de la distribution asymptotique $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)$. Comme le propose Hamilton (1986), la moyenne d'échantillon $\overline{\hat{\boldsymbol{\alpha}}}_{t|t} = \frac{1}{\mathrm{B}}\sum_{a=1}^{B}\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a,\mathbf{Y}\right)$ peut remplacer $\hat{\boldsymbol{\alpha}}_{t|t}(\mathbf{Y})$ en (3.6). Cet auteur ajoute qu'une telle décomposition de l'incertitude du total en une incertitude du filtre et une incertitude des paramètres ressemble à la décomposition bien connue $\mathrm{var}(X) = E[\mathrm{var}(X|Y)] + \mathrm{var}[E(X|Y)]$. Manifestement, cet estimateur EQM repose entièrement sur l'hypothèse d'une normalité asymptotique de l'estimateur du vecteur des hyperparamètres. De plus, cette application produit habituellement des biais significatifs si les séries ne sont pas d'une longueur suffisante, auquel cas la distribution asymptotique normale qui est posée ne pourrait approcher la distribution finie (ordinairement asymétrique) des estimations de maximum de vraisemblance.

Un autre problème est susceptible de se poser avec le traitement asymptotique si on estime que les hyperparamètres sont proches de zéro, ce qui peut advenir des estimations du modèle au départ ou pendant l'application de la procédure même à cause de certaines réalisations extrêmes de $\hat{\rho}$. Dans ce cas, la variance asymptotique de ces hyperparamètres sera très élevée, ce qui viendra gonfler les estimations EQM du signal et de ses composantes inobservées. Il pourrait en résulter un défaut d'inversion de la matrice d'information pour le vecteur des hyperparamètres.

4 L'EPA et son cadre précis de simulation

Nous allons examiner le rendement des cinq méthodes d'estimation EQM par rapport à des séries de la longueur initiale de cette enquête (114 points mensuels de 2001(1) à 2010(6)) et à des séries soit plus courtes de 48 et 80 mois soit plus longues de 200 mois. Pour chacune de ces durées, nous montons une expérience de Monte-Carlo où des séries multiples (1 000) font l'objet d'une simulation en fonction du modèle EPA du nombre de chômeurs. Nous estimons les EQM de chacune de ces séries en prenant $B = 300$ séries bootstrap. Dans le cas de l'approximation asymptotique toutefois, il nous a fallu prévoir au moins $B = 500$ itérations. Nous avons jugé que ce nombre était suffisant pour qu'il y ait convergence des EQM approchées. Nous comparons les EQM issues des cinq méthodes et mises en moyenne sur 1 000 simulations aux moyennes EQM produites par un filtre de Kalman « naïf ». Dans ce cas, au moins 10 000 simulations sont nécessaires pour que les estimations EQM convergent sur une certaine moyenne.

Voici comment nous obtenons paramétriquement les séries artificielles $\mathbf{Y}_t^s$ mentionnées pour des simulations $s = 1, \ldots, 1\,000$ (ou 10 000) : d'abord, nous établissons les estimations $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}$ de maximum de vraisemblance des hyperparamètres par ajustement du modèle SCS aux séries initiales; ensuite, nous tirons aléatoirement des perturbations d'état (on se rappellera que les erreurs d'enquête sont aussi modélisées comme variables d'état) de leur codistribution normale $N\left(\mathbf{0}, \boldsymbol{\Omega}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}\right)\right)$, et produisons les séries par récursion de filtre de Kalman. Comme le système est non stationnaire, les séries produites $\mathbf{Y}_t^s$ peuvent donner des nombres de chômeurs négatifs ou excessivement élevés. Pour éviter tout nombre démesuré de séries aux valeurs négatives, nous appliquons la récursion des variables d'état à partir des estimations lissées des états à un des points les plus hauts des séries observées. De plus, nous écartons les 30 premiers points temporels pour empêcher que les séries ne commencent au même point temporel. Dans l'hypothèse que le chômage aux Pays-Bas ne sera pas de plus de 15 % de toute la population active, nous limitons l'ensemble de données en simulation aux séries dont les valeurs vont de zéro à 1 million de chômeurs (il s'agit d'environ 15 % de la population active des Pays-Bas en 2010), les autres séries étant éliminées. Si nous gardons nos séries artificielles sous la borne supérieure, c'est aussi pour ne pas extrapoler en dehors de la plage initiale des données dans la simulation des erreurs-types z_t^j fondées sur le plan.

Toute série d'estimations ponctuelles ERG en simulation exige sa propre série d'estimations des erreurs-types fondées sur le plan en simulation z_t^j. Les estimations initiales connues des erreurs-types fondées sur le plan $\sqrt{\widehat{\mathrm{Var}}\left(Y_t^j\right)}$ ne conviendraient pas à cette simulation, parce que la variance de l'erreur d'échantillonnage est proportionnelle à l'estimation ponctuelle correspondante. La fonction de variance suivante nous a permis de produire des variances fondées sur le plan pour la série simulée d'estimations ponctuelles (voir les détails à l'annexe B dans Bollineni-Balabay et coll. 2016b) :

$$\ln\left[\widehat{\mathrm{Var}}\left(Y_t^1\right)\right] = \ln\left[\left(z_t^1\right)^2\right] = c + \beta_1 \ln\left(l_t^1\right) + \varepsilon_t^1,\ \varepsilon_t^t \sim N\left(0, \left(\sigma_\varepsilon^1\right)^2\right);$$
$$\ln\left[\widehat{\mathrm{Var}}\left(Y_t^j\right)\right] = \ln\left[\left(z_t^j\right)^2\right] = \psi_j \ln\left[\left(z_{t-3}^{j-1}\right)^2\right] + \beta_j \ln\left(l_t^j\right) + \varepsilon_t^j,\ \varepsilon_t^j \sim N\left(0, \left(\sigma_\varepsilon^j\right)^2\right),\ j = \{2, 3, 4, 5\}, \quad (4.1)$$

où $l_t^j,\ j = \{1, 2, 3, 4, 5\}$ est le signal d'une vague comme somme de la tendance, de la composante saisonnière et du BRE. Les coefficients de régression en (4.1) sont invariants dans le temps et s'obtiennent par régression de $\ln\left(z_t^j\right)^2$ sur $\ln\left(l_t^j\right)$ et $\ln\left(\left(z_{t-3}^{j-1}\right)^2\right)$ de la série initiale de l'EPA. Les exposants servent à désigner la vague à laquelle se rattachent les coefficients. Nous présentons les estimations des coefficients au tableau 4.1 avec la mesure R^2 corrigée de la qualité d'ajustement.

Tableau 4.1
Estimations de régression du processus des erreurs-types fondées sur le plan de sondage

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$\hat{c}$	12,219	-	-	-	-
$\hat{\beta}_j$	0,630	0,468	0,354	0,414	0,413
$\hat{\psi}_j$	-	0,717	0,786	0,749	0,751
$\hat{\sigma}_\varepsilon^j$	0,202	0,204	0,228	0,225	0,267
R^2_{adj}	0,351	0,373	0,386	0,477	0,342

La simulation se déroule de la manière suivante : pour chaque durée de série considérée et chaque simulation s, nous employons cinq signaux simulés $l_{t,s}^j,\ j = \{1, 2, 3, 4, 5\}$, pour produire cinq ensembles d'erreurs-types $z_{t,s}^j$ fondées sur le plan selon le processus défini en (4.1) et avec les coefficients de régression du tableau 4.1. Dès qu'un ensemble de données artificiel est produit, une estimation $\hat{\rho}_s$ est obtenue, après quoi le reste des hyperparamètres est estimé par la méthode du quasi-maximum de vraisemblance. À noter que le même ensemble d'erreurs-types $\mathbf{z}_{t,s}$ fondées sur le plan sert à produire toutes les séries bootstrap dans une simulation particulière.

Pour dégager les EQM réelles, nous mettons le modèle EPA en simulation un grand nombre de fois $(M = 50\,000)$, opération où chacune des itérations est assujettie aux mêmes limites que plus haut (entre zéro et un million de chômeurs). Nous calculons l'EQM réelle en prenant les valeurs réelles de vecteur d'état $\boldsymbol{\alpha}_{m,t}$ qui sont connues pour toute simulation m :

$$\mathbf{EQM}_t^{\text{Réel}} = \frac{1}{\mathrm{M}} \sum_{m=1}^{M} \left[\left(\hat{\boldsymbol{\alpha}}_{m,t}\left(\hat{\boldsymbol{\theta}}_m\right) - \boldsymbol{\alpha}_{m,t} \right) \left(\hat{\boldsymbol{\alpha}}_{m,t}\left(\hat{\boldsymbol{\theta}}_m\right) - \boldsymbol{\alpha}_{m,t} \right)' \right]. \quad (4.2)$$

L'EQM réelle du signal se calcule de la même manière à l'aide des valeurs de signal de la vague $\mathbf{l}_{m,t}$.

5 Résultats

5.1 Autres spécifications de modélisation pour l'EPA

On choisit et évalue habituellement les modèles SCS en employant des tests formels de diagnostic de normalité, d'homoscédasticité et d'indépendance des innovations normalisées. Une paramétrisation parcimonieuse est fondée sur des tests de rapport de vraisemblance logarithmique ou des critères

d'information (d'Akaike, de Bayes, etc.). Toutefois, les résultats de ces tests et critères dépendent des estimations ponctuelles particulières des hyperparamètres plutôt que de leurs distributions entières. Les distributions en simulation de Monte-Carlo (décrite à la section 4) des estimateurs des hyperparamètres nous éclairent davantage sur l'adéquation de la modélisation SCS. Les distributions en simulation nous livrent des indices sur l'éventuelle surspécification d'un modèle, en ce sens que certaines variables d'état pourraient être modélisées comme invariantes dans le temps.

Dans notre étude, nous considérons quatre modèles qui diffèrent pour le nombre d'hyperparamètres à estimer par la méthode du maximum de vraisemblance. Le modèle le plus complet, le modèle 1, est actuellement utilisé par Statistics Netherlands, mais après retrait de la composante de bruit blanc ε_t du paramètre réel de population ξ_t. On a constaté que cette composante avait une variance excessivement élevée et représentait une estimation perturbée d'autres hyperparamètres marginalement significatifs (variances de perturbation du BRE et de la composante saisonnière) dans le cas de l'EPA. En retranchant la composante irrégulière ε_t du modèle, on atténue l'instabilité des deux hyperparamètres précités. Cette formulation implique que le paramètre de population ξ_t n'accuse pas d'irrégularités impossibles à appréhender par la structure stochastique de la tendance et de la composante saisonnière. L'adoption de cette hypothèse peut être favorisée par une rigidité relative du marché du travail. L'évolution des niveaux de chômage est normalement progressive et doit donc être largement intégrée aux mouvements de la tendance stochastique. Les trois autres modèles sont des cas d'espèce du modèle 1, tous avec la composante irrégulière ε_t en moins (voir tableau 5.1).

Tableau 5.1
Hyperparamètres estimés dans les quatre versions du modèle EPA; les variances de perturbation sont estimées à l'échelle logarithmique

Modèles	Description	Paramètres estimés
M1	Modèle complet	$\rho, \sigma^2_{\eta_R}, \sigma^2_{\omega}, \sigma^2_{\eta_\lambda}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}, \sigma^2_{v_4}, \sigma^2_{v_5}$
M2	Modèle saisonnier indépendant du temps	$\rho, \sigma^2_{\eta_R}, \sigma^2_{\eta_\lambda}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}, \sigma^2_{v_4}, \sigma^2_{v_5}$
M3	Modèle BRE indépendant du temps	$\rho, \sigma^2_{\eta_R}, \sigma^2_{\omega}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}, \sigma^2_{v_4}, \sigma^2_{v_5}$
M4	Modèle saisonnier et BRE indépendant du temps	$\rho, \sigma^2_{\eta_R}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}, \sigma^2_{v_4}, \sigma^2_{v_5}$

Les distributions simulées des estimateurs des hyperparamètres dans le modèle 1 montrent que les hyperparamètres de variance pour la composante saisonnière et, en particulier, pour le BRE sont souvent estimés comme étant proches de zéro. Cela cause une bimodalité dans la distribution de ces estimations de variance avec une masse significative concentrée près de zéro. De plus, une tentative d'estimation de $\ln\left(\hat{\sigma}^2_{\omega}\right)$ ainsi que de $\ln\left(\hat{\sigma}^2_{\eta_\lambda}\right)$, comme dans le modèle 1, cause une distorsion dans la distribution des estimateurs de maximum de vraisemblance des autres hyperparamètres, laquelle devrait être normale. Ainsi, la normalité dans $\ln\left(\hat{\sigma}^2_{v_3}\right), \ln\left(\hat{\sigma}^2_{v_4}\right)$ et $\ln\left(\hat{\sigma}^2_{v_5}\right)$ est gravement compromise avec des valeurs aberrantes extrêmes et/ou un énorme coefficient d'applatissement (voir la figure A.1 en annexe où l'axe des x est étiré à cause des valeurs aberrantes), alors que les variances correspondantes sont moins susceptibles de présenter des valeurs extrêmes, étant censées fluctuer autour de l'unité. Si on rend la composante saisonnière invariante dans le temps comme dans le modèle 2, on ne change guère la situation des hyperparamètres de la tendance et du BRE. On pourrait même y voir un traitement moins qu'optimal, car les valeurs aberrantes

sont plus extrêmes et le coefficient d'applatissement est excessif dans la distribution des cinq hyperparamètres des erreurs d'enquête (figure A.2). Par contraste, nous avons pu constater (voir les figures A.3 et A.4) que, dans les deux modèles où la composante BRE est fixe dans le temps (modèles 3 et 4), toutes les estimations des hyperparamètres correspondant aux erreurs d'enquête étaient en distribution normale. Dans le modèle 3, les distributions demeurent asymétriques pour la pente et la composante saisonnière (asymétrie de -0,88 et -0,72 et applatissement de 5,56 et 4,61 respectivement). En fixant à zéro l'hyperparamètre saisonnier dans le modèle 4, l'amélioration est seulement marginale et la distribution de $\ln\left(\hat{\sigma}^2_{\eta_R}\right)$ présente un coefficient négatif d'asymétrie (-0,81) et un coefficient excessif d'applatissement (1,76).

Ces données de simulation semblent indiquer que, dans la modélisation des séries EPA, la préférence pourrait aller au modèle 3 plus parcimonieux, où la seule variance de perturbation BRE est fixée à zéro, mais comme le BRE même dépend du nombre de chômeurs, Statistics Netherlands conserve la variance de cet hyperparamètre à des fins de production afin de garder une souplesse suffisante devant l'évolution progressive du processus sous-jacent.

On peut recourir au test du rapport de vraisemblance pour vérifier si les hyperparamètres de la composante saisonnière et du BRE sont significativement différents de zéro, les modèles 2 à 4 étant imbriqués dans le modèle 1. La variable à tester comporte des valeurs très basses pour les trois autres modèles (0; 0,18 et 0,18 encore pour les modèles 2, 3 et 4, l'absence de différences entre les modèles 2 et 1 et entre les modèles 3 et 4 étant attribuable à la très faible valeur de l'hyperparamètre de la composante saisonnière). Ainsi, ces tests n'indiquent pas que les modèles plus parcimonieux présentent des résultats inférieurs à ceux du modèle 1. Une autre façon d'évaluer l'adéquation des quatre modèles est de les comparer sous l'angle de leur valeur prévisionnelle par la racine carrée des différences quadratiques moyennes (RDQM) entre les estimations ERG et les prédictions des signaux à un pas avant. On peut le faire pour chaque vague séparément : $\text{RDQM}^j = 1/(T-d)\sum_{t=d}^{T}\left(\hat{l}^j_{t|t-1} - Y_t^j\right)^2$, d étant égal à 20, 30 et 60 mois. Les résultats figurant en annexe (tableau B.1) montrent cependant qu'il n'y a guère de différence de rendement des quatre modèles dans leur application à la série initiale. Les modèles plus parcimonieux font voir une légère augmentation de la RDQM.

Les reformulations de modèle ne semblent pas influer sur la distribution de l'estimateur du paramètre autorégressif ρ des erreurs d'enquête sur les 1 000 séries simulées : on approche d'assez près la distribution normale et les valeurs vont de 0 à 0,4 quand $T = 114$, ce qui s'accorde avec l'approximation de sa distribution asymptotique à la sous-section 3.3. L'intervalle des valeurs est un peu plus étendu pour les séries temporelles plus courtes et plus étroites quand $T = 200$. Nous exécutons séparément pour les quatre modèles la procédure de simulation décrite dans la section précédente et l'analyse des méthodes bootstrap.

5.2 Estimation EQM

L'objet de notre étude par simulation est l'estimation EQM de la tendance et du signal de population, ce dernier étant la somme de la tendance et de la composante saisonnière. Nous évaluons le rendement du filtre de Kalman et des cinq méthodes d'estimation EQM à la section 3 en considérant le biais relatif et les EQM des estimateurs EQM. D'abord, nous prenons la moyenne des estimations EQM filtrées en (3.3), (3.4) et

(3.7) sur les 1 000 simulations (la moyenne est indiquée par la barre sur $\overline{\text{EQM}}_{t|t}$), alors que, dans le cas des estimations EQM par filtre de Kalman, nous l'établissons sur 10 000 simulations, comme nous l'avons mentionné au début de la section 4. Ces estimations EQM filtrées et mises en moyenne pour le modèle 3 (sauf pour la méthode AA; voir l'explication plus loin) sont décrites aux figures 5.1 à 5.4 pour $T = 48$, $T = 80$, $T = 114$ et $T = 200$ respectivement. Nous sautons les $d = 30$ premiers points temporels de l'échantillon (d devrait dépasser le nombre de points temporels nécessaires au début de la série pour éliminer l'effet d'une initialisation diffuse par le filtre). À noter que l'analyse est fondée sur des estimations filtrées plutôt que lissées, car ce sont les premières qui reproduisent le mieux le processus de production des chiffres officiels. Les EQM des quatre figures sont en configuration décroissante, comme on pouvait s'y attendre, parce que des estimations filtrées augmentent en précision si on dispose de plus d'information dans le temps pour estimer les variables d'état. Une exception à la règle, ce sont les EQM réelles de la figure 5.2. Une explication possible est que, dans cette application, les EQM des signaux sont proportionnelles aux signaux mêmes par les erreurs-types fondées sur le plan et que les EQM réelles reposent sur un autre ensemble (bien plus étendu) de séries simulées (50 000 pour les EQM réelles et 1 000 pour les EQM estimées). On remarquera que les traits de la figure 5.1 paraissent bien plus lisses, puisqu'ils s'étendent sur moins de points temporels. Ajoutons que, dans les figures 5.2 et 5.3, la configuration semble plus irrégulière, l'échelle de l'axe des y étant plus fine si on compare ces figures aux figures 5.1 et 5.4.

Nous calculons le biais relatif en pourcentage comme $\text{BR}_t^f = 100\%\left(\overline{\text{EQM}}_{t|t}^f \Big/ \text{EQM}_{t|t}^{\text{Réel}} - 1\right)$, où f correspond à une méthode d'estimation particulière et où $\text{EQM}_{t|t}^{\text{Réel}}$ est défini en (4.2). Les biais EQM relatifs en pourcentage et en moyenne dans le temps (après retrait des $d = 30$ premiers points temporels) pour le signal, la tendance et la composante saisonnière sont présentés aux tableaux 5.2, 5.3, 5.4 et 5.5.

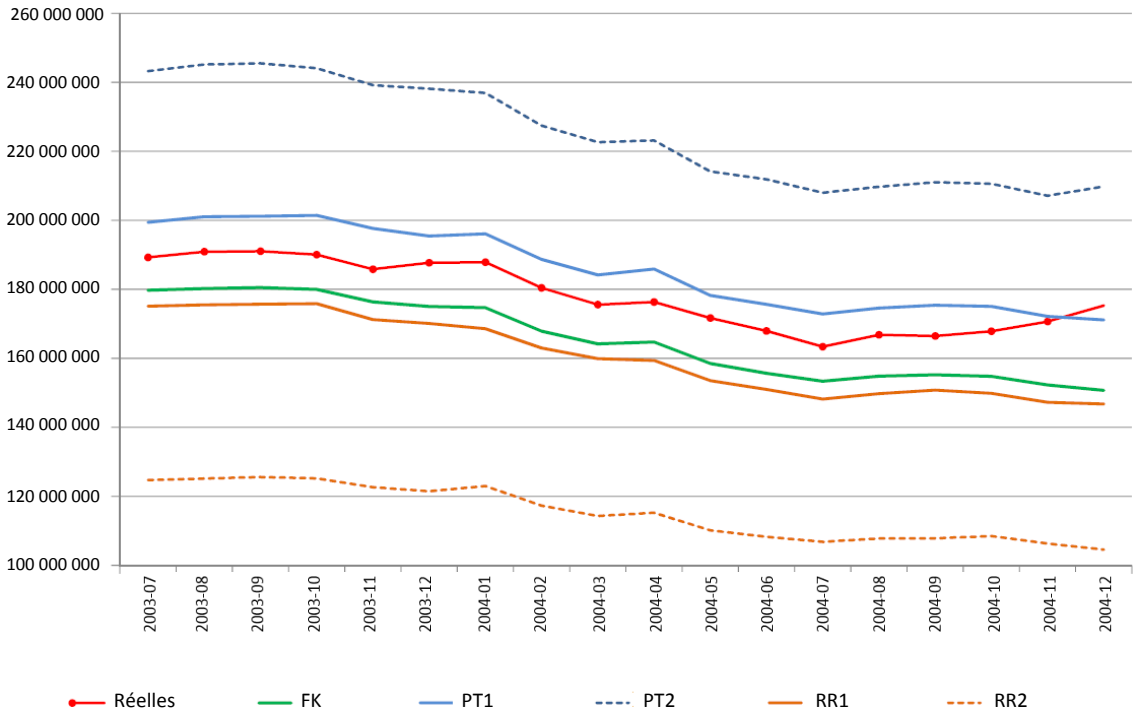

Figure 5.1 EQM réelles et EQM estimées moyennes pour le paramètre réel de population filtré (tendance et composante saisonnière) dans le modèle 3, $T = 48$ mois.

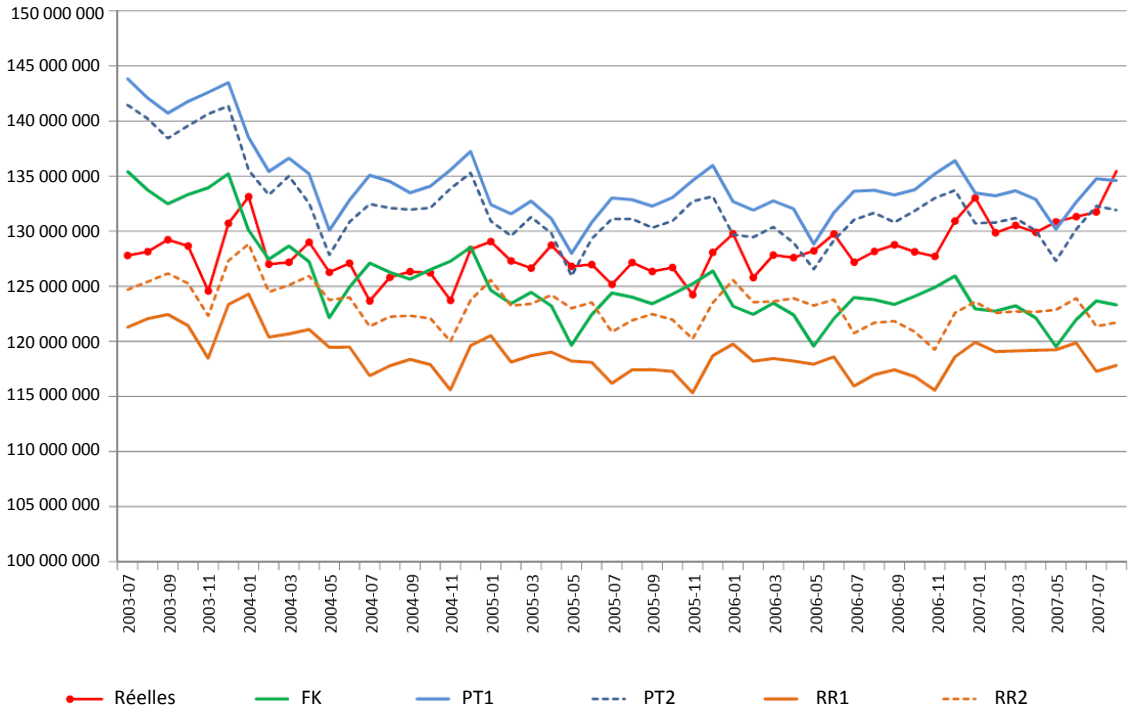

Figure 5.2 EQM réelles et EQM estimées moyennes pour le paramètre réel de population filtré (tendance et composante saisonnière) dans le modèle 3, $T = 80$ mois.

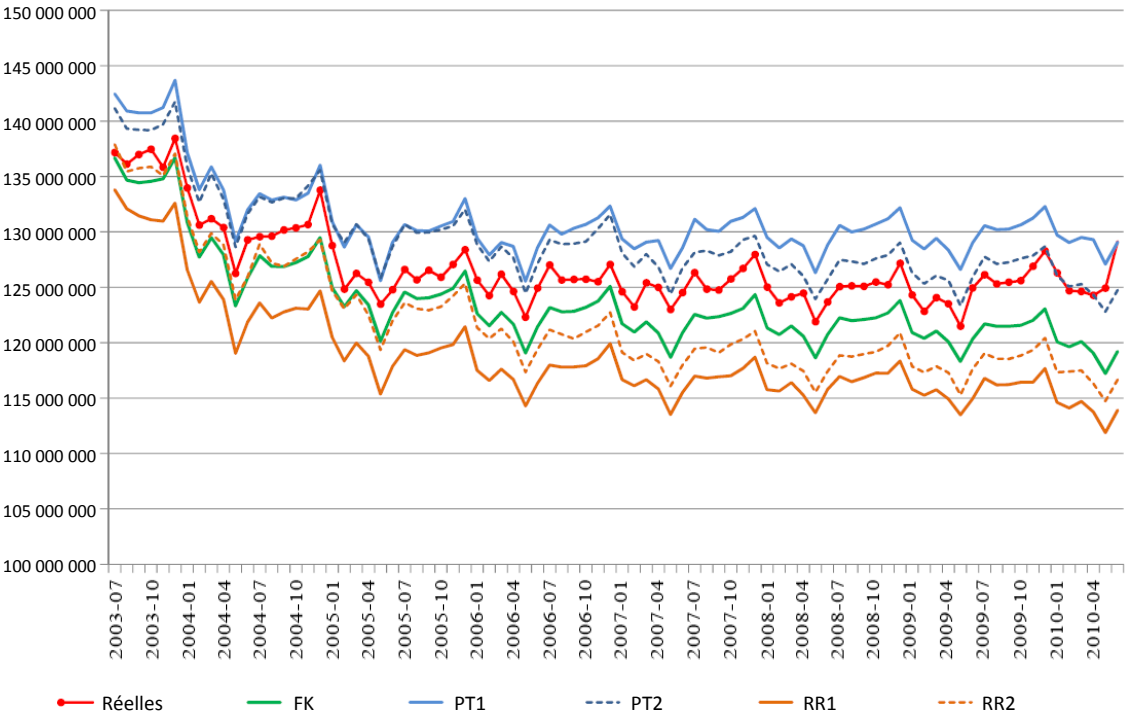

Figure 5.3 EQM réelles et EQM estimées moyennes pour le paramètre réel de population filtré (tendance et composante saisonnière) dans le modèle 3, $T = 114$ mois.

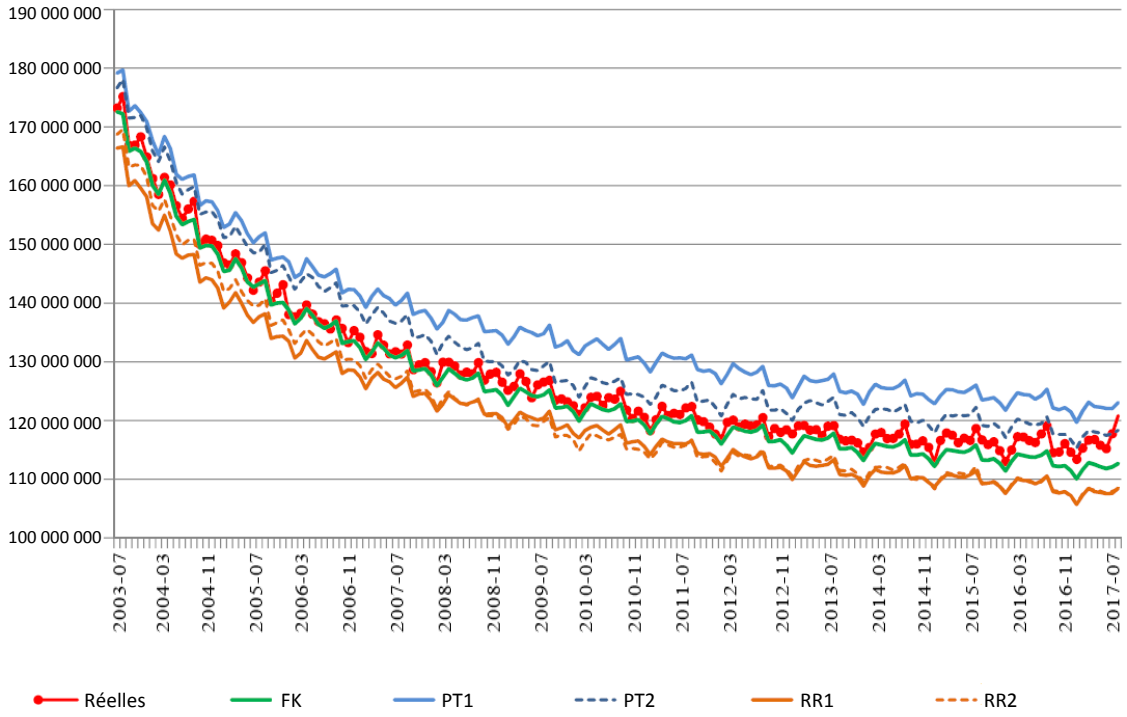

Figure 5.4 EQM réelles et EQM estimées moyennes pour le paramètre réel de population filtré (tendance et composante saisonnière) dans le modèle 3, $T = 200$ mois.

Tableau 5.2
Biais moyen en pourcentage des estimateurs EQM dans le modèle de l'EPA, $t = \{31,\ldots,T\}, T = 48$

Modèles	Signal*				Tendance				Composante saisonnière			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
FK	S.O.	S.O.	-7,1	-7,6	S.O.	S.O.	-6,5	-6,6	S.O.	S.O.	-6,7	-7,0
PT1	S.O.	S.O.	*4,4*	*1,4*	S.O.	S.O.	8,7	6,4	S.O.	S.O.	4,9	2,4
PT2	S.O.	S.O.	26,2	-4,4	S.O.	S.O.	22,4	-3,1	S.O.	S.O.	25,6	-4,6
RR1	S.O.	S.O.	-9,8	-10,8	S.O.	S.O.	-13,9	-13,8	S.O.	S.O.	-9,5	-10,1
RR2	S.O.	S.O.	-35,3	-5,6	S.O.	S.O.	-29,9	-3,2	S.O.	S.O.	-29,7	-5,1

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.3
Biais moyen en pourcentage des estimateurs EQM dans le modèle de l'EPA, $t = \{31,\ldots,T\}$, $T = 80$

Modèles	Signal*				Tendance				Composante saisonnière			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
FK	-3,0	-3,2	-2,1	-2,2	-3,5	-3,8	-2,5	-2,5	8,8	2,5	2,9	2,4
AA	S.O.	S.O.	S.O.	14,9	S.O.	S.O.	S.O.	15,0	S.O.	S.O.	S.O.	14,9
PT1	8,6	6,7	4,9	6,2	10,6	8,9	7,1	8,4	20,8	10,7	10,3	11,1
PT2	*4,8*	*3,7*	*1,4*	*2,1*	4,8	4,9	2,1	2,3	17,3	8,2	6,9	7,1
RR1	-7,2	-9,0	-7,3	-7,2	-9,6	-11,2	-9,6	-9,5	-3,8	-9,0	-6,7	-6,6
RR2	6,7	-3,5	-3,9	-4,2	5,3	-4,1	-4,6	-5,4	18,6	-4,7	-4,1	-4,3

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.4
Biais moyen en pourcentage des estimateurs EQM dans le modèle de l'EPA, $t = \{31,\ldots,T\}$, $T = 114$

Modèles	Signal*				Tendance				Composante saisonnière			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
FK	-2,1	-2,6	-2,4	-2,2	-2,3	-2,7	-2,4	-2,3	2,5	-3,2	-3,1	-2,6
AA	S.O.	S.O.	S.O.	5,2	S.O.	S.O.	S.O.	4,1	S.O.	S.O.	S.O.	12,5
PT1	8,1	5,7	3,3	5,5	10,0	7,9	5,2	7,6	4,9	1,4	1,4	0,3
PT2	*2,2*	*3,2*	*1,9*	*1,5*	3,3	4,3	3,1	2,8	1,2	-2,0	1,0	0,6
RR1	-8,3	-7,8	-6,4	-6,5	-10,7	-9,9	-8,7	-8,9	-3,1	-7,2	-5,5	-5,6
RR2	-1,1	-6,0	-3,9	-3,5	-3,0	-7,6	-5,5	-5,0	7,3	-5,9	-3,2	-3,0

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.5
Biais moyen en pourcentage des estimateurs EQM dans le modèle de l'EPA, $t = \{31,\ldots,T\}$, $T = 200$

Modèles	Signal*				Tendance				Composante saisonnière			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
FK	-1,3	-1,6	-1,3	-1,3	-1,7	-1,8	-1,6	-1,6	3,8	-1,7	-1,6	-1,6
AA	S.O.	S.O.	S.O.	5,9	S.O.	S.O.	S.O.	5,6	S.O.	S.O.	S.O.	5,6
PT1	*6,3*	6,2	6,3	5,5	7,5	7,7	7,8	7,1	10,8	2,6	3,0	3,0
PT2	6,8	*4,0*	*3,0*	*2,3*	7,6	4,9	4,2	3,6	12,5	2,1	1,3	0,6
RR1	-8,0	-8,0	-4,9	-5,9	-10,0	-9,9	-6,8	-7,1	-1,1	-5,3	-3,8	-3,9
RR2	-5,1	-5,6	-4,5	-5,0	-7,0	-7,4	-6,0	-6,4	3,6	-3,1	-3,3	-3,9

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.6
Variance estimée moyenne des EQM des estimateurs EQM pour le nombre de chômeurs dans le modèle de l'EPA (division par 10^{15}), $t = \{31,\ldots,T\}$, $T = 48$

Modèles	Signal*				Tendance				Composante saisonnière			
	M3		M4		M3		M4		M3		M4	
	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}
PT1	3,39	3,46	3,64	3,66	3,61	3,83	3,67	3,81	0,59	0,61	0,64	0,65
PT2	5,03	7,26	3,03	3,10	4,02	5,27	2,56	2,61	1,00	1,50	0,52	0,54
RR1	2,51	2,83	2,68	3,06	2,03	2,51	2,13	2,62	0,44	0,51	0,48	0,55
RR2	1,59	5,93	2,74	2,85	1,52	3,97	2,50	2,56	0,55	1,28	0,50	0,52

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.7
Variance estimée moyenne des EQM des estimateurs EQM pour le nombre de chômeurs dans le modèle de l'EPA (division par 10^{15}), $t = \{31,\ldots,T\}$, $T = 80$

Modèles	Signal*				Tendance				Composante saisonnière			
	M3		M4		M3		M4		M3		M4	
	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}
PT1	2,24	2,29	2,43	2,52	1,82	1,91	1,97	2,09	0,27	0,30	0,27	0,31
PT2	2,20	2,23	2,14	2,18	1,71	1,74	1,66	1,69	0,27	0,28	0,27	0,29
RR1	1.86	1,95	1,74	1,82	1,42	1,56	1,33	1,46	0,22	0,23	0,22	0,23
RR2	1,98	2,01	1,94	1,97	1,57	1,60	1,49	1,54	0,23	0,23	0,23	0,23

* Le signal est la somme de la tendance et de la composante saisonnière.

Voici les principales conclusions de notre étude par simulation :

1. Pour $T = 48$ et en moyenne dans le temps (à partir de $t = 31$), le biais relatif de l'EQM du signal après application du filtre de Kalman est d'environ -7 %. Ce biais tend à décroître à mesure que s'allonge la série. Le biais de filtre de Kalman (FK) est des plus modestes quand $T = 200$ et la situation est telle qu'aucune des méthodes d'estimation n'offre d'amélioration par rapport aux estimations EQM par filtre de Kalman. Nous pourrions toujours appliquer la meilleure méthode d'estimation avec des biais positifs pour dégager une plage de valeurs contenant l'EQM réelle.

2. Nous avons pu voir que la méthode AA (approximation asymptotique) est inapplicable aux modèles comportant des hyperparamètres marginalement significatifs. Quand on estime que certains des hyperparamètres sont proches de zéro, la matrice $\mathbf{I}^{-1}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}} \middle| \rho^{a}\right)$ est numériquement singulière, d'où un échec de la procédure, ou quasi singulière. Dans ce dernier cas, la variance asymptotique devient excessivement élevée et perd donc toute fiabilité. Cela étant dit, la méthode AA serait uniquement envisageable pour le modèle 4. Comme on pouvait s'y attendre, la méthode donne de piètres résultats avec de courtes séries et laisse des biais positifs d'environ 15 %. Le rendement pour $T = 114$ et $T = 200$ est comparable à celui de la méthode bootstrap PT1, mais demeure significativement inférieur à celui de la méthode PT2.

3. Comme on peut immédiatement l'observer, l'emploi du bootstrap RR crée un biais négatif contrairement au bootstrap PT qui engendre un biais positif. À l'encontre de l'affirmation faite par Rodriguez et Ruiz (2012) que leur méthode offre de meilleures propriétés d'échantillon fini que la méthode de Pfeffermann et Tiller (2005), nous pouvons voir dans le cas de l'EPA que les estimations EQM par le bootstrap RR paramétrique ou non créent des biais négatifs plus importants que les estimations EQM par filtre de Kalman à l'échelle des modèles et des longueurs de séries (sauf pour RR2 dans le modèle 4 quand $T = 48$ et dans le modèle 1 quand $T = 80$ et $T = 114$). Alors que Pfeffermann et Tiller (2005) démontrent que leur méthode bootstrap présente des propriétés asymptotiques satisfaisantes, Rodriguez et Ruiz (2012) illustrent la supériorité de leur méthode dans de petits échantillons avec un modèle simple (à marche aléatoire et à bruit). La présente étude par simulation révèle que le bootstrap RR pourrait mal se comporter dans des applications plus complexes. Les méthodes PT n'ont jamais créé de biais négatifs pour l'EPA, ce qui en établit la « prudence » (sauf pour le bootstrap PT2 dans le modèle 4 quand $T = 48$ où le biais négatif demeure inférieur à celui de l'application du filtre de Kalman). Un autre résultat frappant pour $T = 48$ est que le biais positif du bootstrap PT2 et le biais négatif du bootstrap RR prennent des valeurs très élevées dans le modèle 3. Il reste que, avec une série si courte et autant de composantes non stationnaires comme dans le modèle de l'EPA, il est difficile de tirer des estimations fiables des méthodes bootstrap non paramétriques, puisque la période d'initialisation (avec son échantillon diffus) nécessaire à la production non paramétrique d'une série prend plus du quart de sa durée (13 mois sur 48).

4. Pour les séries de longueur $T = 114$ et $T = 80$, les biais positifs engendrés par la méthode PT2 dépassent légèrement les biais FK en valeur absolue dans les modèles comportant des hyperparamètres non significatifs (modèles 1 et 2). Dans les modèles plus stables (modèles 3 et 4), les biais positifs sont inférieurs aux biais négatifs FK en valeur absolue. Pour $T = 48$, nous présentons les résultats bootstrap seulement pour les modèles 3 et 4 (nous ne tenons pas compte des modèles 1 et 2 qui tendent à la surspécification à cause de problèmes numériques). Comme on pouvait s'y attendre, les biais sont plus importants pour une

telle durée des séries : les biais négatifs FK et RR s'accroissent en valeur absolue, tout comme les biais positifs PT, sauf pour le résultat PT2 précité dans le modèle 4.

L'EQM du signal dans le modèle 3, que nous pourrions considérer comme un meilleur choix pour la production des chiffres officiels de l'EPA, est estimée au mieux par la méthode PT2 avec des biais relatifs de 1,4 % et 1,9 % respectivement pour $T = 80$ et $T = 114$. Le bootstrap PT2 serait aussi la meilleure méthode pour $T = 200$, mais comme nous l'avons fait observer, les biais négatifs FK sont déjà des plus modestes pour des séries de cette longueur. Dans le cas de séries très courtes comme $T = 48$, le bootstrap PT1 paramétrique serait le meilleur.

5. Pour les méthodes PT et RR à la fois (sauf pour RR2 dans le modèle 4 avec $T = 48$), les valeurs absolues des biais relatifs sont moindres dans le cas des méthodes non paramétriques par rapport aux méthodes paramétriques. La supériorité du bootstrap non paramétrique peut s'expliquer par une distorsion de la normalité de la distribution des erreurs dans les modèles. Ainsi, notre préférence devrait aller aux bootstraps non paramétriques sauf pour des séries chronologiques très courtes.

6. Il n'y a pas que le biais des estimateurs EQM, puisque leur variabilité nous éclaire grandement aussi sur leur fiabilité. Autant que nous sachions, cet aspect n'a pas encore été exposé dans les études statistiques. Les tableaux 5.6 et 5.7 présentent les variances et les EQM des quatre estimateurs EQM bootstrap pour le signal, la tendance et la composante saisonnière dans le cas des longueurs de série les plus intéressantes, à savoir $T = 48$ et $T = 80$ (nous ne tenons pas compte des modèles 1 et 2, ni de l'approximation asymptotique en raison des problèmes numériques déjà évoqués). Les EQM des deux estimateurs EQM PT sont plus élevées que celles des deux estimateurs EQM RR tant pour le modèle 3 que pour le modèle 4. Si ces derniers semblent d'un rendement supérieur, comme en témoigneraient leurs EQM moindres, c'est que leurs variances sont plus petites. Toutefois, les biais sont parfois assez élevés pour porter les EQM de ces estimateurs EQM presque au niveau des EQM des estimateurs PT. Plus important encore, les biais des estimateurs EQM RR sont le plus souvent négatifs et dépassent fréquemment ceux des estimateurs par filtre de Kalman. Ce phénomène rend les bootstraps RR difficilement applicables dans le cas qui nous occupe.

Outre les résultats de simulation déjà mentionnés, il est également intéressant de voir si les modèles de séries chronologiques structurels (SCS) continuent d'offrir des estimations plus précises que les estimations de variance fondées sur le plan, même après correction de l'incertitude des hyperparamètres. C'est pourquoi nous mettons en comparaison les racines des EQM (REQM) obtenues avec les différentes procédures d'estimation EQM pour la série initiale $(T = 114)$, d'une part, et les erreurs-types (ET) de l'estimateur ERG. De telles différences moyennes des erreurs-types (DMET) dans le modèle m des séries chronologiques $(m = \{1,2,3,4\})$ se définissent ainsi : $\mathrm{DMET}_m^f = 100\,\%/(T-d)\sum_{t=d}^{T}\left[\mathrm{REQM}^f\left(\hat{l}_{t|t}^m\right) - \mathrm{ET}(Y_t)\right]/\mathrm{ET}(Y_t)$. Elles sont présentées au tableau 5.8, $\hat{l}_{t|t}^m$ étant l'estimation filtrée du paramètre réel de population défini comme la tendance et la composante saisonnière dans le modèle m. Nous décrivons les résultats pour le filtre de Kalman (FK) quand nous négligeons l'incertitude des hyperparamètres, ainsi que dans les cas où les cinq méthodes d'estimation EQM sont appliquées dans une prise en compte de cette même incertitude. Nous comparons aussi les REQM réelles en (4.2) aux erreurs-types ERG (« Réel » en ligne au tableau 5.8). À noter que le BRE et, en particulier, les estimations saisonnières des hyperparamètres par l'ensemble de données initial de l'EPA sont plutôt petits. Il n'y a donc pas de différences dignes de mention entre les

estimations ponctuelles du signal dans les quatre modèles. La méthode AA, la moins sûre, produit des erreurs-types surestimées (par rapport à la diminution de 18 % à 20 % pour les REQM réelles) à cause des matrices d'information quasi singulières des estimations de maximum de vraisemblance des hyperparamètres. Vu ce phénomène, on devrait se sentir plus en confiance dans l'utilisation des estimateurs PT. Bien que notre étude par simulation indique que le bootstrap PT2 est normalement d'un meilleur rendement que le bootstrap paramétrique PT1, pour cette série en particulier les ET dégagées par le bootstrap PT1 sont les plus proches des REMQ réelles avec une diminution d'environ 20 % des erreurs-types de l'estimation ERG. Ainsi, la modélisation permet une baisse significative de la variance comparativement à une approche plus classique fondée sur le plan, et ce, même après avoir pris en compte l'incertitude des hyperparamètres.

Tableau 5.8
Différences moyennes en pourcentage des erreurs-types (DMET) entre les estimateurs par la régression généralisée et les estimateurs de modélisation pour la série initiale de l'EPA, $d = 30$; augmentation en pourcentage des ET par filtre de Kalman après application de la correction EQM (entre parenthèses)

	Modèle 1	Modèle 2	Modèle 3	Modèle 4
FK	-24,1	-24,1	-24,5	-24,5
Valeur réelle	-20,0 (5,56)	-20,1 (5,5)	-20,6 (5,4)	-20,7 (5,3)
AA	-18,8 (6,9)	-19,0 (6,7)	-19,1 (7,1)	-19,5 (6,6)
PT1	-20,1 (5,2)	-20,1 (5,2)	-21,1 (4,6)	-21,2 (4,4)
PT2	-22,9 (1,6)	-21,2 (3,8)	-22,2 (3,1)	-22,5 (2,6)
RR1	-26,5 (-3,2)	-26,6 (-3,4)	-26,5 (-2,7)	-26,5 (-2,7)
RR2	-24,0 (-0,1)	-25,4 (-1,8)	-25,6 (-1,4)	-25,7 (-1,6)

6 Observations en conclusion

Les organismes nationaux de statistique s'intéressent de plus en plus à l'utilisation de modèles de séries chronologiques structurels (SCS) pour la production des chiffres mensuels de la population active. Aux Pays-Bas, un tel modèle est appliqué depuis 2010. Le modèle SCS représente une sorte d'estimation sur petits domaines (EPD) où l'information tirée d'échantillons de périodes antérieures permet d'obtenir des estimations plus précises, et de tenir compte du plan de sondage avec renouvellement de panel, lequel est souvent employé dans les enquêtes sur la population active.

Si l'on ne tient pas compte de l'incertitude des hyperparamètres dans les EQM des estimations fondées sur des modèles SCS, on se trouve à sous-estimer les EQM des estimations de domaines. Le biais qui se crée lorsqu'on écarte ainsi l'incertitude des hyperparamètres peut être important, plus particulièrement quand les séries sont courtes, ce qui est souvent le cas dans les organismes nationaux de statistique. La plupart des applications des procédures EPD dans les études spécialisées reposent sur des modèles multiniveaux, pratique courante lorsqu'il s'agit de tenir compte de l'incertitude des hyperparamètres. Les études consacrées au modèles SCS dans le contexte des estimations sur petits domaines sont plutôt limitées et la plupart des applications ne tiennent pas compte de cette incertitude dans les estimations EQM. L'importance du biais dans les EQM obtenues dépend de la structure du modèle et de la longueur de la série. Le présent article décrit une simulation de Monte-Carlo appliquée au modèle SCS qu'utilise Statistics Netherlands pour estimer le chômage mensuel. Cette simulation a un double but. D'abord, elle établit la

quantité de biais dans les EQM de l'EPA quand on néglige l'incertitude des hyperparamètres. De plus, nous comparons notre simulation à plusieurs méthodes d'estimation EQM disponibles dans la documentation spécialisée pour le cadre de modèles SCS et établissons ainsi la meilleure méthode pour l'EPA des Pays-Bas. En deuxième lieu, nous jugeons que la simulation des distributions des estimateurs des hyperparamètres permet de mieux comprendre la dynamique des composantes inobservées de le modèle SCS et donc de vérifier la nécessité de modéliser les composantes comme variant dans le temps. Dans le cas de l'EPA, la simulation fait voir l'intérêt éventuel d'adopter une version plus restreinte du modèle où le biais de renouvellement de l'échantillon serait invariant dans le temps et où le bruit blanc de population serait négligé. Pour cette double raison, nous recommandons d'effectuer une simulation comme celle que nous décrivons dans le processus de mise en œuvre du modèle servant à la production des statistiques officielles.

La comparaison des méthodes d'estimation EQM jette en outre un nouvel éclairage sur leurs propriétés. L'approximation asymptotique est inapplicable aux cas où les hyperparamètres sont proches de zéro, parce que la matrice d'information des estimations des hyperparamètres devient (presque) singulière. Les bootstraps non paramétriques, parce qu'ils dépendent moins d'hypothèses de normalité, sont d'un meilleur rendement que les bootstraps paramétriques selon Pfeffermann et Tailler (2005) et Rodriguez et Ruiz (2012) à la fois sauf si les séries sont très courtes. Notre constatation première est que les bootstraps PT présentent des biais positifs et sont invariablement d'un rendement supérieur à celui des bootstraps RR dont les biais sont généralement négatifs et plus importants (en valeur absolue) que dans l'application du filtre de Kalman. Elle contredit Rodriguez et Ruiz (2012) qui affirment la supériorité de leur méthode lorsque les séries chronologiques sont courtes. On peut penser que leurs résultats sont purement heuristiques, étant fondés sur un modèle simple (marche aléatoire et bruit), alors que Pfeffermann et Tiller (2005) démontrent que leur méthode bootstrap produit des estimations EQM avec un biais d'un bon ordre.

Les variances des estimateurs EQM PT sont plus élevées que celles des estimateurs RR correspondants. Les différences entre ces deux types d'estimateurs varient de modestes à modérées (les EQM des seconds sont inférieures de 28 % à 8 % aux EQM des premiers selon le modèle et la longueur de la série). Aspect plus important encore, la tendance des estimateurs RR à engendrer des biais négatifs parfois supérieurs à ceux de l'application du filtre de Kalman rend inapplicables ces méthodes bootstrap. Ainsi, on devrait généralement envisager de recourir aux méthodes PT pour d'autres données d'enquête, quoique leur rendement le cède occasionnellement à celui des méthodes RR.

Dans le cas des séries chronologiques très courtes, les bootstraps non paramétriques ne seraient pas un choix possible pour un modèle qui aurait la complexité que nous présentons. Il reste que le bootstrap paramétrique PT corrige les EQM aux biais négatifs jusqu'à dégager un léger biais positif (de 1,4 % à 4,4 % selon le modèle). Pour la présente durée de série de 114 mois, il est possible d'abaisser de -2,4 % à 1,9 % le biais EQM négatif grâce à la méthode non paramétrique de Pfeffermann et Tiller (2005) dans le modèle où le BRE est invariant dans le temps. Les racines des EQM réelles par filtre de Kalman sont inférieures d'environ 20 % aux erreurs-types des estimations ERG dans les quatre modèles appliqués aux données de l'EPA. En général, les biais des estimations EQM par filtre de Kalman sont relativement modestes dans l'application de l'EPA, aussi parraîtrait-il suffisant de s'en remettre à ces estimations naïves pour la publication des chiffres officiels.

Remerciements

Nous remercions Statistics Netherlands d'avoir financé cette étude. Nous remercions également le rédacteur adjoint et les examinateurs anonymes d'avoir lu attentivement notre manuscrit et formulé de précieuses observations. Les points de vue exprimés dans la présente sont ceux des auteurs et ne reflètent pas nécessairement les politiques de Statistics Netherlands.

Annexes

A. Densités simulées des hyperparamètres dans les quatre versions du modèle de l'EPA

Nous présentons en annexe les fonctions de densité des hyperparamètres obtenues par simulation quand les quatre versions du modèle de l'EPA (voir le tableau 5.1) servent de processus de génération de données. L'axe des x présente les hyperparamètres de variance à l'échelle logarithmique et l'axe des y porte les valeurs de fréquence. L'axe des x peut être étiré à cause des valeurs aberrantes.

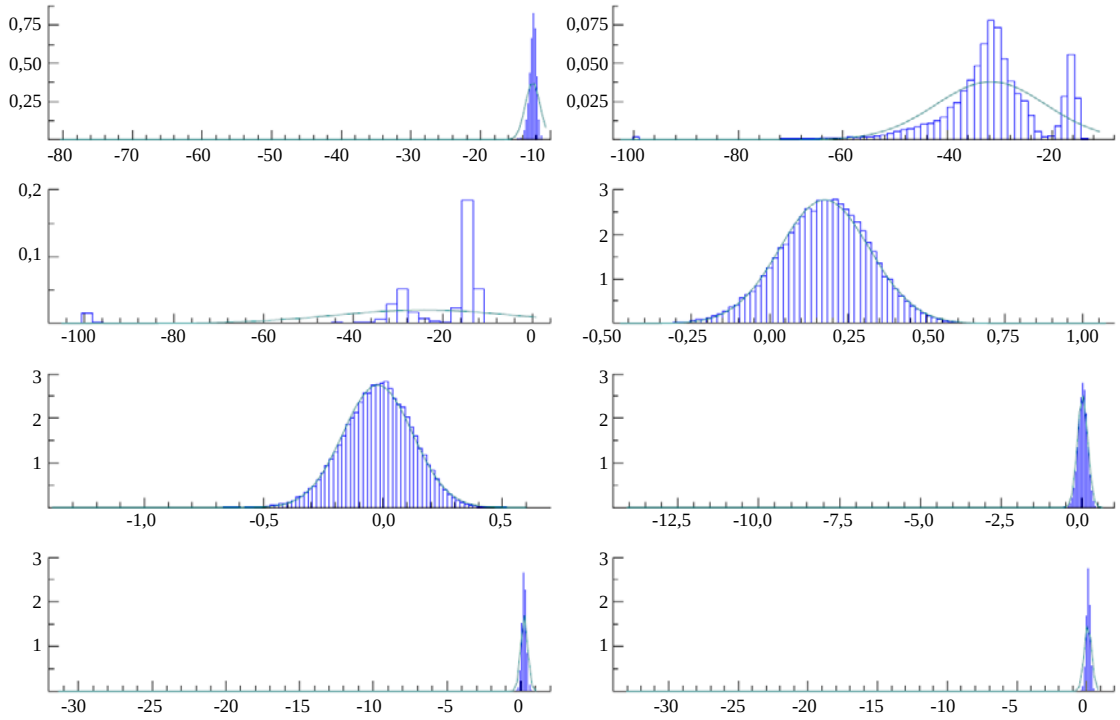

Figure A.1 Distribution des hyperparamètres sous le modèle complet de l'EPA (modèle 1), de gauche à droite sur l'axe des x : $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_\gamma)$, $\ln(\hat{\sigma}^2_\lambda)$, $\ln(\hat{\sigma}^2_{v_t^t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; densité normale avec les mêmes moyenne et variance superposée; 50 000 simulations, $T = 114$.

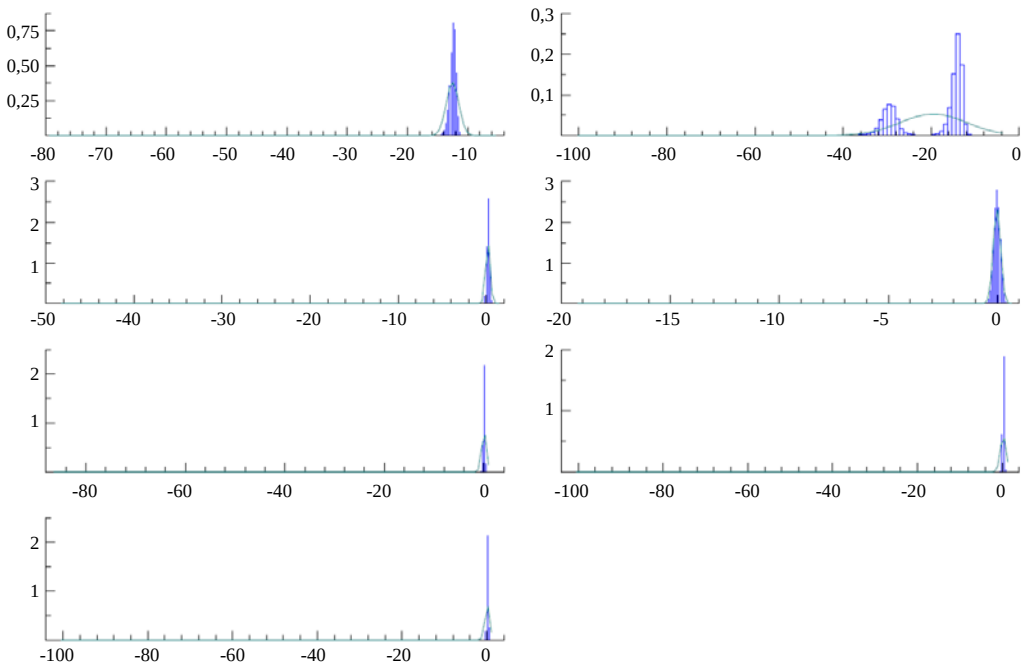

Figure A.2 Distribution des hyperparamètres sous le modèle 2, de gauche à droite sur l'axe des x : $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_\lambda)$, $\ln(\hat{\sigma}^2_{v_t^t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; densité normale avec les mêmes moyenne et variance superposée; 50 000 simulations, $T = 114$.

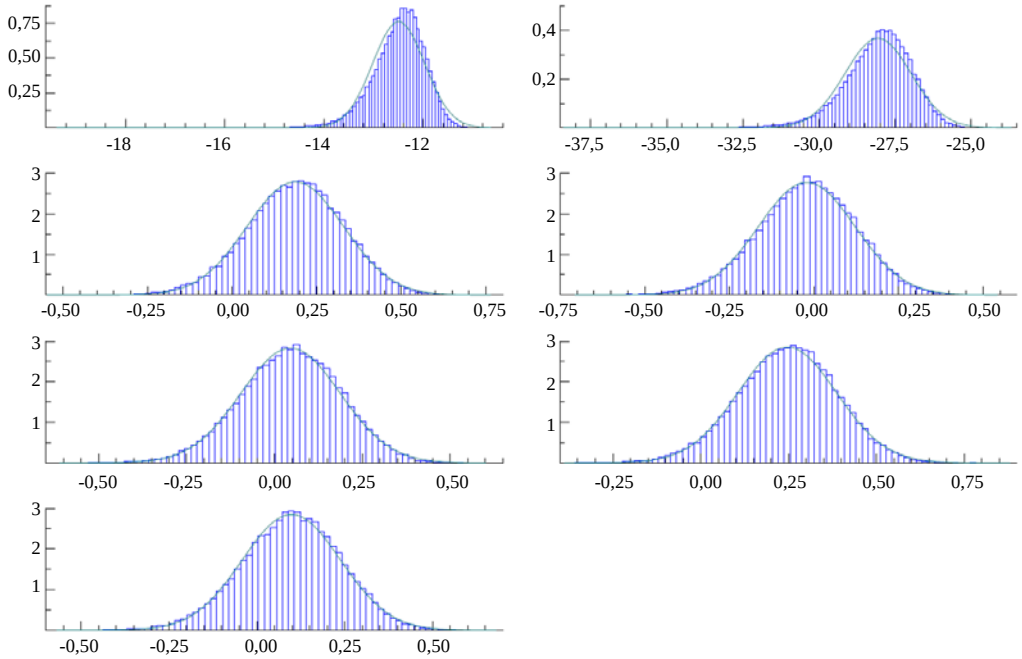

Figure A.3 Distribution des hyperparamètres sous le modèle 3, de gauche à droite sur l'axe des x : $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_\gamma)$, $\ln(\hat{\sigma}^2_{v_t^t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; densité normale avec les mêmes moyenne et variance superposée; 50 000 simulations, $T = 114$.

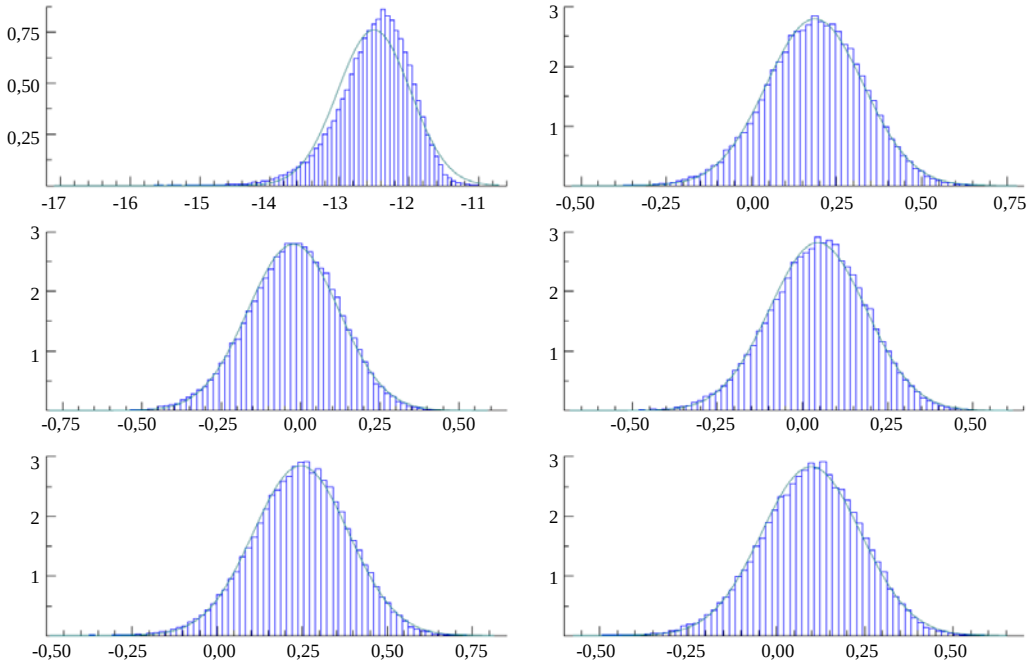

Figure A.4 Distribution des hyperparamètres sous le modèle 4, de gauche à droite sur l'axe des x : $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_{v_t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; densité normale avec les mêmes moyenne et variance superposée; 50 000 simulations, $T = 114$.

B. Rendement prévisionnel des quatre modèles de l'EPA

Tableau B.1
Racine des écarts quadratiques moyens des estimations par la régression généralisée du nombre de chômeurs par prédiction « un pas avant » et par vague

Vague	Modèle 1			Modèle 2			Modèle 3			Modèle 4		
	$d = 20$	$d = 30$	$d = 60$	$d = 20$	$d = 30$	$d = 60$	$d = 20$	$d = 30$	$d = 60$	$d = 20$	$d = 30$	$d = 60$
1	34 370	33 582	34 641	34 370	33 582	34 641	34 518	33 754	34 881	34 525	33 757	34 885
2	30 130	29 770	29 410	30 130	29 770	29 410	30 138	29 780	29 418	30 144	29 779	29 409
3	35 792	32 631	34 654	35 792	32 631	34 654	35 714	32 535	34 499	35 716	32 532	34 499
4	39 647	38 556	36 797	39 647	38 556	36 797	39 753	38 640	36 891	39 743	38 633	36 889
5	38 271	37 622	36 341	38 271	37 622	36 341	38 183	37 528	36 225	38 177	37 523	36 226

Bibliographie

Bailar, B. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

Bartlett, M.S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8, 27-41.

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.

Bollineni-Balabay, O., van den Brakel, J. et Palm, F. (2016a). Multivariate state space approach to variance reduction in series with level and variance breaks due to survey redesigns. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 179, 377-402.

Bollineni-Balabay, O., van den Brakel, J. et Palm, F. (2016b). State space time series modelling of the Dutch Labour Force Survey: Model selection and MSE estimation, - Extended version. Document de travail, Statistics Netherlands, Heerlen. https://www.cbs.nl/en-gb/background/2016/41/state-space-time-series.

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

Doornik, J. (2007). *An Object-Oriented Matrix Programming Language Ox 5*. Timberlake Consultants Press, Londres.

Durbin, J., et Koopman, S.J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89, 603-615.

Durbin, J., et Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. Numéro 38. Oxford University Press.

EUROSTAT (2015). Task force on monthly unemployment - revised report. Working group labour market statistics.

Hamilton, J. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics,* 33, 387-397.

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

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

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

Krieg, S., et van den Brakel, J. (2012). Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. *Computational Statistics & Data Analysis,* 56, 2918-2933.

Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête,* 13, 2, 211-220. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1987002/article/14607-fra.pdf.

ONS (2015). A state space model for LFS estimates: Agreeing the target and dealing with wave specific bias. Rapport de la 29e réunion du Comité consultatif de la méthodologie des services statistiques du gouvernement. http://www.ons.gov.uk/ons/guide-method/method-quality/advisory-committee/previous-meeting-papers-and-minutes/mac-29-papers.pdf.

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

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28, 40-68.

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 Tiller, R. (2005). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26, 893-916.

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 and Economic Statistics*, 16, 339-348.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rodriguez, A., et Ruiz, E. (2012). Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters. *Computational Statistics and Data Analysis*, 56, 62-74.

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

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

van den Brakel, J., 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., 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.

Zhang, M., et Honchar, O. (2016). Predicting survey estimates by state space models using multiple data sources. Article pour le Comité consultatif de la méthodologie de l'*Australian Bureau of Statistics*.