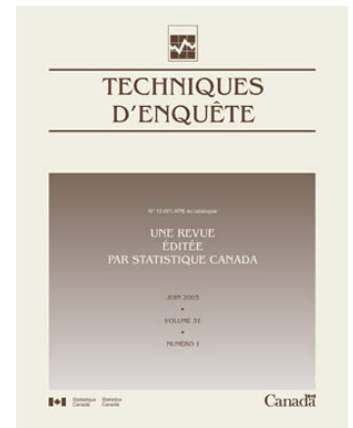


Techniques d'enquête

Variance de l'estimateur par la régression généralisée en situation d'erreur de mesure

par Jan van den Brakel et John Michiels

Date de diffusion : le 29 juin 2026



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 infostats@statcan.gc.ca

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

- | | |
|---|----------------|
| • Service de renseignements statistiques | 1-800-263-1136 |
| • Service national d'appareils de télécommunications pour les malentendants | 1-800-363-7629 |
| • Télécopieur | 1-514-283-9350 |

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.

Publication autorisée par la ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par la ministre de l'Industrie, 2026

L'utilisation de la présente publication est assujettie aux modalités de l'[entente de licence ouverte](#) de Statistique Canada.

Une [version HTML](#) est aussi disponible.

This publication is also available in English.

Variance de l'estimateur par la régression généralisée en situation d'erreur de mesure

Jan van den Brakel et John Michiels¹

Résumé

Sauf dans le cas d'un échantillonnage à deux phases, l'approximation classique de la variance de l'estimateur par la régression généralisée (estimateur GREG) part de l'hypothèse selon laquelle les totaux de population utilisés dans le système de pondération sont observés sans erreur. Si le modèle de pondération de l'estimateur GREG contient des totaux de population observés soumis à des sources d'erreur de mesure autres que l'erreur d'échantillonnage des estimations de première phase, cette incertitude sera alors ignorée par l'approximation de la variance de l'estimateur GREG. Le présent article propose, pour l'estimateur GREG, une approximation de la variance qui tient compte de l'incertitude supplémentaire résultant de l'erreur de mesure dans un ou plusieurs des totaux de population utilisés dans le système de pondération. Cette approche a été mise au point pour l'Enquête sur la population active aux Pays-Bas (EPAP) et y est appliquée. Les publications mensuelles de l'EPAP sont obtenues à l'aide d'un modèle de séries chronologiques qui corrige le biais de groupe de renouvellement ainsi que les discontinuités causées par d'importants remaniements et l'impossibilité d'effectuer des interviews sur place pendant la COVID-19. Les estimations par la régression généralisée des chiffres trimestriels sont étalonnées par rapport à la moyenne des publications mensuelles afin d'assurer une uniformité numérique entre les tableaux des publications mensuelles et trimestrielles. L'approximation classique de la variance de l'estimateur GREG part de l'hypothèse selon laquelle ces totaux de population sont observés sans erreur. Il en découle une sous-estimation de la variance de l'estimateur GREG. L'approximation de la variance proposée dans l'article aboutit à des erreurs-types plus réalistes pour les estimations trimestrielles par la régression généralisée.

Mots-clés : Calage; Enquête sur la population active; estimation de la variance; uniformité numérique entre les tableaux d'estimations.

1. Introduction

Les statistiques officielles publiées par les instituts nationaux de statistique reposent principalement sur des enquêtes par sondage et l'estimateur par la régression généralisée (GREG) (Särndal, Swensson et Wretman, 1992). L'estimateur GREG utilise de l'information auxiliaire dont les distributions dans la population sont connues pour améliorer la précision des estimations. Pour ce faire, les poids de sondage, définis comme l'inverse des probabilités d'inclusion du plan d'échantillonnage, sont calibrés de telle sorte que les sommes des variables auxiliaires pondérées des unités d'échantillonnage sont exactement égales aux distributions dans la population. Dans la plupart des cas, le principe de départ est que les totaux de population connus de ces variables auxiliaires sont exempts d'erreurs, puisqu'ils sont tirés de registres. Une exception est l'échantillonnage à deux phases, où les estimations des totaux de population fondées sur un grand échantillon de première phase sont utilisées dans le système de pondération des estimations fondées sur l'échantillon de deuxième phase. La variance de l'estimateur GREG dans le cas d'un échantillonnage à deux phases tient compte de l'incertitude supplémentaire liée à l'utilisation d'estimations tirées de la première phase dans le système de pondération des estimations obtenues à partir de la deuxième phase (Särndal et Swensson, 1987; Hidiroglou et Särndal, 1998; Singh et Kumar, 2010). Une méthode connexe,

1. Jan van den Brakel, Statistics Netherlands, C.P. 2454, 2490 AA Den Haag, Pays-Bas, et Maastricht University, C.P. 616, 6200 MD Maastricht, Pays-Bas. Courriel : ja.vandenbrakel@cbs.nl; John Michiels, Statistics Netherlands, C.P. 2454, 2490 AA Den Haag, Pays-Bas.

proposée par Renssen et Nieuwenbroek (1997) ainsi que par Berger, Muñoz et Rancourt (2009), est l'approximation de la variance qui tient compte de l'incertitude liée à l'utilisation, dans l'estimateur GREG, d'estimations pour des variables communes qui sont fondées sur au moins deux échantillons probabilistes.

Outre l'échantillonnage à deux phases, il existe d'autres situations où les totaux de population utilisés dans le système de pondération de l'estimateur GREG comportent de l'incertitude, qui n'est pas prise en compte dans l'approximation classique de la variance de l'estimateur GREG. L'estimateur GREG appliqué aux chiffres trimestriels de l'Enquête sur la population active aux Pays-Bas (EPAP) en est un exemple.

En fonction de l'EPAP, des chiffres mensuels, trimestriels et annuels sont publiés concernant la population active occupée et au chômage aux Pays-Bas. L'EPAP est fondée sur un plan avec renouvellement de panel. Les tableaux des publications mensuelles sont colligés à l'aide d'un modèle de séries chronologiques structurel (MSCS) multivarié afin de résoudre divers problèmes observés liés aux enquêtes par panel telles que l'EPAP. On utilise ce modèle comme une forme d'estimation pour petits domaines en augmentant la taille d'échantillon efficace du dernier mois au moyen des renseignements sur les échantillons observés au cours des mois précédents. Le modèle tient également compte du biais de groupe de renouvellement (BGR) (voir Bailer, 1975) et des effets systématiques découlant de trois remaniements importants du processus d'enquête ainsi que de l'impossibilité d'effectuer des interviews sur place pendant les confinements liés à la pandémie de COVID-19 (Van den Brakel et Krieg, 2015; Van den Brakel, Souren et Krieg, 2022).

Les chiffres trimestriels et annuels sont principalement calculés à l'aide de l'estimateur GREG le plus couramment utilisé. Les estimations par domaine fondées sur un modèle qui concernent le nombre de personnes actives occupées et au chômage au cours d'un mois sont incluses sous forme de termes de pondération dans l'estimateur GREG pour les diffusions trimestrielles et annuelles. Cette pratique garantit l'uniformité entre les chiffres mensuels, trimestriels et annuels sur la population active et permet de corriger, du moins en partie, les effets du BGR, des discontinuités et de la COVID dans les estimations par la régression généralisée des chiffres trimestriels et annuels sur la population active.

L'incertitude supplémentaire liée à l'utilisation d'estimations des totaux de population, tirées d'un modèle de séries chronologiques, n'est pas prise en compte dans les variances de l'estimateur GREG. Il peut s'ensuivre une forte sous-estimation de la variance des estimations trimestrielles par la régression généralisée, en particulier si celles-ci sont étroitement liées au tableau de pondération fondé sur les estimations des séries chronologiques mensuelles. Dans le présent article, nous proposons, pour l'estimateur GREG, une nouvelle approximation de la variance qui tient compte de l'incertitude supplémentaire liée à l'utilisation, dans le système de pondération, de totaux de population observés présentant des erreurs de mesure. Cette méthode est illustrée en étant appliquée aux chiffres trimestriels sur la population active issus de l'EPAP.

L'article est structuré comme suit. La section 2 présente brièvement l'estimateur GREG classique, principalement dans le but de définir la notation. La section 3 décrit l'EPAP, le MSCS pour les chiffres mensuels et l'estimateur GREG pour les chiffres trimestriels. Certains renseignements concernant le MSCS appliqué aux chiffres mensuels sont fournis afin d'expliquer la raison pour laquelle le tableau des publications mensuelles est inclus dans le système de pondération de l'estimateur GREG pour les chiffres trimestriels. À la section 4, une expression analytique de la variance de l'estimateur GREG qui tient compte de

l'incertitude relative aux totaux de population du système de pondération est proposée. La section 5 présente les résultats d'une application aux chiffres trimestriels de l'EPAP. L'article se termine par une discussion à la section 6.

2. Estimateur par la régression généralisée

L'estimateur GREG a pour but d'estimer les totaux de population, qui sont définis comme la somme des valeurs d'une variable d'intérêt, par exemple y , pour tous les éléments de la population cible U :

$$t_y = \sum_{i \in U} y_i. \quad (2.1)$$

À cette fin, un échantillon probabiliste s de taille n est prélevé d'une population cible dont tous les éléments présentent des probabilités de premier ordre non nulles π_i pour l'unité de population i et des probabilités d'inclusion de deuxième ordre $\pi_{ii'}$ pour les unités de population i et i' . L'estimateur GREG pour t_y est tiré d'un modèle de régression linéaire qui définit la relation entre la variable cible et un ensemble de variables auxiliaires;

$$y_i = \boldsymbol{\beta}' \mathbf{x}_i + \varepsilon_i, \quad (2.2)$$

où y_i est la réponse de la variable cible de l'unité d'échantillonnage i , $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})'$ est un vecteur à p dimensions contenant les variables auxiliaires de l'unité d'échantillonnage i , $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ est un vecteur à p dimensions contenant les coefficients de régression et ε_i est un résidu. On suppose que $E_{\varepsilon}(\varepsilon_i) = 0$ et $\text{Var}_{\varepsilon}(\varepsilon_i) = v_i \sigma^2$, où $E_{\varepsilon}(\cdot)$ et $\text{Var}_{\varepsilon}(\cdot)$ désignent l'espérance et la variance à l'égard du modèle (2.2) et v_i est un coefficient d'échelle qui est connu pour chaque unité d'échantillonnage. Dans Särndal et coll. (1992), au chapitre 6, il est indiqué que l'estimateur GREG pour t_y peut être exprimé comme suit :

$$\hat{t}_y^R = \hat{t}_y^{\pi} + \hat{\boldsymbol{\beta}}'(\mathbf{t}_x - \hat{\mathbf{t}}_x^{\pi}), \quad (2.3)$$

où $\hat{t}_y^{\pi}$ est l'estimateur de Narain-Horvitz-Thompson (Narain, 1951, et Horvitz et Thompson, 1952) pour t_y , c'est-à-dire

$$\hat{t}_y^{\pi} = \sum_{i \in s} \frac{y_i}{\pi_i},$$

$\mathbf{t}_x$ est un vecteur à p dimensions contenant les totaux de population des variables auxiliaires, c'est-à-dire

$$\mathbf{t}_x = \sum_{i \in U} \mathbf{x}_i,$$

et $\hat{\mathbf{t}}_x^{\pi}$ est l'estimateur de Narain-Horvitz-Thompson pour $\mathbf{t}_x$, que l'on définit comme suit :

$$\hat{\mathbf{t}}_x^{\pi} = \sum_{i \in s} \frac{\mathbf{x}_i}{\pi_i}.$$

Dans (2.3), un estimateur des coefficients de régression est défini comme suit :

$$\hat{\beta} = \left(\sum_{i \in S} \frac{\mathbf{x}_i \mathbf{x}_i'}{\pi_i v_i} \right)^{-1} \sum_{i \in S} \frac{\mathbf{x}_i y_i}{\pi_i v_i}.$$

La variance de l'estimateur GREG est obtenue en calculant approximativement (2.3) avec une linéarisation par série de Taylor du premier degré (Särndal et coll., 1992, chapitre 6) :

$$\hat{t}_y^R = \hat{t}_y^\pi + \beta' (\mathbf{t}_x - \hat{\mathbf{t}}_x^\pi) + \text{le reste} \doteq \hat{t}_y^\pi + \beta' (\mathbf{t}_x - \hat{\mathbf{t}}_x^\pi) \equiv \hat{t}_y^{R0}. \quad (2.4)$$

La variance de $\hat{t}_y^{R0}$ est utilisée comme approximation de la variance de $\hat{t}_y^R$ et est définie comme suit :

$$V(\hat{t}_y^{R0}) = \sum_{i \in U} \sum_{j \in U} (\pi_{ii'} - \pi_i \pi_j) \frac{e_i}{\pi_i} \frac{e_j}{\pi_j}, \quad (2.5)$$

où $e_i = y_i - \beta' \mathbf{x}_i$. Un estimateur de la variance est défini comme suit :

$$\hat{V}(\hat{t}_y^{R0}) = \sum_{i \in S} \sum_{j \in S} \frac{(\pi_{ii'} - \pi_i \pi_j)}{\pi_{ii'}} \frac{\hat{e}_i}{\pi_i} \frac{\hat{e}_j}{\pi_j}, \quad (2.6)$$

où $\hat{e}_i = y_i - \hat{\beta}' \mathbf{x}_i$.

3. Enquête sur la population active aux Pays-Bas

3.1 Plan de sondage

L'objectif de l'EPAP est de fournir des renseignements fiables sur la population active des Pays-Bas. Avant 2000, l'EPAP était conçue comme une enquête transversale. Depuis octobre 1999, l'EPAP est menée selon un plan avec renouvellement de panel. Chaque mois, un échantillon est intégré au panel. Les unités d'échantillonnage sont observées cinq fois à intervalles trimestriels, ce qui correspond à un plan avec renouvellement de panel mensuel de type 1-2-1 (5). Jusqu'en 2021, un échantillon stratifié par grappes à deux degrés d'adresses était prélevé chaque mois. Les strates correspondent à des régions géographiques. Les municipalités sont considérées comme des unités primaires d'échantillonnage et les adresses, comme des unités secondaires d'échantillonnage. Tous les ménages résidant à une même adresse, jusqu'à un maximum de trois ménages, sont inclus dans l'échantillon. Tous les membres du ménage âgés de 15 ans et plus sont inclus dans l'échantillon. Des sous-populations différentes font l'objet d'un suréchantillonnage afin d'améliorer la précision des diffusions officielles, par exemple, les adresses où vivent des personnes officiellement inscrites à l'agence de placement, ainsi que les sous-populations présentant de faibles taux de réponse. En 2021, dans le plan d'échantillonnage, on est passé d'un échantillon stratifié par grappes à deux degrés de ménages à un échantillon stratifié par grappes à deux degrés de personnes. Les strates correspondent aux mêmes régions géographiques que celles utilisées dans l'ancien plan. Les municipalités sont considérées comme des unités primaires d'échantillonnage et les personnes, comme des unités secondaires d'échantillonnage.

Dès le moment où l'EPAP est passée à un plan avec renouvellement de panel, environ 8 000 répondants ont été observés dans les échantillons mensuels de la première vague. Ce nombre a diminué progressivement pour s'établir à 6 500 personnes en 2012. Après le deuxième remaniement, en 2012, la taille des échantillons mensuels de la première vague a augmenté : environ 8 500 répondants ont été recensés dans les échantillons mensuels de la première vague. Depuis 2021, la taille moyenne des échantillons mensuels de la première vague est d'environ 5 000 personnes. Les taux de réponse de la première vague sont d'environ 55 %. Dans les vagues de suivi, environ 90 % des répondants de la vague précédente sont observés.

Depuis l'introduction du plan avec renouvellement de panel en 2000, le plan de sondage a été remanié trois fois. Le premier remaniement a eu lieu en 2010, lorsque la collecte de données de la première vague a changé : on est passé des interviews sur place (interviews sur place assistées par ordinateur [IPAO]) exclusivement à un plan à mode mixte combinant des interviews téléphoniques (interviews téléphoniques assistées par ordinateur [ITAO]) et des IPAO. Pour ce faire, il a fallu remanier en profondeur le questionnaire. En 2012, la collecte de données est passée à un plan à mode mixte séquentiel qui commence par des interviews sur le Web (interviews Web assistées par ordinateur [IWAO]), et le suivi des cas de non-réponse reposait sur des ITAO et des IPAO. Lors du remaniement de 2021, le plan d'échantillonnage est passé d'un échantillon fondé sur les ménages à un échantillon fondé sur les personnes. La collecte de données effectuée lors des vagues de suivi était fondée uniquement sur des ITAO, et ce, de l'introduction du plan avec renouvellement de panel jusqu'en 2021. Lors du remaniement de 2021, un plan à mode mixte séquentiel fondé sur les IWAO et les ITAO a été mis en œuvre pour les vagues de suivi.

3.2 Modèle de séries chronologiques pour les chiffres mensuels sur la population active

Jusqu'en 2010, Statistics Netherlands publiait chaque mois des chiffres trimestriels constamment mis à jour sur la population active en utilisant l'estimateur GREG. Cependant, la taille de l'échantillon mensuel était considérée comme étant trop petite pour produire des chiffres mensuels précis sur la population active à l'aide de cette méthode. Voilà pourquoi un MSCS multivarié a été mis en place en juin 2010 pour la production des chiffres mensuels sur la population active. Ce modèle est utilisé pour l'estimation pour petits domaines; il présente les estimations mensuelles les plus exactes possible en intégrant les renseignements des échantillons provenant des périodes de référence précédentes. Il permet également de corriger les effets du BGR ainsi que les effets systématiques observés dans les résultats découlant des remaniements de l'Enquête et de l'impossibilité d'effectuer des IPAO pendant la crise de la COVID-19. La présente sous-section donne une vue d'ensemble de ce modèle.

L'échantillon observé pour la j^{e} fois sera désormais appelé « j^{e} vague ». En raison du plan avec renouvellement, les données sont recueillies chaque mois à partir de cinq échantillons indépendants. Chaque mois, un nouvel échantillon est intégré au panel. Il s'agit de la première vague, c'est-à-dire que l'échantillon est observé pour la première fois. La deuxième vague correspond à l'échantillon qui a été intégré au panel il y a trois mois et qui est observé pour la deuxième fois. De même, les échantillons des troisième, quatrième

et cinquième vagues sont ceux qui ont intégré le panel il y a 6, 9 et 12 mois, et qui sont observés pour les troisième, quatrième et cinquième fois, respectivement. Les échantillons observés pour la cinquième fois quittent le panel.

L'objectif est d'estimer les totaux de population mensuels des chiffres sur la population active définis par (2.1). Dans ce cas-ci, y désigne les indicateurs qui définissent la situation d'activité du particulier i au cours du mois t . Supposons que $\hat{t}_{y_i}^{[j]}$ désigne l'estimateur GREG, défini par (2.3), pour un total de population inconnu (2.1) au cours du mois t , en fonction de l'échantillon qui est observé pour la j^e fois. Voir le système de pondération de cet estimateur GREG à l'annexe A.1. Le plan avec renouvellement de panel 1-2-1 (5) fait en sorte que, chaque mois, cinq estimations par la régression généralisée sont observées et peuvent être recueillies dans un vecteur à cinq dimensions, soit $\hat{y}_t = (\hat{t}_{y_i}^{[1]}, \dots, \hat{t}_{y_i}^{[5]})'$. Il est alors possible de construire une série chronologique à cinq dimensions, qui constitue les données d'entrée du MSCS suivant :

$$\hat{y}_t = \mathbf{1}_{[5]} \theta_t + \lambda_t + \sum_{r=1}^3 \Delta_r^r \gamma^r + \delta_t^{\text{COV}} \gamma^{\text{COV}} + \varepsilon_t. \quad (3.1)$$

Il s'agit d'un prolongement du modèle proposé par Pfeiffermann (1991). La première composante θ_t de (3.1) désigne le paramètre de population inconnu, et $\mathbf{1}_{[5]}$ est un vecteur colonne à cinq dimensions dont chaque élément est égal à un. Cette composante indique que $\hat{y}_t$ contient cinq estimations par la régression généralisée pour le paramètre de population au cours du mois t . Le paramètre de population est modélisé à l'aide d'un « MSCS de base », c'est-à-dire

$$\theta_t = L_t + S_t + I_t, \quad (3.2)$$

où L_t est le « modèle de tendance lisse » pour la variation à faible fréquence dans la série du paramètre de population, S_t est le modèle saisonnier trigonométrique pour les effets mensuels dans la série et I_t est une composante de bruit blanc pour la variation inexpliquée du paramètre de population; pour en savoir plus, voir Durbin et Koopman, 2012, chapitre 3. Cette première composante peut être interprétée comme une forme d'estimation pour petits domaines (Rao et Molina, 2015), puisqu'elle repose sur des renseignements tirés d'échantillons observés au cours de périodes de référence antérieures pour produire des estimations plus stables et plus exactes des chiffres mensuels sur la population active.

La deuxième composante de (3.1), c'est-à-dire λ_t , permet de modéliser le BGR attribuable au plan avec renouvellement de panel. Le BGR désigne le phénomène selon lequel il existe des différences systématiques entre les résultats des vagues du panel (Bailar, 1975), qui sont le résultat net des différences dans les questionnaires et les modes de collecte des données utilisés dans les diverses vagues, de l'attrition du panel et des effets du panel. Dans la présente application, on suppose que la première vague n'est pas touchée par un BGR et qu'elle fournit donc les estimations les plus fiables pour θ_t ; pour obtenir une justification, voir Van den Brakel et Krieg (2009). Les quatre autres composantes contiennent des marches aléatoires, désignées par $\lambda_t^{[j]}$ ($j = 2, \dots, 5$), et permettent de modéliser la différence systématique entre la première vague et les quatre vagues de suivi. Par conséquent, $\lambda_t = (0, \lambda_t^{[2]}, \lambda_t^{[3]}, \lambda_t^{[4]}, \lambda_t^{[5]})'$. Étant donné que le BGR de la

première vague est égal à zéro, les estimations du modèle de séries chronologiques pour θ_t sont étalonnées par rapport au niveau de la série de l'estimateur GREG de la première vague. De cette manière, les estimations sur la population active sont comparables aux estimations obtenues selon le plan transversal utilisé avant l'introduction du plan avec renouvellement de panel en octobre 1999.

La troisième composante permet de modéliser les discontinuités dans la série de données d'entrée qui découlent des trois remaniements importants de l'Enquête qui ont eu lieu, respectivement, en 2010, en 2012 et en 2021. Les discontinuités sont modélisées à l'aide d'interventions de niveau, c'est-à-dire $\Delta_t' = \text{diag}(\delta_t^{r,[1]}, \delta_t^{r,[2]}, \delta_t^{r,[3]}, \delta_t^{r,[4]}, \delta_t^{r,[5]})$, qui désigne une matrice diagonale contenant des variables fictives $\delta_t^{r,[j]}$ qui passent de zéro à un au moment où l'Enquête à la vague j passe de l'ancien plan au nouveau lors des remaniements $r=1$ en 2010, $r=2$ en 2012 et $r=3$ en 2021. De plus, $\gamma^r = (\gamma^{r,[1]}, \gamma^{r,[2]}, \gamma^{r,[3]}, \gamma^{r,[4]}, \gamma^{r,[5]})'$ sont des vecteurs à cinq dimensions qui contiennent des estimations des discontinuités dans les cinq vagues pendant les remaniements $r=1$ en 2010, $r=2$ en 2012 et $r=3$ en 2021.

La quatrième composante de (3.1), $\delta_t^{\text{COV}} \gamma^{\text{COV}}$, contient une correction pour la perte de réponses en raison de l'arrêt des IPAO lors de la première vague durant le confinement lié à la crise du coronavirus en 2020 et en 2021. Pour cette composante, $\delta_t^{\text{COV}} = (\delta_t^{\text{COV}}, 0, 0, 0, 0)'$ est un vecteur à cinq dimensions qui contient une intervention de niveau uniquement pour la première vague. L'indicateur δ_t^{COV} est égal à un pendant les mois de confinement sans IPAO et à zéro autrement. Le coefficient γ^{COV} est une approximation de la différence systématique lors de la première vague qui découle de la perte des IPAO lors de cette vague. Une estimation de γ^{COV} est obtenue en modélisant la série chronologique de la première vague fondée sur toutes les réponses avec une série fondée uniquement sur les réponses obtenues par ITAO et par Internet et une série de personnes dénombrées en tant que bénéficiaires dans une série chronologique structurelle multivariée. Pour obtenir plus de renseignements, voir Van den Brakel et coll. (2022).

La dernière composante de (3.1), c'est-à-dire ε_t , permet de modéliser les erreurs d'enquête et de tenir compte de l'hétéroscédasticité attribuable, par exemple, à la variation de la taille des échantillons au fil du temps et à la corrélation sérielle découlant du chevauchement partiel des échantillons dans le plan avec renouvellement de panel. Les erreurs d'échantillonnage sont empilées dans un vecteur à cinq dimensions $\varepsilon_t = (\varepsilon_t^{[1]}, \varepsilon_t^{[2]}, \varepsilon_t^{[3]}, \varepsilon_t^{[4]}, \varepsilon_t^{[5]})'$. Pour tenir compte de l'hétéroscédasticité, par exemple en raison de la variation de la taille des échantillons au fil du temps, les erreurs d'échantillonnage sont mises à l'échelle avec les erreurs-types des estimations par la régression généralisée de la série de données d'entrée, c'est-à-dire $\varepsilon_t^{[j]} = \sqrt{\text{var}(\hat{y}_t^{[j]})} \tilde{\varepsilon}_t^{[j]}$. Les erreurs-types des estimations par la régression généralisée sont estimées à partir des données d'enquête. L'erreur d'échantillonnage mise à l'échelle pour la première vague, c'est-à-dire $\tilde{\varepsilon}_t^{[1]}$, est un terme d'erreur dont la distribution est normale et indépendante et qui n'est pas corrélé avec les observations passées, puisque la première vague est observée pour la première fois. Les erreurs d'échantillonnage mises à l'échelle des vagues de suivi sont modélisées à l'aide d'un modèle autorégressif (3) afin de tenir compte de la corrélation sérielle avec les observations passées. Pour en savoir plus, voir Van den Brakel et Krieg (2015).

Le modèle (3.1) peut être exprimé dans ce que l'on appelle la représentation en espace d'états et ajusté à l'aide du filtre de Kalman afin d'obtenir des estimations optimales des variables d'état; voir, par exemple, Durbin et Koopman (2012). L'analyse est réalisée à l'aide d'un logiciel développé dans OxMetrics en combinaison avec les sous-routines de SsfPack 3.0; voir Doornik (2009) et Koopman, Shephard et Doornik (2008).

Les paramètres de population estimés par le modèle de séries chronologiques sont la population active au chômage, la population active occupée et la population active totale. Ces trois paramètres sont estimés à l'échelle nationale et ventilés en six domaines, en fonction d'un classement recoupé entre le genre et l'âge donnant trois catégories. Les variables d'intérêt sont la tendance (L_t) et le signal. Ce dernier est défini comme la tendance et la composante saisonnière ($L_t + S_t$). Ces estimations sont corrigées pour tenir compte des discontinuités, c'est-à-dire que les tendances publiées sont définies comme $L_t + \gamma^{1,[1]} + \gamma^{2,[1]} + \gamma^{3,[1]}$ et les signaux, comme $L_t + S_t + \gamma^{1,[1]} + \gamma^{2,[1]} + \gamma^{3,[1]}$. Rappelons que $\gamma^{r,[1]}$ est la discontinuité pour la première vague du r^{e} remaniement. Étant donné que les estimations de l'EPAP sont étalonnées au niveau de la première vague, si l'on part de l'hypothèse selon laquelle le BGR de la première vague est égal à zéro, les estimations de la tendance et du signal sont corrigées pour tenir compte des discontinuités de la première vague. De cette manière, la série complète pour la tendance et le signal se situent au niveau de mesure du plan de sondage le plus récent.

La population active occupée, la population active au chômage et la population active totale à l'échelle nationale ainsi que leur ventilation en six domaines définissent un ensemble de 21 paramètres. Le modèle (3.1) est appliqué de façon distincte à chacun de ces 21 paramètres. De ce fait, les tableaux des publications mensuelles ne sont pas uniformes sur le plan numérique. Par conséquent, une fonction de Lagrange est appliquée aux estimations filtrées afin de garantir que la somme des personnes actives occupées et de celles au chômage est égale à la population active totale à l'échelle nationale et dans les six domaines, et que la somme des six domaines est égale au total national. Pour en savoir plus, voir Van den Brakel et Krieg (2015).

3.3 Estimateur par la régression généralisée pour les estimations trimestrielles

Les estimations trimestrielles de l'EPAP sont fondées sur l'estimateur GREG. En raison de l'utilisation d'un plan mensuel 1-2-1 (5), les observations de chaque trimestre sont fondées sur 15 échantillons indépendants, puisqu'aucune des vagues observées au cours d'un même trimestre ne se chevauche. En conséquence, la taille d'échantillon est suffisamment grande pour produire des estimations suffisamment précises à l'aide d'une approche d'inférence fondée sur le plan de sondage telle que l'estimateur GREG. Le système de pondération de l'estimateur GREG trimestriel pour l'EPAP repose sur un ensemble complexe de tableaux contenant des variables auxiliaires sociodémographiques dont les totaux de population sont tirés de registres.

Contrairement au modèle de séries chronologiques utilisé pour produire les chiffres mensuels sur la population active, l'estimateur GREG ne tient pas compte du BGR, des effets des trois remaniements ou de la perte de réponses en raison de l'arrêt des IPAO pendant la période de la pandémie. Afin de garantir l'uniformité des estimations trimestrielles de la situation sur le marché du travail avec les estimations mensuelles, le système de pondération de l'estimateur GREG trimestriel contient également les moyennes sur trois mois issues des tableaux des publications mensuelles. De cette manière, les estimations trimestrielles correspondant aux chiffres mensuels seront uniformes sur le plan numérique, et les chiffres trimestriels qui constituent des améliorations ou sont liés aux chiffres mensuels seront corrigés pour tenir compte de ces effets systématiques. Le tableau supplémentaire du système de pondération qui contient les moyennes sur trois mois issues des tableaux des publications mensuelles est ci-après appelé « marge du MSCS ». Le système de pondération des chiffres trimestriels est précisé à l'annexe A.2.

L'ajout d'un tableau de population au système de pondération de l'estimateur GREG complique l'estimation de la variance. La variance de l'estimateur GREG part de l'hypothèse selon laquelle les totaux de population des moyennes sur trois mois issues des tableaux des publications mensuelles sont observés sans erreur. Le fait d'ignorer la variance des estimations du modèle de séries chronologiques conduit à une sous-estimation considérable de la variance des estimations trimestrielles par la régression généralisée. À l'heure actuelle, la variance des chiffres trimestriels sur la population active est estimée à l'aide d'un estimateur GREG sans les moyennes sur trois mois issues des tableaux des publications mensuelles dans le système de pondération. Il s'agit d'une approche pragmatique qui pourrait mener à la surestimation de la variance. Dans la section suivante, nous proposons une approximation de la variance qui tient compte de l'incertitude des totaux de population dans le système de pondération de l'estimateur GREG.

4. Variance de l'estimateur par la régression généralisée en situation d'erreur de mesure

L'approximation de la variance (2.5) et son estimateur (2.6) ignorent tous deux l'incertitude supplémentaire qui apparaît dans les estimations par la régression généralisée lorsque le système de pondération contient des composantes observées présentant des erreurs de mesure. Dans la présente application, cette incertitude provient du fait que le système de pondération contient des totaux de population estimés à l'aide d'un modèle de séries chronologiques pour des séries d'estimations issues d'une enquête mensuelle répétée. Dans la présente section, nous élaborons une méthodologie permettant de tenir compte de cette incertitude dans le contexte d'un étalonnage des estimations trimestrielles par la régression généralisée par rapport aux moyennes sur trois mois des prédictions du modèle de séries chronologiques. Il est entendu que cette méthodologie est plus générale et peut s'appliquer à des situations où les estimations sont étalonnées par rapport à des totaux de population tirés d'autres sources de données indépendantes.

Pour tenir compte de cette incertitude supplémentaire, le vecteur contenant les totaux de population des variables auxiliaires $\mathbf{t}_x$ est divisé en un terme pour lequel les totaux de population réels sont connus, disons

$\mathbf{t}_a$ (où l'indice a signifie « administratif »), et en un terme pour lequel les totaux de population réels sont estimés, disons $\tilde{\mathbf{t}}_m$ (où l'indice m représente l'estimation du modèle), c'est-à-dire $\mathbf{t}_x = (\tilde{\mathbf{t}}'_m \mathbf{t}'_a)'$. Cela donne :

$$\hat{t}_y^R = \hat{t}_y^\pi + \hat{\beta}'(\mathbf{t}_x - \hat{\mathbf{t}}_x^\pi) = \hat{t}_y^\pi + \hat{\beta}'_m(\tilde{\mathbf{t}}_m - \hat{\mathbf{t}}_m^\pi) + \hat{\beta}'_a(\mathbf{t}_a - \hat{\mathbf{t}}_a^\pi).$$

Dans ce cas-ci, $\hat{\mathbf{t}}_m^\pi$ et $\hat{\mathbf{t}}_a^\pi$ sont les estimateurs de Narain-Horvitz-Thompson pour $\mathbf{t}_m$ et $\mathbf{t}_a$, et $\hat{\beta}_m$ et $\hat{\beta}_a$ sont les estimations correspondantes pour les coefficients de régression. D'après une approximation du premier degré de Taylor, il s'ensuit que :

$$\hat{t}_y^R \doteq \hat{t}_y^\pi + \beta'_m(\tilde{\mathbf{t}}_m - \hat{\mathbf{t}}_m^\pi) + \beta'_a(\mathbf{t}_a - \hat{\mathbf{t}}_a^\pi) \equiv \hat{t}_y^{R0}.$$

Une expression de la variance de $\hat{t}_y^{R0}$ doit tenir compte de deux sources de variation : l'erreur d'échantillonnage liée au plan d'échantillonnage de l'EPAP et l'erreur de mesure du modèle de séries chronologiques. On y parvient en conditionnant sur l'erreur de mesure des modèles de séries chronologiques à l'aide de la décomposition suivante :

$$\text{Var}(\hat{t}_y^{R0}) = E_m \text{Var}_s(\hat{t}_y^{R0} | m) + \text{Var}_m E_s(\hat{t}_y^{R0} | m), \quad (4.1)$$

où E_m et Var_m désignent l'espérance et la variance par rapport au modèle de séries chronologiques, et E_s et Var_s désignent l'espérance et la variance par rapport au plan d'échantillonnage. Pour le premier terme de (4.1), il s'ensuit que la variance de l'estimateur par la régression, conditionnellement au modèle de séries chronologiques, est égale à la variance de l'estimateur par la régression traitant les totaux de population obtenus avec le modèle de séries chronologiques comme des valeurs fixes connues. En conséquence, il s'ensuit, pour le premier terme de (4.1), que :

$$E_m \text{Var}_s(\hat{t}_y^{R0} | m) = E_m \text{Var}_s(\hat{t}_y^{R0}) = \text{Var}_s(\hat{t}_y^{R0}).$$

Pour le deuxième terme de (4.1), il s'ensuit que :

$$\text{Var}_m E_s(\hat{t}_y^{R0} | m) = \text{Var}_m E_s(\hat{t}_y^\pi + \beta'_m(\tilde{\mathbf{t}}_m - \hat{\mathbf{t}}_m^\pi) + \beta'_a(\mathbf{t}_a - \hat{\mathbf{t}}_a^\pi) | m).$$

En prenant l'espérance par rapport au plan d'échantillonnage et en utilisant $E_s \hat{t}_q^\pi = t_q$, (pour $q = y, a, m$), nous supposons que

$$\text{Var}_m E_s(\hat{t}_y^{R0} | m) = \text{Var}_m(t_y + \beta'_m(\tilde{\mathbf{t}}_m - \mathbf{t}_m)).$$

Étant donné que t_y et $\mathbf{t}_m$ sont des totaux de population finie, qui sont des constantes par rapport au modèle de séries chronologiques, il s'ensuit que

$$\text{Var}_m E_s(\hat{t}_y^{R0} | m) = \beta'_m \text{Var}_m(\tilde{\mathbf{t}}_m) \beta_m, \quad (4.2)$$

où $\text{Var}_m(\tilde{\mathbf{t}}_m)$ est une matrice de covariance contenant les (co)variances des estimations du modèle de séries chronologiques sur la diagonale, qui sont obtenues à partir du logiciel utilisé pour produire les chiffres

mensuels sur la population active. Dans la présente application, $\text{Var}_m(\tilde{\mathbf{t}}_m)$ est une matrice diagonale, puisque les estimations des séries chronologiques $\tilde{\mathbf{t}}_m$ sont obtenues en appliquant le MSCS (3.1) à chaque catégorie de $\mathbf{t}_m$ de façon distincte. Il convient de prendre note que dans le cas où la variable cible t_y est incluse dans le tableau défini par $\mathbf{t}_m$, une régression de la variable cible y sera effectuée en fonction de la variable cible elle-même. Dans ce cas, le coefficient de régression correspondant est égal à un et tous les autres coefficients de régression sont égaux à zéro. Dans ce cas, $t_y - \boldsymbol{\beta}'_m \mathbf{t}_m = t_y - t_y = 0$. Dans le dernier paragraphe de la présente section, nous expliquons davantage ce cas particulier et, à l'annexe A.3, nous présentons une démonstration.

En conséquence, nous obtenons l'approximation suivante de la variance pour l'estimateur GREG :

$$\text{Var}(\hat{t}_y^{R0}) = \text{Var}_s(\hat{t}_y^{R0}) + \boldsymbol{\beta}'_m \text{Var}_m(\tilde{\mathbf{t}}_m) \boldsymbol{\beta}_m, \quad (4.3)$$

où $\text{Var}_s(\hat{t}_y^{R0})$, qui est définie dans (2.5), peut être estimée à l'aide de (2.6).

Pour justifier l'approximation de la variance (4.3), considérons le cas où l'estimateur GREG est utilisé pour estimer le tableau à partir du système de pondération correspondant aux composantes estimées par le modèle de séries chronologiques, c'est-à-dire $\mathbf{t}_m$. Disons que k désigne le nombre de catégories de $\mathbf{t}_m$ et que l désigne le nombre de catégories de $\mathbf{t}_a$, c'est-à-dire les composantes du système de pondération pour lesquelles les totaux de population sont connus. Si l'estimateur GREG est appliqué pour estimer le tableau $\mathbf{t}_m$, alors $\hat{t}_y^{R0}$ dans (2.4) devient un vecteur de k dimensions, disons $\hat{\mathbf{t}}_m^{R0}$. À l'annexe A.3, nous montrons que dans le cas en question, les coefficients de régression dans (2.4) forment une matrice $(k+l) \times k$ égale à

$$\mathbf{B} = \begin{bmatrix} \mathbf{I}_{[k \times k]} \\ \mathbf{O}_{[l \times k]} \end{bmatrix} \quad (4.4)$$

pour laquelle on suppose, sans perte de généralité, que les variables de pondération sont ordonnées de telle sorte que celles correspondant à la marge du MSCS viennent en premier. De plus, $\mathbf{I}_{[k \times k]}$ désigne la matrice d'identité à k dimensions et $\mathbf{O}_{[l \times k]}$ désigne une matrice $l \times k$ dont chaque élément est égal à zéro. Cela signifie que chaque estimateur par la régression pour les totaux dans $\hat{\mathbf{t}}_m^{R0}$ possède un coefficient de régression égal à un pour la variable auxiliaire du système de pondération correspondant à la variable cible, tandis que les $(k+l-1)$ autres coefficients de régression sont égaux à zéro. Cela laisse entendre que

$$\hat{\mathbf{t}}_m^{R0} = \hat{\mathbf{t}}_m^\pi + \mathbf{I}_{[k \times k]}(\tilde{\mathbf{t}}_m - \hat{\mathbf{t}}_m^\pi) + \mathbf{O}_{[l \times k]}(\mathbf{t}_a - \hat{\mathbf{t}}_a^\pi) = \tilde{\mathbf{t}}_m.$$

Cela laisse aussi entendre, pour le résidu dans (2.5) et (2.6), que $e_i = y_i - \boldsymbol{\beta}' \mathbf{x}_i = x_{i,k'} - \boldsymbol{\beta}'_{k'} \mathbf{x}_i = 0$ pour toutes les $i \in s$, où $\boldsymbol{\beta}_{k'}$ désigne la k' ème colonne de $\mathbf{B}$ dont le k' ème élément est égal à un et les $(k+l-1)$ éléments restants sont égaux à zéro. Il s'ensuit donc que la variance de l'estimateur GREG fondée sur (2.6) serait égale à zéro, ce qui est non souhaitable et inexact, puisque $\hat{\mathbf{t}}_m^{R0}$ est exactement égal à $\tilde{\mathbf{t}}_m$, qui est une estimation basée sur un modèle de séries chronologiques et est soumise à des incertitudes. En raison de (4.4), la variance de l'estimateur GREG fondée sur (4.3) est exactement égale à $\text{Var}_m(\tilde{\mathbf{t}}_m)$ et correspond donc correctement à l'incertitude des estimations du modèle de séries chronologiques $\tilde{\mathbf{t}}_m$.

Il existe dans la littérature quelques méthodologies connexes qui tiennent compte de l'incertitude supplémentaire présente lorsque l'estimateur GREG contient des marges soumises à des erreurs. Comme nous l'avons mentionné dans l'introduction, l'exemple classique d'échantillonnage à deux phases est celui où l'estimateur GREG pour la deuxième phase repose sur les totaux de population estimés à partir des observations de la première phase. Särndal et Swensson (1987), Hidiroglou et Särndal (1998) et Singh et Kumar (2010) proposent des approximations de la variance qui tiennent compte de cette incertitude supplémentaire.

Renssen et Nieuwenbroek (1997) ont étendu le concept de l'échantillonnage à deux phases afin d'harmoniser les estimations des variables communes dans au moins deux enquêtes par sondage. Il s'agit entre autres d'une approximation de la variance qui tient compte de l'incertitude liée à l'utilisation d'estimations pour les variables communes dans l'estimateur GREG pour des enquêtes par sondage distinctes. Berger et coll. (2009) approfondissent l'estimateur de la variance de Renssen et Nieuwenbroek (1997) en fournissant des données détaillées sur les échantillons à chevauchement partiel et sans chevauchement. Il s'agit entre autres de la variance pour les estimateurs composites dans les plans avec renouvellement de panel. La principale différence entre la méthodologie proposée dans le présent article et celle de Renssen et Nieuwenbroek (1997) et de Berger et coll. (2009) est qu'elles permettent d'en arriver à des approximations de la variance des estimateurs GREG qui reposent sur des variables communes issues de plusieurs enquêtes par sondage, en fonction de l'approche d'inférence classique fondée sur le plan de sondage pour l'échantillonnage probabiliste. Par conséquent, elles peuvent tenir compte de la dépendance découlant d'un chevauchement partiel des échantillons entre les sources. En revanche, la méthodologie proposée dans le présent article peut également prendre en compte l'erreur de mesure dans les totaux de population du système de pondération qui n'est pas liée à l'erreur d'échantillonnage.

5. Résultats

5.1 Mesures d'évaluation

Dans la présente sous-section, nous définissons des mesures d'évaluation afin de comparer les estimations ponctuelles et les erreurs-types des estimateurs GREG. Disons que $\hat{t}_{y,t}^{R(+m)}$ désigne l'estimation par la régression généralisée pour le trimestre t avec la marge du MSCS, c'est-à-dire les moyennes sur trois mois issues des tableaux des publications mensuelles dans le système de pondération, et que $\hat{t}_{y,t}^{R(-m)}$ désigne l'estimation par la régression généralisée pour le trimestre t sans ce tableau dans le système de pondération. Disons que la période $t = 1$ désigne le premier trimestre des estimations par la régression généralisée de la période incluse dans cette comparaison. De la même manière, T désigne le dernier trimestre des estimations par la régression généralisée de la période incluse dans cette comparaison. En conséquence, les estimations par la régression généralisée du trimestre T sont incluses dans la comparaison. Pour évaluer l'effet sur les estimations ponctuelles, nous calculons les mesures suivantes :

- Estimations de la moyenne :

$$EM = \frac{1}{T} \sum_{t=1}^T \hat{t}_{y,t}^{R(x)}, x = [+m, -m].$$

- Différence relative moyenne :

$$DR = \frac{1}{T} \sum_{t=1}^T \frac{\hat{t}_{y,t}^{R(+m)} - \hat{t}_{y,t}^{R(-m)}}{\hat{t}_{y,t}^{R(-m)}} \times 100 \%.$$

- Différence relative absolue moyenne :

$$DRA = \frac{1}{T} \sum_{t=1}^T \frac{|\hat{t}_{y,t}^{R(+m)} - \hat{t}_{y,t}^{R(-m)}|}{\hat{t}_{y,t}^{R(-m)}} \times 100 \%.$$

Pour évaluer l'effet sur les erreurs-types, disons que $ET(\hat{t}_{y,t}^{R(+m)})$ désigne l'erreur-type de l'estimateur GREG avec la marge du MSCS lorsque l'approximation de la variance ignore l'incertitude de la marge du MSCS dans le système de pondération. De la même manière, $ET(\hat{t}_{y,t}^{R(-m)})$ désigne l'erreur-type de l'estimateur GREG sans la marge du MSCS lorsque l'approximation classique de la variance ne tient pas compte de l'incertitude dans les totaux de population du système de pondération. Enfin, $ET(\hat{t}_{y,t}^{R(me)})$ désigne les erreurs-types de l'estimateur GREG avec la marge du MSCS lorsque l'approximation de la variance tient compte de l'incertitude de la marge du MSCS dans le système de pondération. Les mesures suivantes sont calculées :

- Erreurs-types moyennes :

$$ETM = \frac{1}{T} \sum_{t=1}^T ET(\hat{t}_{y,t}^{R(x)}), x = [+m, -m, me].$$

Il est entendu que *me* désigne les erreurs-types de l'estimateur GREG qui tient compte de l'incertitude du tableau des publications mensuelles dans le système de pondération et qu'il fait toujours référence à l'estimateur GREG pour lequel la marge du MSCS est comprise dans le système de pondération.

- Différence relative moyenne entre les erreurs-types :

$$DRET(x1,x2) = \frac{1}{T} \sum_{t=1}^T \frac{ET(\hat{t}_{y,t}^{R(x1)}) - se(\hat{t}_{y,t}^{R(x2)})}{ET(\hat{t}_{y,t}^{R(x2)})} \times 100 \%, x1, x2 = [+m, -m, me].$$

5.2 L'effet de la pondération des chiffres trimestriels en fonction des tableaux des publications mensuelles

Dans la présente sous-section, nous comparons les estimations ponctuelles fondées sur l'estimateur GREG et une pondération avec et sans la marge du MSCS. Ainsi, nous examinons dans quelle mesure le prolongement du modèle de pondération avec les publications mensuelles influe sur les estimations ponctuelles des chiffres trimestriels. L'examen portera sur 12 trimestres, soit ceux de 2017, de 2018 et de 2019.

Les estimations ponctuelles de la variable « situation sur le marché du travail » pour l'estimateur GREG avec et sans les tableaux des publications mensuelles dans le modèle de pondération sont comparées dans le tableau 5.1 et présentées dans la figure 5.1. Dans cet exemple, les prédictions des séries chronologiques

mensuelles correspondent au niveau de mesure du plan de sondage utilisé de 2012 à 2021. Par conséquent, la correction des estimations trimestrielles par la régression généralisée, à l'aide de la marge du MSCS, correspond au BGR moyen pour les quatre vagues de suivi de cette période. Pour la population active occupée, les estimations fondées sur la première vague sont systématiquement inférieures à celles fondées sur les vagues de suivi. Il en découle un BGR négatif pour la population active occupée. Par conséquent, les estimations trimestrielles par la régression généralisée avec la marge du MSCS donnent lieu à une correction à la baisse. Pour la population active au chômage, le BGR était faible à cette époque. Les différences entre les estimations par la régression généralisée avec et sans la marge du MSCS sont donc faibles. Les autres groupes comprennent diverses catégories de la population inactive. Pour ces groupes, le BGR est le complément du BGR de la population active occupée, puisque le BGR des personnes au chômage est presque nul. Par conséquent, les estimations trimestrielles par la régression généralisée avec la marge du MSCS donnent lieu à une correction du biais à la hausse.

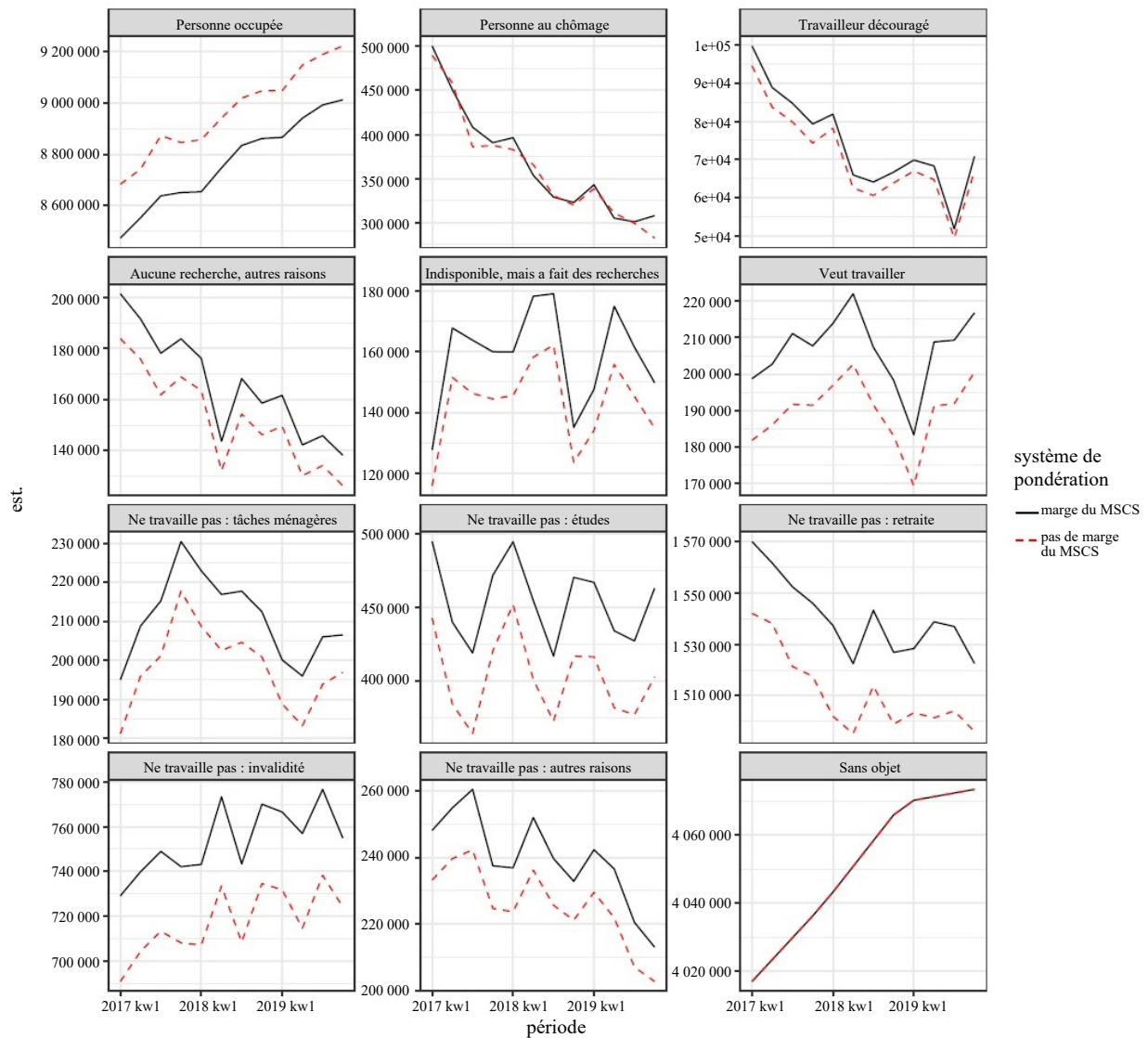
L'intégration de la marge du MSCS dans le système de pondération a un effet marqué sur certaines variables cibles. Les différences relatives (absolues) sont les plus importantes pour les estimations ponctuelles des catégories « Ne travaille pas : études » et « Indisponible, mais a fait des recherches ». Elles s'élèvent à 9,1 % et à 7,3 %, respectivement. Pour les deux catégories « Personne occupée » et « Personne au chômage », qui figurent également dans le tableau des publications mensuelles, l'effet sur les estimations ponctuelles est relativement faible. La différence absolue pour « Personne occupée » est d'environ 130 000, ce qui reste néanmoins substantiel. On voit la nécessité d'assurer une uniformité numérique entre les estimations mensuelles et trimestrielles, ainsi que la mesure dans laquelle les estimations trimestrielles bénéficient de la correction tenant compte du BGR et des discontinuités qui est appliquée aux chiffres mensuels à l'aide du modèle de séries chronologiques.

Tableau 5.1
Effet sur les mesures ponctuelles et les erreurs-types pour « situation sur le marché du travail »

Catégorie	EM(+m) chiffres	EM(-m) chiffres	DR %	DRA %
Personne occupée	8 768 846	8 897 074	-1,4	1,4
Personne au chômage	367 392	361 486	1,7	2,5
Travailleur découragé	73 290	70 964	3,2	3,2
Aucune recherche, autres raisons	164 637	155 059	6,2	6,2
Indisponible, mais a fait des recherches	152 885	142 442	7,3	7,3
Veut travailler	205 787	195 523	5,2	5,2
Ne travaille pas : tâches ménagères	217 006	211 615	2,6	2,6
Ne travaille pas : études	454 345	416 703	9,1	9,1
Ne travaille pas : retraite	1 538 607	1 514 953	1,6	1,6
Ne travaille pas : invalidité	765 337	750 500	2,0	2,0
Ne travaille pas : autres raisons	232 533	224 346	3,6	3,6

Note : Estimations de la moyenne (EM); différence relative absolue moyenne (DRA); différence relative moyenne (DR).

Figure 5.1 Estimations trimestrielles par la régression généralisée avec et sans la marge du MSCS dans le système de pondération pour la variable cible « situation sur le marché du travail », 2017 à 2019



5.3 Erreurs-types de l'estimateur par la régression généralisée en situation d'erreur de mesure

Dans la présente sous-section, l'approximation analytique de la variance proposée, qui tient compte de l'erreur de mesure dans le système de pondération, c'est-à-dire l'expression (4.3), est évaluée et comparée à deux versions différentes de l'approximation classique de la variance de l'estimateur GREG. La première version concerne les erreurs-types fondées sur le système de pondération qui comprend la marge du MSCS, elle-même fondée sur le MSCS mensuel. Cette approximation mènera à une sous-estimation des erreurs-types si la variable cible présente une corrélation élevée ou coïncide avec les estimations du tableau de la marge du MSCS. La deuxième version concerne les erreurs-types fondées sur le système de pondération quand la marge du MSCS est exclue. Cette comparaison a pour but d'examiner dans quelle mesure cette

dernière approche peut être utilisée comme approximation pragmatique pour les erreurs-types des chiffres trimestriels.

La contribution de l'erreur de mesure à l'approximation classique de la variance de l'estimateur GREG pour la situation sur le marché du travail est présentée à la figure 5.2. Les ETM et les DRET sont résumées dans le tableau 5.2.

On peut constater que l'effet de la prise en compte de l'incertitude de la marge du MSCS à l'aide de la correction proposée pour l'erreur de mesure est marqué pour certaines variables (comparer les lignes pointillées vertes avec les lignes pleines noires). Les catégories « Personne occupée » et « Personne au chômage » présentent les variations les plus importantes, mais pour les catégories « Ne travaille pas : études » et « Ne travaille pas : retraite », les erreurs-types sont sous-estimées si l'incertitude de la marge du MSCS est ignorée. Pour les catégories « Personne occupée » et « Personne au chômage », les coefficients de régression sont égaux à un pour la catégorie correspondant à la variable cible, tandis que les autres coefficients sont nuls. Sans le terme de correction de l'erreur de mesure, les estimations par la régression généralisée produisent des erreurs-types nulles, car la procédure de pondération tient compte des marges du MSCS. Avec le terme de correction de l'erreur de mesure, les estimations des erreurs-types reviennent au niveau des erreurs-types des moyennes sur trois mois issues des tableaux des publications mensuelles. Voir également la discussion portant sur l'équation (4.4) à la section 4.

Les erreurs-types de l'estimateur GREG qui tient compte de l'incertitude des marges du MSCS restent inférieures aux erreurs-types des estimations par la régression généralisée selon le système de pondération sans les marges du MSCS (comparer les lignes pointillées vertes avec les lignes discontinues rouges). Les erreurs-types sont environ 25 % plus faibles pour les personnes occupées pour la période de 2017 à 2019 et environ 40 % plus faibles pour les personnes au chômage, comme le montre le tableau 5.2. Pour les variables cibles qui coïncident avec la marge du MSCS, la surestimation de l'erreur-type obtenue avec un estimateur GREG qui exclut la marge du MSCS du système de pondération peut être considérable.

Les autres catégories présentant une réduction importante de l'erreur-type sont « Ne travaille pas : études » et « Ne travaille pas : retraite ». Pour certaines autres catégories, le terme de pondération supplémentaire entraîne une augmentation des erreurs-types estimées; la catégorie la plus notable à cet égard est « Indisponible, mais a fait des recherches », dont l'augmentation relative de l'erreur-type est de 4,5 %, comme le montre le tableau 5.2. En ce qui concerne les autres catégories de la situation sur le marché du travail telles que « Travailleur découragé » ou « Ne travaille pas : autres raisons », nous pouvons constater que la prise en compte de l'incertitude de la marge du MSCS entraîne des variations relativement plus faibles des erreurs-types. Cela s'explique probablement par le fait que la relation entre ces variables cibles et les totaux de population dans la marge du MSCS est moins forte.

Tableau 5.2
Effet sur les mesures ponctuelles et les erreurs-types pour la situation sur le marché du travail

Catégorie	ETM(+m) chiffres	ETM(-m) chiffres	ETM(me) chiffres	DRET (+m,-m) %	DRET (+m,me) %	DRET (-m,me) %
Personne occupée	0	16 568	13 280	-100,0	-100,0	24,8
Personne au chômage	0	8 480	6 064	-100,0	-100,0	39,7
Travailleur découragé	4 706	4 586	4 720	2,6	-0,3	-2,8

Note : Erreur-type moyenne (ETM); différence relative moyenne entre les erreurs-types (DRET).

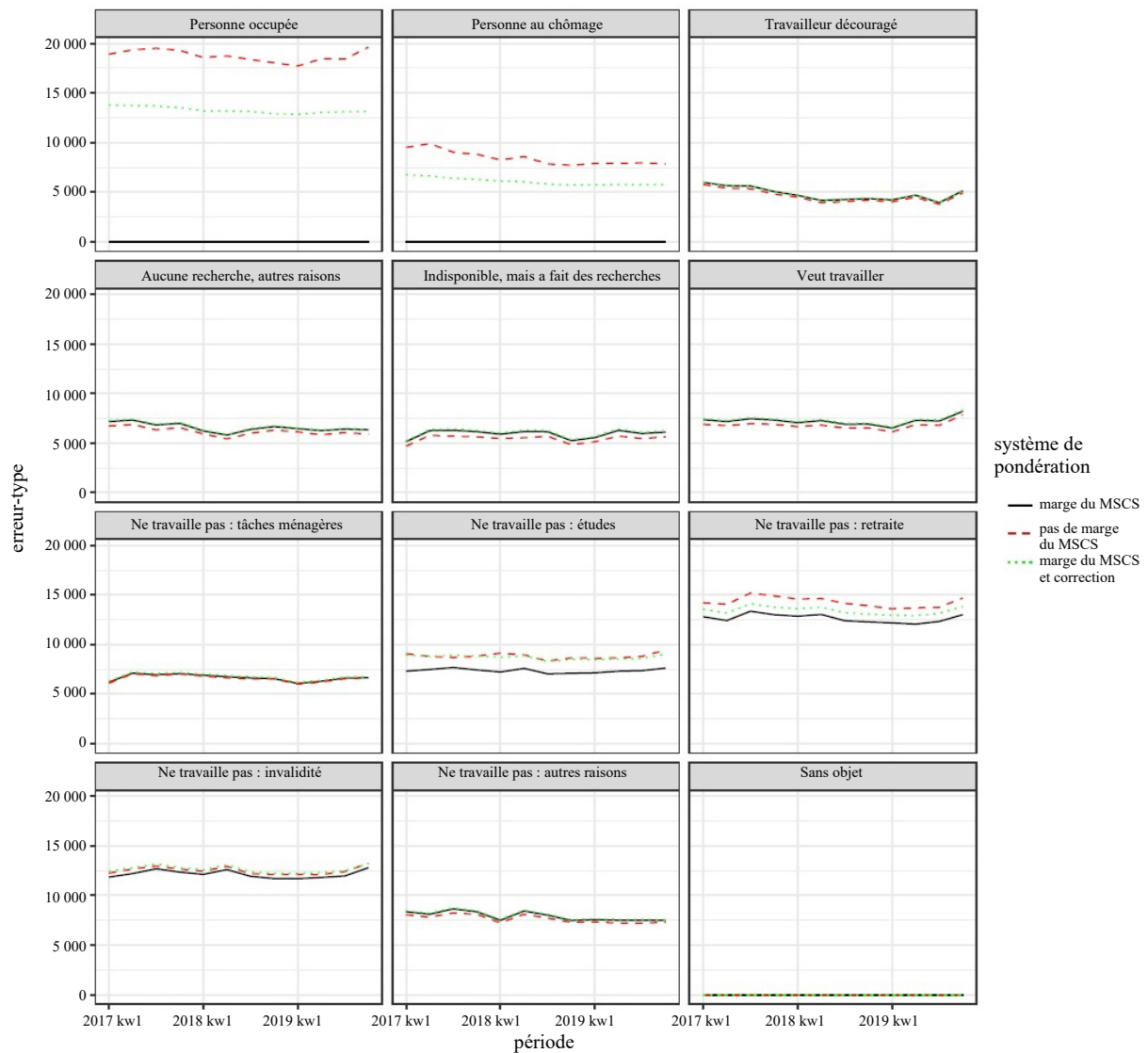
Tableau 5.2 (suite)

Effet sur les mesures ponctuelles et les erreurs-types pour la situation sur le marché du travail

Catégorie	ETM(+m) chiffres	ETM(-m) chiffres	ETM(me) chiffres	DRET (+m,-m) %	DRET (+m,me) %	DRET (-m,me) %
Aucune recherche, autres raisons	6 623	6 407	6 710	3,4	-1,3	-4,5
Indisponible, mais a fait des recherches	5 850	5 597	6 003	4,5	-2,5	-6,7
Veut travailler	7 338	7 155	7 447	2,6	-1,5	-3,9
Ne travaille pas : tâches ménagères	6 875	6 969	6 967	-1,3	-1,3	0,0
Ne travaille pas : études	7 522	9 255	8 820	-18,7	-14,7	4,9
Ne travaille pas : retraite	12 343	13 980	13 321	-11,7	-7,3	4,9
Ne travaille pas : invalidité	12 173	12 477	12 496	-2,4	-2,6	-0,2
Ne travaille pas : autres raisons	7 829	7 680	7 914	1,9	-1,1	-2,9

Note : Erreur-type moyenne (ETM); différence relative moyenne entre les erreurs-types (DRET).

Figure 5.2 Erreurs-types de l'estimateur par la régression généralisée sans (noir et rouge) et avec (vert) la correction des erreurs de mesure pour la variable cible « situation sur le marché du travail »



Note : Modèle de séries chronologiques structurel (MSCS).

6. Conclusion

Dans le présent article, nous proposons une approximation de la variance pour les estimateurs GREG dont le système de pondération contient des totaux de population observés présentant des erreurs de mesure. Les méthodes existantes dans la littérature qui tiennent compte de l'incertitude dans le système de pondération de l'estimateur GREG sont fondées sur un échantillonnage à deux phases où les estimations des totaux de population provenant de la première phase servent d'information auxiliaire à l'estimateur GREG pendant la deuxième phase. Dans ce cas, le plan d'échantillonnage peut être utilisé pour obtenir une approximation de la variance de l'estimateur GREG qui tient compte de cette incertitude supplémentaire. Pour en savoir plus, voir par exemple Särndal et coll. (1992), chapitre 9. Une méthodologie connexe, proposée par Renssen et Nieuwenbroek (1997) et Berger et coll. (2009), est l'approximation de la variance qui tient compte de l'incertitude liée à l'utilisation d'estimations pour les variables communes dans l'estimateur GREG qui sont fondées sur au moins deux échantillons probabilistes.

La situation examinée dans le présent article est différente. Nous examinons les estimateurs GREG dont le système de pondération contient des totaux de population dans le but d'assurer une uniformité numérique entre les estimations par la régression généralisée et ces totaux de population, même si les totaux de population observés présentent des erreurs. Cette incertitude peut être attribuable à des erreurs d'échantillonnage fondées sur le plan, comme dans Renssen et Nieuwenbroek (1997) et Berger et coll. (2009). Cependant, elle peut également tenir compte d'autres sources d'incertitude. Mentionnons, par exemple, l'incertitude associée aux estimations basées sur un modèle provenant de modèles de séries chronologiques ou de modèles d'estimation pour petits domaines. Une situation pratique dans laquelle cela se produit est lorsque les estimations par la régression généralisée sont étalonnées par rapport à des totaux de population corrigés en fonction du BGR ou des discontinuités. Si ces totaux de population sont fondés sur des estimations et sont donc soumis à une erreur de mesure, il est alors important que la variance des estimations par la régression généralisée tienne également compte de cette incertitude supplémentaire. Contrairement à l'échantillonnage à deux phases, la variance de l'estimateur GREG ne peut pas tenir compte de cette incertitude supplémentaire, car cette source d'erreur de mesure n'est pas liée au plan d'échantillonnage. Par conséquent, nous proposons une approximation de la variance de l'estimateur GREG qui contient une composante additionnelle de variance. Nous partons naturellement de l'hypothèse selon laquelle le niveau d'incertitude des totaux de population est connu *a priori*.

La méthodologie proposée dans le présent article pour tenir compte de l'incertitude supplémentaire pendant l'étalonnage des estimations trimestrielles par la régression généralisée par rapport aux moyennes sur trois mois des prédictions issues du modèle de séries chronologiques pour la population active néerlandaise peut être appliquée de manière plus générale aux cas où les estimations sont étalonnées par rapport aux totaux de population tirés d'autres sources de données indépendantes. Wulansari (2026) a appliqué la méthodologie proposée dans le présent article au contexte de l'estimation pour petits domaines afin de calculer

approximativement la variance du modèle de Fay-Herriot dans le cas d'un échantillonnage complexe et en présence d'erreurs de mesure dans les covariables.

La méthodologie proposée est appliquée aux estimations trimestrielles de l'EPAP. L'équipe de l'EPAP utilise un MSCS multivarié pour produire les chiffres mensuels officiels sur la population active. Les moyennes trimestrielles sur trois mois issues des tableaux des publications mensuelles sont incluses dans le système de pondération de l'estimateur GREG pour les chiffres trimestriels, le but étant d'assurer une uniformité numérique entre les tableaux des publications mensuelles et trimestrielles. Avec l'approximation de la variance classique de l'estimateur GREG, les tableaux des publications mensuelles sont traités comme s'ils étaient observés sans erreur. En conséquence, la variance des variables cibles trimestrielles qui sont étroitement liées au tableau des publications mensuelles sera considérablement sous-estimée. Dans le cas le plus extrême, la variance aura tendance à être nulle pour les variables cibles qui sont également comprises dans le tableau des publications mensuelles. Dans la méthodologie proposée, l'approximation de la variance de l'estimateur GREG est augmentée par la somme des variances des estimations du MSCS du tableau des publications mensuelles, multipliée par les valeurs au carré des coefficients de régression du tableau des données de sortie mensuelles de l'estimateur GREG.

Une approche plus pragmatique consiste à utiliser l'approximation classique de la variance de l'estimateur GREG avec un système de pondération sans le tableau des publications mensuelles. Par rapport à l'approximation de la variance proposée, cette solution de rechange ne mène pas à une surestimation des erreurs-types (ou ne mène qu'à une légère surestimation) pour les variables cibles trimestrielles qui ne sont pas liées ou ne sont liées que faiblement aux variables du tableau des données de sortie mensuelles. Pour les variables cibles trimestrielles qui présentent une corrélation élevée avec les variables du tableau des données de sortie mensuelles, cette approche peut faire en sorte de surestimer les erreurs-types de jusqu'à 40 %.

Statistics Netherlands utilisera l'approximation de la variance élaborée dans le présent article pour publier les erreurs-types des chiffres trimestriels de l'EPAP. L'un des inconvénients de l'approximation de la variance proposée est qu'elle nécessite l'évaluation d'une composante additionnelle de variance. Pour ce faire, il faut appliquer une combinaison linéaire des coefficients de régression du tableau des publications mensuelles issu du système de pondération et des erreurs-types des estimations mensuelles du modèle de séries chronologiques. Pour les variables cibles qui ne sont pas liées ou qui ne sont liées que faiblement au tableau des publications mensuelles utilisé dans le système de pondération, les erreurs-types des chiffres trimestriels seront estimées à partir de la variance de l'estimateur GREG sans tenir compte du tableau des publications mensuelles.

Pour les chiffres trimestriels qui sont étroitement liés au tableau des publications mensuelles utilisé dans le système de pondération, il est nécessaire de prendre en compte l'incertitude supplémentaire qui y est associée. Cette opération est relativement simple pour les variables cibles qui coïncident avec l'une des

catégories du tableau des publications mensuelles, puisque les coefficients de régression sont égaux à un dans ce cas. Par conséquent, l'erreur-type du chiffre trimestriel est égale à l'erreur-type des moyennes sur trois mois issues des estimations du modèle de séries chronologiques. Étant donné que les erreurs-types des estimations des séries chronologiques mensuelles sont très stables au fil du temps, il n'est pas nécessaire de les recalculer pour chaque trimestre.

Pour les chiffres trimestriels qui constituent des classements croisés détaillés des catégories du tableau des publications mensuelles issu du système de pondération, il est également nécessaire de tenir compte de l'incertitude supplémentaire dans les estimations des séries chronologiques mensuelles. Dans ce cas, les coefficients de régression ne seront pas exactement égaux à un, de sorte qu'une estimation à partir de l'estimateur GREG sera nécessaire. Un projet de suivi portera sur la stabilité des variances des estimations par la régression généralisée des chiffres trimestriels de l'EPAP au fil du temps. En fonction de leur stabilité au fil du temps, ces erreurs-types devront être mises à jour périodiquement, mais pas nécessairement à une fréquence trimestrielle.

Remerciements

Les auteurs souhaitent remercier le rédacteur associé anonyme et les deux examinateurs d'avoir lu attentivement leur manuscrit et d'avoir fourni des commentaires constructifs qui se sont avérés utiles pour améliorer l'article. Les points de vue exprimés dans l'article sont ceux des auteurs et ne reflètent pas nécessairement les politiques de Statistics Netherlands. Le projet de recherche est financé par Eurostat en vertu de l'accord de subvention 2023-NL-LFS, n° 101139082.

Annexe

A.1 Modèle de pondération pour les chiffres mensuels sur la population active

La présente annexe précise le modèle de pondération de l'estimateur GREG utilisé pour produire les séries de données d'entrée du modèle de séries chronologiques à la sous-section 3.2. Les nombres entre parenthèses correspondent au nombre de catégories précisées par la variable auxiliaire. Les estimations mensuelles par la régression généralisée par vague qui composent les séries de données d'entrée pour le modèle de séries chronologiques (3.1) sont fondées sur le système de pondération suivant :

Nationalité(3) + Genre(2) + Âge(21) + Type de ménage(3) + Région(44) + Âge*Genre(7) +
Prestations d'assurance-emploi(5) + Revenu(3).

La plupart des noms de variable sont évidents, à l'exception de la variable *Prestations d'assurance-emploi(5)*, qui désigne la période pendant laquelle un répondant reçoit des prestations d'assurance-emploi, répartie en cinq catégories distinctes.

A.2 Modèle de pondération pour les chiffres trimestriels sur la population active

La présente annexe précise le modèle de pondération de l'estimateur GREG utilisé pour produire les chiffres trimestriels à la sous-section 3.3. Ce modèle est défini comme suit :

Genre(2)*Nationalité(20) + Grandes municipalités(197)*ÂgeGenre(3) + Genre(2)*Âge(43) + Âge(5)*Mode(3)*Genre(2) + Prestations d'assurance-emploi(5) + Prestations d'assurance-emploi(2)*Province(12) + Revenu(6) + Type de revenu(3)*Province(12) + Type de ménage(3) + Situation sur le marché du travail(3)*Genre(2)*Âge(3).

Les nombres entre parenthèses correspondent au nombre de catégories précisées par la variable auxiliaire. La plupart des noms de variable sont évidents, à l'exception des suivants :

- ÂgeGenre(3) : répartition en trois catégories (moins de 15 ans ou plus de 64 ans, hommes de 15 à 64 ans, femmes de 15 à 64 ans).
- Situation sur le marché du travail(3)*Genre(2)*Âge(3) : la marge du MSCS ou la moyenne sur trois mois provenant des tableaux des publications mensuelles pour la situation sur le marché du travail classée en trois catégories selon le genre et l'âge.

A.3 Démonstration de la formule (4.4)

La présente annexe montre que la matrice des coefficients de régression de l'estimateur GREG est égale à (4.4) si l'estimateur GREG est utilisé pour estimer le tableau à partir du système de pondération correspondant aux composantes estimées avec le modèle de séries chronologiques. Tout d'abord, l'estimateur GREG est défini pour un vecteur de totaux de population :

$$\hat{\mathbf{t}}_y^{R0} = \hat{\mathbf{t}}_y^{\pi} + \mathbf{B}'(\mathbf{t}_x - \hat{\mathbf{t}}_x^{\pi}),$$

où $\hat{\mathbf{t}}_y^{R0}$ et $\hat{\mathbf{t}}_y^{\pi}$ sont deux vecteurs colonnes à q dimensions contenant les estimateurs par la régression généralisée et de Horvitz-Thompson pour les q totaux de population d'intérêt et $\mathbf{B}$ est une matrice à $p \times q$ dimensions contenant les coefficients de régression du modèle de régression multivariée qui motive l'estimateur GREG. Le modèle de régression multivariée est défini comme suit :

$$\mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{E},$$

où $\mathbf{Y}$ est une matrice $n \times q$ dont chaque rangée correspond aux q valeurs cibles ($y_{i,1}, \dots, y_{i,q}$) de l'unité d'échantillonnage $i = 1, \dots, n$, $\mathbf{X}$ est une matrice $n \times p$ dont chaque rangée correspond aux p variables auxiliaires issues du système de pondération ($x_{i,1}, \dots, x_{i,p}$) de l'unité d'échantillonnage $i = 1, \dots, n$ et $\mathbf{E}$ est une matrice $n \times q$ contenant les résidus correspondants du modèle de régression. L'estimateur GREG pour $\mathbf{B}$ est désormais défini en notation matricielle comme suit :

$$\hat{\mathbf{B}} = [\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y},$$

où $\boldsymbol{\Sigma}$ est une matrice diagonale $n \times n$ comprenant les éléments diagonaux $\pi_i \nu_i$, c'est-à-dire le produit des probabilités d'inclusion π_i et le coefficient d'échelle de la variance du modèle de régression linéaire ν_i de l'unité d'échantillonnage $i = 1, \dots, n$. Rappelons, d'après la section 4, que le vecteur à p dimensions contenant les totaux de population pour les variables auxiliaires $\mathbf{t}_x$ peut être divisé en un vecteur pour lequel les totaux de population réels sont connus $\mathbf{t}_a$ d'une longueur l et en un vecteur pour lequel les totaux de population des tableaux des publications mensuelles sont estimés à l'aide du modèle de séries chronologiques $\tilde{\mathbf{t}}_m$ d'une longueur k . Ainsi, $\mathbf{t}_x = (\tilde{\mathbf{t}}_m' \mathbf{t}_a')'$ et $\mathbf{X} = (\mathbf{X}_m \mathbf{X}_a)$, où $\mathbf{X}_m$ est une matrice $n \times k$ dont chaque rangée correspond aux k variables auxiliaires de $\tilde{\mathbf{t}}_m$ et où $\mathbf{X}_a$ est une matrice $n \times l$ dont chaque rangée correspond aux l variables auxiliaires de $\mathbf{t}_a$.

Considérons le cas où $\hat{\mathbf{t}}_y^{R0}$ correspond à la composante du modèle de pondération que l'on estime à l'aide du modèle de séries chronologiques $\mathbf{t}_m$. Dans ce cas, $\hat{\mathbf{t}}_y^{R0} = \hat{\mathbf{t}}_m^{R0}$. Dans ce cas, nous pouvons montrer que

$$\hat{\mathbf{B}} = \begin{bmatrix} \mathbf{I}_{[k \times k]} \\ \mathbf{O}_{[l \times k]} \end{bmatrix}.$$

Démonstration :

Dans ce cas, $\mathbf{Y} = \mathbf{X}_m$ et $\mathbf{X} = (\mathbf{X}_m \mathbf{X}_a)$, ce qui mène à l'expression suivante pour $\hat{\mathbf{B}}$:

$$\hat{\mathbf{B}} = [(\mathbf{X}_m \mathbf{X}_a)' \boldsymbol{\Sigma}^{-1} (\mathbf{X}_m \mathbf{X}_a)]^{-1} (\mathbf{X}_m \mathbf{X}_a)' \boldsymbol{\Sigma}^{-1} \mathbf{X}_m.$$

En précisant les opérations matricielles, nous obtenons :

$$\hat{\mathbf{B}} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_{11} \\ \mathbf{X}_{21} \end{bmatrix},$$

où

$$\mathbf{X}_{11} = \mathbf{X}_m' \boldsymbol{\Sigma}^{-1} \mathbf{X}_m,$$

$$\mathbf{X}_{21} = \mathbf{X}_a' \boldsymbol{\Sigma}^{-1} \mathbf{X}_m,$$

$$\mathbf{X}_{12} = \mathbf{X}_m' \boldsymbol{\Sigma}^{-1} \mathbf{X}_a,$$

$$\mathbf{X}_{22} = \mathbf{X}_a' \boldsymbol{\Sigma}^{-1} \mathbf{X}_a.$$

En prenant l'inverse de la matrice partitionnée (voir par exemple Mardia, Kent et Bibby, 1979, section A.2.4), nous obtenons :

$$\hat{\mathbf{B}} = \begin{bmatrix} \mathbf{A} & -\mathbf{A}\mathbf{X}_{12}\mathbf{X}_{22}^{-1} \\ -\mathbf{X}_{22}^{-1}\mathbf{X}_{21}\mathbf{A} & (\mathbf{X}_{22} - \mathbf{X}_{21}\mathbf{X}_{11}^{-1}\mathbf{X}_{12})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{11} \\ \mathbf{X}_{21} \end{bmatrix}, \quad (\text{A.1})$$

où $\mathbf{A} = (\mathbf{X}_{11} - \mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21})^{-1}$.

En utilisant la formule (A.2.4f) de Mardia, Kent et Bibby (1979), nous avons :

$$(\mathbf{X}_{22} - \mathbf{X}_{21}\mathbf{X}_{11}^{-1}\mathbf{X}_{12})^{-1} = \mathbf{X}_{22}^{-1} + \mathbf{X}_{22}^{-1}\mathbf{X}_{21}\mathbf{A}\mathbf{X}_{12}\mathbf{X}_{22}^{-1}. \quad (\text{A.2})$$

En insérant (A.2) dans (A.1) et en précisant davantage les opérations matricielles, nous obtenons :

$$\hat{\mathbf{B}} = \begin{bmatrix} \mathbf{A}\mathbf{X}_{11} - \mathbf{A}\mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21} \\ -\mathbf{X}_{22}^{-1}\mathbf{X}_{21}\mathbf{A}\mathbf{X}_{11} + \mathbf{X}_{22}^{-1}\mathbf{X}_{21} + \mathbf{X}_{22}^{-1}\mathbf{X}_{21}\mathbf{A}\mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21} \end{bmatrix}.$$

De là, il s'ensuit, pour la première composante, que

$$\mathbf{A}\mathbf{X}_{11} - \mathbf{A}\mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

et, pour la deuxième composante, que

$$\begin{aligned} -\mathbf{X}_{22}^{-1}\mathbf{X}_{21}\mathbf{A}\mathbf{X}_{11} + \mathbf{X}_{22}^{-1}\mathbf{X}_{21} + \mathbf{X}_{22}^{-1}\mathbf{X}_{21}\mathbf{A}\mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21} &= -\mathbf{X}_{22}^{-1}\mathbf{X}_{21}\mathbf{A}(\mathbf{X}_{11} - \mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21}) + \mathbf{X}_{22}^{-1}\mathbf{X}_{21} \\ &= -\mathbf{X}_{22}^{-1}\mathbf{X}_{21}\mathbf{A}\mathbf{A}^{-1} + \mathbf{X}_{22}^{-1}\mathbf{X}_{21} = \mathbf{O}. \end{aligned}$$

En recueillant les deux résultats, nous obtenons

$$\hat{\mathbf{B}} = \begin{bmatrix} \mathbf{I} \\ \mathbf{O} \end{bmatrix},$$

ce qui prouve (4.4).

Bibliographie

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

Berger, Y.G., Muñoz, J.F. et Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals – An application to the extended regression estimator and the regression composite estimator. *Computational Statistics and Data Analysis*, 53, 2596-2604.

Doornik, J. (2009). *An Object-Oriented Matrix Programming Language Ox 6*. Timberlake Consultants Press.

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

- Hidioglou, M.A. et Särndal, C.-E. (1998). [Emploi des données auxiliaires dans l'échantillonnage à deux phases](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1998001/article/3905-fra.pdf). *Techniques d'enquête*, 24(1), 11-20. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1998001/article/3905-fra.pdf>.
- Horvitz, D.G. et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Koopman, S.J., Shephard, N. et Doornik, J. (2008). *Ssfpack 3.0: Statistical Algorithms for Models in State-Space Form*. Timberlake Consultants, Press London.
- Mardia, K.V., Kent, J.T. et Bibby, J.M. (1979). *Multivariate Analysis*. London: Academic Press.
- Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.
- Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, 163-175.
- Rao, J.N.K. et Molina, I. (2015). *Small Area Estimation, 2nd edition*. New York: John Wiley & Sons, Inc.
- Renssen, R.H. et Nieuwenbroek, N. (1997). Alligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.
- Särndal, C.-E. et Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 279-294.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- Singh, H.P. et Kumar, S. (2010). Estimation of mean in presence of non-response using two phase sampling scheme. *Statistical Papers*, 51, 559-582.
- van den Brakel, J.A. et Krieg, S. (2015). [Remédier aux petites tailles d'échantillon, au biais de groupe de renouvellement et aux discontinuités dans les plans de sondage avec renouvellement de panel](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015002/article/14231-fra.pdf). *Techniques d'enquête*, 41(2), 281-312. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015002/article/14231-fra.pdf>.

- van den Brakel, J.A. et Krieg, S. (2009). [Estimation du taux de chômage mensuel par modélisation structurelle de séries chronologiques dans un plan de sondage avec renouvellement de panel](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009002/article/11040-fra.pdf). *Techniques d'enquête*, 35(2), 193-207. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009002/article/11040-fra.pdf>.
- van den Brakel, J.A., Souren, M. et Krieg, S. (2022). Estimating monthly Labour Force Figures during the COVID-19 pandemic in the Netherlands. *Journal of the Royal Statistical Society, Series A*, 185, 1560-1583.
- Wulansari, I.Y. (2026). *Small Area Estimation with Measurement Error in Covariates: Simulation Studies and Applications in Official Statistics*. Thèse de doctorat (à venir), School of Mathematical and Physical Sciences, University of Technology Sydney, Australie.