Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Modélisation de la mesure intra-annuelle dans les données administratives et d'enquête couplées

par Jetske Marcelis, Arnout van Delden, Sander Scholtus et Femke Kessels

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Modélisation de la mesure intra-annuelle dans les données administratives et d'enquête couplées

Jetske Marcelis, Arnout van Delden, Sander Scholtus et Femke Kessels[1]

Résumé

Au Bureau central de la statistique des Pays-Bas (CBS), pour certains secteurs économiques, deux séries d'indices de chiffre d'affaires intra-annuels partiellement indépendantes sont disponibles : une série mensuelle fondée sur des données d'enquête et une série trimestrielle fondée sur les données de la taxe sur la valeur ajoutée pour les petites unités et sur des données d'enquête réutilisées pour les autres unités. Le CBS vise à étalonner la série mensuelle d'indices de chiffre d'affaires aux données trimestrielles du recensement à une fréquence trimestrielle. Pour l'heure, cela n'est pas réalisable, car les données fiscales ont une distribution trimestrielle différente, le chiffre d'affaires étant relativement grand au quatrième trimestre de l'année et plus faible au premier trimestre. Dans la présente étude, nous cherchons à décrire cette tendance trimestrielle présentant un écart au niveau micro. Nous avons élaboré auparavant un modèle de mélange utilisant des niveaux de chiffre d'affaires absolus pouvant expliquer en partie les distributions trimestrielles. Étant donné que les niveaux de chiffre d'affaires absolus diffèrent entre les deux séries, nous utilisons dans la présente étude un modèle fondé sur les niveaux de chiffre d'affaires trimestriels relatifs au cours d'une année.

Mots clés : modèles de mélange; erreurs de mesure; erreurs de déclaration; données fiscales; tendances saisonnières.

1. Introduction

Comme d'autres pays, le Bureau central de la statistique des Pays-Bas utilise les données de la taxe sur la valeur ajoutée (TVA) pour estimer les niveaux et les variations de chiffre d'affaires intra-annuels. Dans bon nombre de ces cas, l'organisme statistique a d'abord utilisé des enquêtes pour estimer les variations ou les niveaux de chiffre d'affaires, puis a commencé à étudier la possibilité de les remplacer par la TVA. Quand on compare les valeurs de chiffre d'affaires dérivées de la TVA à celles obtenues par une enquête, on trouve des erreurs de mesure dans les deux types de sources. Des exemples d'études sur les erreurs de mesure dans les données de la TVA sont donnés dans Țîru et coll. (2019) et Lewis et Woods (2013).

Dans le présent article, nous traitons d'une forme particulière d'erreur de mesure, susceptible de se produire dans les données de la TVA, à savoir l'occurrence de ce qu'on appelle les effets trimestriels, soit des valeurs de chiffre d'affaires relativement importantes au quatrième trimestre de l'année et des valeurs relativement faibles au premier trimestre. À notre connaissance, ces tendances se produisent parce qu'à la clôture de leur année comptable, les entreprises apportent des corrections pour s'assurer de l'exactitude du montant annuel de la TVA. Ce phénomène se produit vraisemblablement dans d'autres pays aussi. Nous pourrions étudier ces tendances saisonnières puisque, pour certains secteurs économiques aux Pays-Bas, deux séries d'indices de chiffre d'affaires intra-annuels partiellement indépendantes sont disponibles : une série mensuelle et une série trimestrielle. La série d'indices mensuelle est fondée sur une enquête-échantillon et sert à produire des résultats pour les statistiques à court terme (SCT). La série d'indices trimestrielle comprend le chiffre d'affaires tiré des données de la taxe sur la valeur ajoutée (TVA) pour les petites entreprises et les entreprises simples ainsi que la réutilisation des données d'enquête mensuelles (agrégées à un trimestre) pour les entreprises plus complexes. Parce que les deux sources incluses dans la série trimestrielle couvrent presque toutes les unités de la population cible, elles sont aussi appelées données du recensement. À partir des données trimestrielles du recensement, on calcule les totaux annuels du chiffre d'affaires qui servent ensuite à caler les résultats des statistiques structurelles sur les entreprises (SSE). Ces SSE sont ensuite entrées dans la comptabilité nationale. Cette étape de calage s'explique par le fait que les données du recensement sur le chiffre d'affaires sont considérées

[1] Tous les auteurs, Bureau central de la statistique des Pays-Bas, P.O. Box 24500, La Haye, Pays-Bas, 2490 HA (j.marcelis@cbs.nl, a.vandelden@cbs.nl (auteur correspondant), s.scholtus@cbs.nl, flm.kessels@cbs.nl).

comme étant de meilleure qualité puisqu'elles ne sont pas sujettes aux erreurs d'échantillonnage. La comptabilité nationale est publiée dans différentes diffusions : les diffusions les plus tardives sont basées sur les SSE, tandis que les diffusions hâtives sont basées sur les SCT. Par conséquent, les différences entre la série de données d'enquête mensuelle et la série du recensement trimestrielle peuvent entraîner des différences entre les diffusions hâtives et tardives de la comptabilité nationale.

Le Bureau central de la statistique des Pays-Bas aimerait étalonner la série mensuelle d'indices de chiffre d'affaires aux données trimestrielles du recensement à une fréquence trimestrielle. Cela permettrait non seulement d'assurer la cohérence entre les chiffres trimestriels de l'enquête et les données du recensement, mais aussi de réduire les ajustements de la comptabilité nationale, car les inexactitudes de la série d'enquêtes sont corrigées chaque trimestre et ne s'accumulent pas. La série de données d'enquête est sujette à la variance, surtout dans les industries à petite taille d'échantillon. De plus, on utilise un échantillon défini par un seuil d'inclusion dans lequel les petites entreprises ne sont pas observées, ce qui pourrait entraîner un (petit) biais.

Quand on applique cet étalon, les taux de croissance de la série mensuelle des diffusions finales sont ajustés sous réserve que les indices ajustés par trimestre soient identiques aux indices trimestriels des données du recensement. Nous avions auparavant comparé différentes méthodes d'étalonnage, comme l'ajustement par le quotient et la méthode de Denton (Daalmans, 2018). En 2016, le CBS a étalonné la série mensuelle de 2015 avec la série trimestrielle de la même année au moyen de la méthode de Denton (voir Bikker et coll., 2013; Denton, 1971). Cela a donné lieu à des résultats inattendus puisque, pour la majorité des secteurs économiques, les taux de croissance d'une année à l'autre (sur 12 mois) du chiffre d'affaires trimestriel provenant des données de l'enquête-échantillon ont été ajustés à la baisse au premier trimestre de l'année et à la hausse au quatrième trimestre de l'année (voir Van Delden et Scholtus, 2017). À titre d'exemple, pour le secteur économique Commerce de détail, les ajustements des taux de croissance sur 12 mois du chiffre d'affaires trimestriel pour les quatre trimestres ont été de –0,5, +0,5, +0,2 et +1,0. Un examen plus attentif a montré que cet effet était dû en partie à l'utilisation de la méthode de Denton et en partie aux effets trimestriels des données de la TVA.

Van Delden et Scholtus (2017) ont effectué une première analyse des différences de chiffre d'affaires trimestriel entre les deux sources en 2014 et 2015. Ils ont analysé les données de TVA et les données d'enquête observées couplées au niveau de l'entreprise, pour toutes les entreprises pour lesquelles des données observées des deux sources étaient disponibles. Ils ont analysé les données en utilisant un modèle de régression linéaire pour les données trimestrielles avec une pente pouvant varier pour chacun des trimestres et une méthode de régression pouvant s'adapter aux valeurs aberrantes. Pour traiter les valeurs aberrantes, ils ont testé un modèle de régression linéaire robuste ainsi qu'un modèle de mélange à deux groupes ressemblant à celui de Di Zio et Guarnera (2013). Dans ce modèle à deux groupes, un groupe avait une variance résiduelle de taille moyenne et un deuxième groupe avait une variance résiduelle importante. Les effets saisonniers constatés au moyen de ce modèle à deux groupes étaient relativement petits et n'étaient pas entièrement uniformes dans les différents secteurs économiques. Pour ce qui est de l'étape suivante, Van Delden et coll. (2020) ont élaboré un modèle de mélange élargi dans lequel ils cherchent à expliquer l'effet trimestriel observé en utilisant différents groupes dont la taille de l'effet trimestriel et la taille et la structure des (co-)variances peuvent varier. Van Delden et coll. (2020) ont comparé un modèle à six groupes avec le modèle précédent à deux groupes pour le secteur économique Placement professionnel; ils ont constaté qu'ils pouvaient expliquer plus d'effets trimestriels au moyen du modèle à six groupes.

Les résultats pour d'autres secteurs économiques (Commerce de détail, Fabrication, Construction) ont montré que ce modèle de mélange élargi pouvait toutefois expliquer seulement une partie des effets saisonniers. L'un des problèmes auxquels nous nous heurtons est que les niveaux de chiffre d'affaires annuels totaux indiqués par la TVA sont plus élevés que ceux indiqués dans l'enquête-échantillon (voir la **Figure 3-1** ci-dessous). Pour autant que nous sachions, le chiffre d'affaires au niveau de l'entreprise basé sur les déclarations de TVA est parfois un peu trop important, car il n'est pas corrigé pour tenir compte des livraisons internes entre les unités juridiques d'une entreprise. Par ailleurs, le chiffre d'affaires indiqué par l'enquête est parfois trop petit, par exemple quand les entreprises déclarent dans l'enquête uniquement le chiffre d'affaires de leurs activités principales, sans déclarer celui de leurs activités secondaires. Le modèle de mélange élargi a parfois pris en charge des groupes de la population dont le taux de chiffre d'affaires indiqué par la TVA était plus élevé que le taux de chiffre d'affaires de l'enquête, alors que nous cherchions à expliquer les effets saisonniers trimestriels. C'est pourquoi nous présentons dans l'article un nouveau modèle de mélange, dans lequel nous utilisons les valeurs de chiffre d'affaires relatif d'une même année. Nous éliminons ainsi les effets des différences de taux de chiffre d'affaires annuel. Le présent article vise à déterminer dans quelle mesure

le nouveau modèle parvient à décrire les tendances trimestrielles présentant un écart. Cela doit finalement permettre de corriger les données de la TVA pour tenir compte de ces effets trimestriels.

Le reste de l'article s'organise comme suit. Nous commençons par la section 2 qui décrit le modèle de mélange fondé sur un chiffre d'affaires relatif. Ensuite, à la section 3, nous décrivons les données empiriques pour lesquelles nous essayons d'expliquer les effets trimestriels. La section 4 présente des essais du nouveau modèle de mélange et, à la section 5, nous appliquons le modèle aux données de Placement professionnel. Enfin, la section 6 analyse les résultats.

2. Modèle de mélange pour différences de chiffre d'affaires relatif

Dans cette section, nous présentons un modèle de mélange pour détecter les entreprises dont les modèles de chiffre d'affaires trimestriel présentent un écart entre les données d'enquête et celles de la TVA. Pour une année donnée, supposons que $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ et $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})'$ désignent les vecteurs observés de quatre valeurs trimestrielles de chiffre d'affaires de l'entreprise i, respectivement dans les données d'enquête et de TVA. Nous divisons chaque vecteur par son total (chiffre d'affaires annuel) pour obtenir des vecteurs de valeurs de chiffre d'affaires trimestriel relatif, $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})'$ et $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$, avec $\sum_{k=1}^{4} y_{ik} = \sum_{k=1}^{4} x_{ik} = 1$. Par souci de simplicité, nous supposons tous les $y_{ik} \geq 0$ et $x_{ik} \geq 0$ et nous élaborons le modèle de mélange uniquement pour les entreprises qui satisfont à cette hypothèse. Les $\boldsymbol{y}_i$ et $\boldsymbol{x}_i$ sont des exemples de ce qu'on appelle des *données compositionnelles*; voir p. ex. Aitchison (1986).

La distribution de Dirichlet à $(p-1)$ dimensions offre un moyen souple de modéliser les données compositionnelles, c.-à-d. les vecteurs $\boldsymbol{a} = (a_1, \ldots, a_p)'$ avec tous les $a_k \geq 0$ et $\sum_{k=1}^{p} a_k = 1$. La fonction de densité de la probabilité de cette distribution est donnée par :

$$f_{Dir}(\boldsymbol{a}; \boldsymbol{\beta}) = \frac{\Gamma\left(\sum_{k=1}^{p} \beta_k\right)}{\prod_{k=1}^{p} \Gamma(\beta_k)} \prod_{k=1}^{p} a_k^{\beta_k - 1}, \tag{1}$$

où $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ désigne un vecteur de paramètres positifs et $\Gamma(u) = \int_0^\infty v^{u-1} e^{-v} \mathrm{d}v$ est la fonction gamma. Soit $\beta_{tot} = \sum_{k=1}^{p} \beta_k$ et $\beta_k^* = \beta_k / \beta_{tot}$. Les premiers et deuxièmes moments centraux marginaux de la distribution de Dirichlet dans (1) sont donnés par $E(a_k) = \beta_k^*$ et $\mathrm{var}(a_k) = \beta_k^*(1 - \beta_k^*)/(\beta_{tot} + 1)$. Nous constatons qu'en augmentant (diminuant) tous les β_k par le même facteur, nous pouvons obtenir une distribution avec le même point central, mais des variances plus petites (plus grandes).

Pour modéliser les différences entre $\boldsymbol{y}_i$ et $\boldsymbol{x}_i$ au moyen des distributions de Dirichlet, nous définissons un vecteur de différence transformé :

$$\boldsymbol{d}_i = \frac{1}{4} \iota_4 - \frac{\boldsymbol{y}_i - \boldsymbol{x}_i}{4}, \tag{2}$$

où $\iota_4 = (1,1,1,1)'$. Les éléments du vecteur $\boldsymbol{d}_i$ satisfont à $\sum_{k=1}^{4} d_{ik} = 1$ avec $0 \leq d_{ik} \leq 1/2$. En l'absence de différences systématiques entre les distributions de $\boldsymbol{y}_i$ et $\boldsymbol{x}_i$, nous nous attendons à ce que $E(d_{ik}) = 1/4$ pour tous les k.

Comme dans Van Delden et coll. (2020), nous autorisons plusieurs sous-populations (groupes) d'entreprises, chacune ayant une relation différente entre les valeurs d'enquête et celles de la TVA. Notre modèle proposé pour $\boldsymbol{d}_i$ est un mélange de distributions de Dirichlet :

$$f(\boldsymbol{d}_i) = \prod_{g=1}^{G} \left\{ \alpha_g \cdot f_{Dir}\left(\boldsymbol{d}_i; \kappa_g \left(\frac{1}{4}\iota_4 + \boldsymbol{\delta}_g\right)\right)\right\}^{z_{gi}}. \tag{3}$$

Ici, G indique le nombre de groupes et $z_{gi} \in \{0,1\}$ est un indicateur notant si l'unité i appartient au groupe g (avec $\sum_{g=1}^{G} z_{gi} = 1$). Dans chaque groupe, on suppose une distribution de Dirichlet ayant la forme (1) avec un vecteur de paramètre de forme $\boldsymbol{\beta}_g = \kappa_g(\iota_4/4 + \boldsymbol{\delta}_g)$. Le paramètre scalaire κ_g détermine l'ampleur de la variance dans le groupe g, tandis que les paramètres $\boldsymbol{\delta}_g = (\delta_{g1}, \delta_{g2}, \delta_{g3}, \delta_{g4})'$ décrivent les différences trimestrielles systématiques potentielles entre les données d'enquête et les données de la TVA. Nous utilisons la restriction naturelle selon laquelle $\sum_{k=1}^{4} \delta_{gk} = 0$. Dans certains groupes, nous pouvons ajouter la restriction $\boldsymbol{\delta}_g = \boldsymbol{0}$, ce qui indique qu'il n'y a pas de différences systématiques pour les entreprises de ce groupe. Enfin, les paramètres du modèle $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_G)'$

désignent les tailles relatives des différents groupes dans la population, avec $\sum_{g=1}^{G} \alpha_g = 1$. L'ensemble complet des paramètres du modèle est noté par $\boldsymbol{\theta}$ et contient $\boldsymbol{\alpha}$, $\kappa_1, \ldots, \kappa_G$ et tous les $\boldsymbol{\delta}_g$ qui n'ont pas été restreints à $\mathbf{0}$.

En pratique, les indicateurs de groupe z_{gi} ne sont pas observés. Pour estimer un modèle de mélange de forme (3), nous pouvons utiliser un algorithme de maximisation de l'espérance (conditionnelle) (E(C)M) (McLachlan et Peel, 2000). Dans cet algorithme, deux étapes sont répétées jusqu'à la convergence :

- Étape E : compte tenu des estimations des paramètres actuels $\boldsymbol{\theta}$, évaluer $\tau_{gi} = E\left(z_{gi} \middle| \boldsymbol{d}_i, \boldsymbol{\theta}\right) = P\left(z_{gi} = 1 \middle| \boldsymbol{d}_i, \boldsymbol{\theta}\right)$.
- Étape M : mettre à jour les estimations des paramètres en maximisant la fonction de log-vraisemblance basée sur (3), tous les z_{gi} inconnus étant remplacés par leur valeur espérée τ_{gi}.

Pour simplifier les calculs dans l'étape M, les paramètres $\kappa_1, \ldots, \kappa_G$ et $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_G$ sont mis à jour séparément, conditionnellement aux valeurs actuelles des autres paramètres, ce qui en fait un algorithme ECM plutôt qu'un algorithme EM. L'algorithme nécessite des valeurs de départ pour $\boldsymbol{\theta}$ et peut converger vers une solution sous-optimale en fonction de ces valeurs de départ. Pour trouver la solution optimale, il faut essayer plusieurs ensembles de valeurs de départ et seule la solution ayant la meilleure valeur de la fonction de log-vraisemblance est retenue.

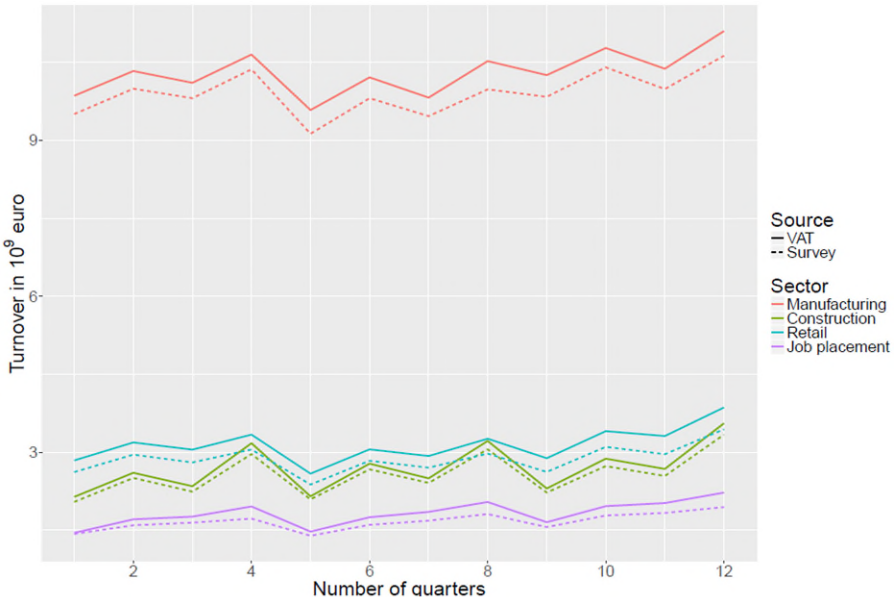
Afin de trouver la meilleure spécification du modèle (3) pour les données disponibles – ce qui comprend la sélection du nombre de groupes G et la détermination des groupes contenant des différences systématiques représentées par $\boldsymbol{\delta}_g$ – nous pouvons ajuster plusieurs modèles et comparer leur critère d'information d'Akaike (AIC) et leur critère d'information bayésien (BIC) en fonction de la log-vraisemblance et du nombre de paramètres du modèle. La troisième mesure d'ajustement possible est l'ICL-BIC, qui étend le BIC pour tenir également compte de la mesure dans laquelle le modèle estimé est capable d'attribuer des unités à un seul groupe en se fondant sur τ_{gi}. Pour en savoir plus sur ces mesures d'ajustement, voir McLachlan et Peel (2000).

3. Données empiriques

Dans cet article, nous utilisons les données de quatre secteurs économiques (Fabrication, Construction, Commerce de détail et Placement professionnel) et de trois années (2014 à 2016) pour étudier les différences de tendances saisonnières entre les données d'enquête et les données de la TVA. Pour les secteurs de la fabrication, de la construction et du commerce de détail, nous disposons de données d'enquête mensuelles. La sortie des statistiques à court terme du Placement professionnel était fondée sur les données d'enquête trimestrielles de 2014 à 2016, mais pour les années ultérieures, la sortie des SCT était fondée sur les données du recensement, car la série d'enquêtes est terminée. Nous avons inclus les données sur le placement professionnel parce que ce secteur a des effets saisonniers clairs (voir Van Delden et Scholtus, 2019) et que ces données convenaient particulièrement bien à l'élaboration du modèle. Par conséquent, la plupart des résultats de la section 0 se rapportent au placement professionnel. Puisque les données que nous avons utilisées ont été décrites dans Van Delden et Scholtus (2017) et dans Van Delden et coll. (2020), nous les présenterons très brièvement. Chacun des quatre secteurs économiques est divisé en plusieurs sous-secteurs. Nous estimons les tendances saisonnières par secteur économique plutôt que par sous-secteur, car les différences entre les tendances saisonnières des sous-secteurs se sont révélées trop subtiles pour être estimées.

Pour les entreprises qui ont répondu à l'enquête-échantillon, nous avons couplé les valeurs de chiffre d'affaires de la TVA. Nous avons inclus seulement les entreprises qui ont répondu à l'enquête-échantillon pour les quatre trimestres de l'année et pour lesquelles le chiffre d'affaires de la TVA était disponible pour les quatre trimestres de l'année. Pour les unités où $Y_{ik}/X_{ik} \geq 100$ ou $Y_{ik}/X_{ik} \leq 0{,}01$, nous avons supposé qu'il y avait une grande erreur dans Y_{ik} ou X_{ik}; ces unités ont donc été omises. De plus, nous avons omis les sous-secteurs des secteurs économiques pour lesquels les estimations des variations ou du niveau de chiffre d'affaires fondées sur la TVA ont été jugées non fiables en raison des différences de définition entre le chiffre d'affaires indiqué par la TVA et l'enquête-échantillon (Van Delden et coll., 2016). Les poids de post-strate (w_{ki}=$w_{k\ell}$) pour les entreprises i dans le trimestre q de la strate ℓ ont été calculés comme étant le rapport de la taille de la population ($N_{k\ell}$) à la taille des unités incluses ($n_{k\ell}$). Pour ce qui est de la strate, nous avons utilisé une combinaison de classes de taille à un chiffre et de sous-secteurs. Notons que tous les résultats sur les données de la TVA et de l'échantillon présentés dans notre étude se rapportent aux petites unités et aux unités simples (les deux incluant les unités et les populations correspondantes), seules unités pour lesquelles nous avons les deux sources.

Figure 3-1
Chiffre d'affaires total estimé des petites unités et des unités simples, à partir des données de TVA et d'enquête, quand les deux types de données sont disponibles. Les trimestres sont numérotés à partir du premier trimestre de 2014.



Nous avons estimé les niveaux totaux de chiffre d'affaires trimestriels comme étant $\widehat{\boldsymbol{Y}}_{\boldsymbol{k}} = \sum_{\boldsymbol{i}} \boldsymbol{w}_{\boldsymbol{ki}} \boldsymbol{y}_{\boldsymbol{ki}}$ pour le chiffre d'affaires de l'enquête et $\widehat{\boldsymbol{X}}_{\boldsymbol{k}} = \sum_{\boldsymbol{i}} \boldsymbol{w}_{\boldsymbol{ki}} \boldsymbol{x}_{\boldsymbol{ki}}$ pour le chiffre d'affaires de la TVA (**Figure 3-1**). Pour l'ensemble des trimestres et des secteurs économiques, les totaux de chiffre d'affaires estimés étaient plus élevés pour les données de la TVA que pour les données d'enquête. De plus, la différence entre le chiffre d'affaires de la TVA et des enquêtes est souvent plus grande au quatrième trimestre de l'année et plus petite au premier trimestre de l'année. Cela se voit particulièrement dans le secteur du placement professionnel (ligne continue violette).

4. Mise à l'essai du modèle

Au moyen de données simulées, nous avons effectué des essais pour savoir dans quelle mesure le modèle de mélange de la section 2 peut être estimé de façon fiable. Ces données simulées sont générées à partir d'un mélange de distributions de Dirichlet. Nous avons basé les valeurs des paramètres de cette distribution sur les estimations des paramètres d'un modèle de mélange appliquées aux ensembles de données disponibles (c.-à-d. les ensembles de données des quatre secteurs économiques en 2014, 2015 et 2016) afin d'obtenir des données simulées réalistes. Ainsi, nous avons d'abord appliqué un modèle à trois groupes à ces données afin de déterminer des valeurs communes pour les tailles d'échantillon et les paramètres libres du modèle de mélange. Nous avons créé deux ensembles différents de valeurs de paramètres pour examiner les données ayant un effet trimestriel relativement fort (fondé sur le secteur Placement professionnel) et les données ayant un effet trimestriel relativement faible (fondé sur le secteur Fabrication). Ces ensembles sont affichés dans le **Tableau 4-1**. Pour les valeurs $\boldsymbol{\delta}_{\boldsymbol{g}}$, les valeurs du troisième groupe sont affichées, car c'est le seul groupe ayant un effet systématique (c.-à-d. $\boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \mathbf{0}$ dans ces simulations). De plus, nous avons effectué toutes les simulations en utilisant la taille d'échantillon des ensembles de données originaux sur lesquels les ensembles de paramètres sont fondés (voir la colonne « taille » dans le **Tableau 4-1**) et en utilisant la plus petite taille d'échantillon des ensembles de données disponibles (750 unités) afin de déterminer si les simulations donnent des résultats semblables en cas d'échantillon plus petit. Nous avons réalisé trois études par simulations (voir les sections 4.1 – 4.3). Pour chaque étude par simulations, nous avons généré 100 fois des données à partir d'un mélange de distributions de Dirichlet avec ces ensembles de paramètres.

Tableau 4-1

Valeurs de paramètres basées sur les données de Fabrication (ensemble 1) et Placement professionnel (ensemble 2) qui serviront à former des distributions de Dirichlet prédéfinies à partir desquelles les données simulées seront générées.

	Taille	$\boldsymbol{\alpha_g}$			$\boldsymbol{\kappa_g \cdot 10^3}$			$\boldsymbol{\delta_{3k} \cdot 10^{-4}}$			
		α_1	α_2	α_3	κ_1	κ_2	κ_3	δ_{31}	δ_{32}	δ_{33}	δ_{34}
Ensemble 1	1 100	0,25	0,15	0,60	0,15	700	5	-20	0	0	20
Ensemble 2	2 250	0,25	0,20	0,55	0,5	2 000	20	-3,5	-0,25	0	3,75

4.1 Trouver les estimations des paramètres

Dans la première étude par simulations, nous avons cherché à déterminer dans quelle mesure les valeurs estimées des paramètres se rapprochaient des valeurs réelles, étant donné un nombre exact de groupes. Nous avons examiné quatre scénarios différents qui forment un plan de sondage 2 × 2, dans lequel les paramètres étaient des tailles de groupe égales ($\alpha_1, \alpha_2, \alpha_3 = 1/3$) comparativement à des tailles de groupe inégales ($\alpha_1, \alpha_2, \alpha_3$ selon le **Tableau 4-1**) et de bonnes valeurs de départ comparativement à de mauvaises valeurs de départ. Dans l'ensemble, l'algorithme ECM a de bonnes performances : il a estimé correctement les vraies valeurs des paramètres à partir de valeurs de départ raisonnables. Il y avait seulement un léger biais dans l'estimation de certaines valeurs $\boldsymbol{\delta_g}$ dans les données de l'ensemble de paramètres 2. Pour les simulations avec des tailles de groupe inégales et de mauvaises valeurs de départ (scénario 4), l'algorithme ECM pouvait atteindre un maximum local dans le cas des données basées sur l'ensemble de paramètres 2. Cela a entraîné un biais et une variance dans les estimations des paramètres. La variabilité des estimations des valeurs α_g et κ_g était petite dans les scénarios 1, 2 et 3. Cependant, les erreurs-types des valeurs $\boldsymbol{\delta_g}$ étaient relativement grandes dans tous les scénarios. Un examen plus attentif des résultats a montré que la variabilité et le léger biais dans les estimations pour $\boldsymbol{\delta_g}$ dans les scénarios 1, 2 et 3 sont principalement dus au fait que les valeurs de $\boldsymbol{\delta_g}$ dans les échantillons diffèrent de celles de la population, ce qui indique un effet d'échantillonnage. En particulier, nous n'avons pas constaté de grande amélioration de l'exactitude des estimations quand les vrais indicateurs de groupe z_{gi} ont été utilisés. Pour les simulations avec la plus petite taille d'échantillon (750 unités), les estimations des paramètres ressemblaient tout aussi bien aux vraies valeurs des paramètres, mais la variance augmentait par rapport aux simulations avec la taille d'échantillon initial.

4.2 Trouver le nombre de groupes

Dans la deuxième étude par simulations, nous avons cherché à savoir si l'algorithme ECM pouvait récupérer le vrai nombre de groupes à partir de valeurs de départ raisonnables. Nous avons utilisé différents modèles dans lesquels nous faisions passer le nombre de groupes de deux à sept. Sept était le nombre maximal de groupes déterminé dans la méthode antérieure de Van Delden et coll. (2020). Pour comparer les performances des différents modèles, nous avons calculé les critères AIC, BIC et ICL-BIC. Nous avons ensuite déterminé si le modèle ayant le bon nombre de groupes avait les meilleures mesures d'ajustement dans la plupart des simulations. Pour les deux ensembles de paramètres et les deux tailles d'échantillon, les résultats ont montré que l'algorithme ECM était en mesure de récupérer le vrai nombre de groupes dans presque toutes les simulations, voire toutes, selon la mesure d'ajustement.

4.3 Tester l'effet des valeurs de départ

Dans la troisième étude par simulations, nous avons analysé davantage l'effet de différentes valeurs de départ sur les performances du modèle de mélange. Nous avons constaté que dans les simulations ayant des valeurs de départ raisonnables, l'algorithme a correctement estimé les paramètres, tandis que dans les simulations ayant de mauvaises valeurs de départ, l'algorithme ECM a parfois atteint des maximums locaux. En cas de maximum local, les mesures d'ajustement correspondantes étaient moins optimales. L'algorithme était le plus sensible aux valeurs de départ pour κ_g. C'est pourquoi dans ce qui suit (section 5), nous avons utilisé plusieurs ensembles de valeurs de départ, et plus particulièrement utilisé des valeurs de départ suffisamment variables pour κ_g, et nous avons sélectionné le modèle final en fonction des meilleures mesures d'ajustement.

5. Appliquer le modèle au Placement professionnel

Afin d'évaluer dans quelle mesure l'algorithme pouvait détecter les tendances saisonnières, nous l'avons appliqué au secteur Placement professionnel pour les années 2014 à 2016. Comme nous l'avons mentionné ci-dessus, nous avons utilisé le placement professionnel en raison des tendances saisonnières claires qu'il présente (Van Delden et Scholtus, 2019). Nous avons appliqué des modèles de mélange ayant la forme (3) sur deux à six groupes. Les modèles avec de deux à quatre groupes comportaient un groupe à effet systématique, tandis que les modèles à cinq ou six groupes comportaient deux groupes à effet systématique. Dans tous les ensembles de données, les modèles à deux et à trois groupes ont donné des résultats inférieurs aux résultats des modèles à quatre ou six groupes. C'est pourquoi nous analyserons uniquement les modèles à quatre, cinq et six groupes.

5.1 Modèles de base

Les groupes des modèles à quatre, cinq et six groupes sont spécifiés dans le **Tableau 5.1-2**. Dans le modèle à quatre groupes, il y avait trois groupes dont le degré d'erreur de mesure variait et un groupe ayant des effets trimestriels (différences systématiques entre la distribution trimestrielle des données de la TVA et des données d'enquête). Pour obtenir le modèle à cinq groupes, le modèle à quatre groupes a été élargi d'un groupe supplémentaire avec des effets trimestriels. Le modèle à six groupes comprend également un groupe supplémentaire sans effet trimestriel. De plus, il y avait de petites différences dans la taille de la variance entre les groupes des trois modèles.

Tableau 5.1-2.
Spécification des modèles à quatre, cinq et six groupes sur les données de placement professionnel de 2014, 2015 et 2016. Cette spécification est conforme aux valeurs de départ et aux estimations finales des paramètres. Le nombre de signes plus indique la taille relative de la variance (+++++ très grande, + très petite).

	Modèle à quatre groupes		Modèle à cinq groupes		Modèle à six groupes	
	Variance	Effet systématique	Variance	Effet systématique	Variance	Effet systématique
Groupe 1	+++++	Non	+++++	Non	+++++	Non
Groupe 2	+	Non	+	Non	++	Non
Groupe 3	+++	Non	++++	Non	++++	Non
Groupe 4	+++	Oui	+++	Oui	+++	Oui
Groupe 5			++	Oui	+	Non
Groupe 6					++++	Oui

5.2 Résultats

Pour ce qui est du nombre optimal de groupes, les trois années ont donné les mêmes résultats : le modèle à six groupes était préférable selon l'AIC et le BIC, tandis que le modèle à cinq groupes était préférable selon l'ICL-BIC.

Nous avons utilisé la procédure suivante pour analyser dans quelle mesure le modèle pouvait détecter les tendances saisonnières : nous avons déterminé le rapport entre le chiffre d'affaires absolu total selon les données d'enquête (Y_k) et de la TVA (X_k) par trimestre de l'année, sans et avec ajustement des effets trimestriels détectés par le modèle. Nous avons utilisé les valeurs $\boldsymbol{\delta}$ estimées des groupes ayant un effet trimestriel pour calculer les valeurs trimestrielles ajustées de chiffre d'affaires de la TVA à partir des valeurs originales X_{ki} et leurs totaux ajustés correspondants, notés par $\tilde{X}_k$. Pour les totaux, les chiffres d'affaires sont pondérés au moyen des poids de post-strate (w_{ki}).

Dans la **Figure 5.2-1**, les rapports Y_k/X_k sont affichés pour le modèle à six groupes, pour les trois années séparément. Ces rapports ont été déterminés pour les deux groupes présentant des différences systématiques, pour tous les groupes (indiqués par une couleur), et pour le chiffre d'affaires de la TVA ajusté et non ajusté (indiqué par des lignes continues et pointillées, respectivement). Dans tous les cas, l'ajustement de X_{ki} n'est appliqué qu'au(x) groupe(s) présentant des différences systématiques. Dans ces graphiques, les rapports Y_k/X_k avec le chiffre d'affaires de la TVA non ajusté déterminé pour tous les groupes ont montré deux types de différences entre les deux séries chronologiques. Premièrement, les rapports Y_k/X_k n'étaient pas égaux à un, mais inférieurs à un. Cela indique que $X_k > Y_k$, c'est-à-dire que les totaux de chiffre d'affaires de la TVA étaient plus élevés que les totaux de chiffre d'affaires d'enquête, ce

qui correspond à une différence de niveau. Deuxièmement, les rapports Y_k/X_k présentaient un écart par rapport à une ligne horizontale : les rapports variaient par trimestre, ce qui correspond aux différences trimestrielles. Donc, pour notre application, nous cherchions à ce que les rapports Y_k/X_k après l'ajustement soient plus horizontaux que les rapports non ajustés. La F**igure 5.2-1** montre que l'ajustement du chiffre d'affaires de la TVA a effectivement entraîné des effets trimestriels plus petits. Autrement dit, l'ajustement donne une tendance plus stable des rapports Y_k/X_k au cours de l'année, ce qui signifie que le modèle de mélange actuel a expliqué en partie les effets saisonniers. Particulièrement pour 2014, nous avons constaté que le deuxième groupe (la ligne verte) présentait un effet saisonnier clair, qui a été bien ajusté par le modèle.

Figure 3-1

Les rapports Y_k/X_k sont fondés sur le modèle à six groupes pour le placement professionnel de 2014 à 2016. Les rapports sont déterminés pour les deux groupes présentant des différences systématiques et pour tous les groupes, ainsi que pour le chiffre d'affaires de la TVA ajusté et non ajusté.

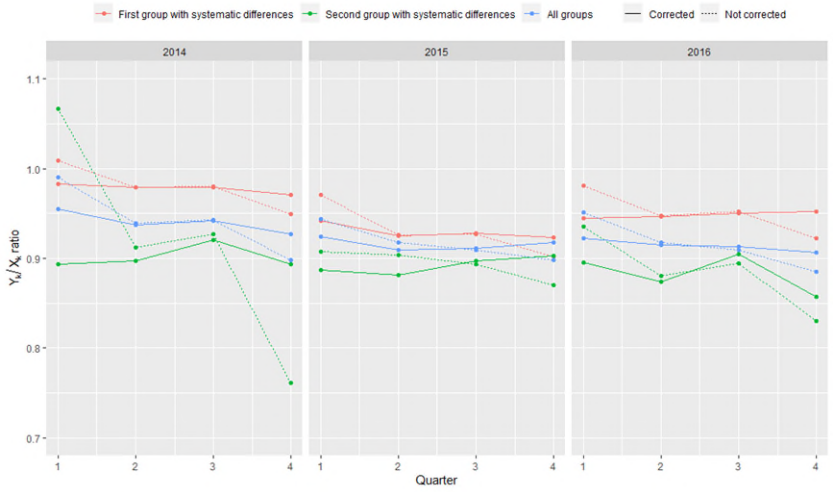
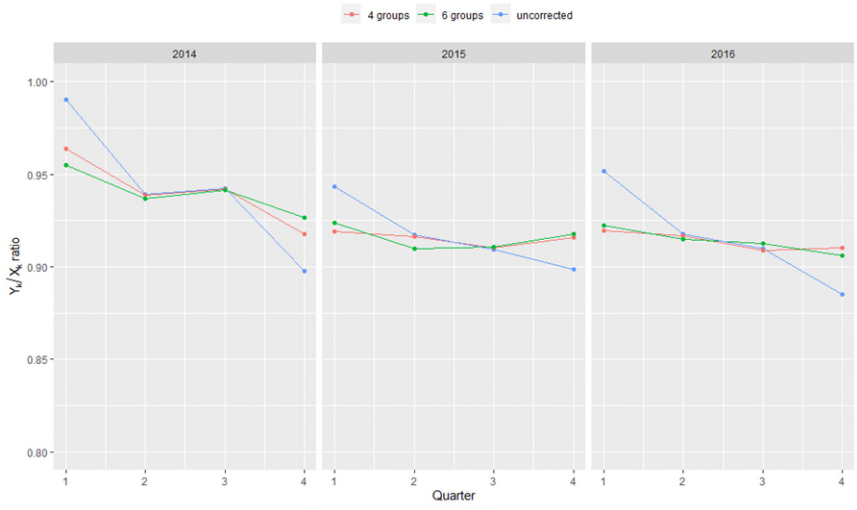


Figure 5.2-2

Rapports Y_k/X_k avec le chiffre d'affaires de la TVA ajusté pour Placement professionnel de 2014 à 2016 pour le modèle à quatre et six groupes. De plus, les rapports Y_k/X_k avec le chiffre d'affaires de la TVA non ajusté sont inclus.



Ensuite, nous avons comparé les rapports Y_k/X_k des modèles à quatre, cinq et six groupes. Dans la **Figure 5.2.-2**, nous avons seulement affiché les rapports des modèles à quatre et à six groupes, car ces modèles pouvaient expliquer davantage les effets saisonniers que le modèle à cinq groupes. Ici, les rapports ont été déterminés pour tous les groupes.

Pour les deux modèles, l'ajustement de X_{ki} a donné des lignes présentant des écarts moindres entre les trimestres que la ligne du chiffre d'affaires de la TVA non ajusté. Les deux modèles ont par conséquent expliqué une partie des effets saisonniers. En particulier pour 2014, le modèle à six groupes a eu de meilleures performances que le modèle à quatre groupes.

6. Conclusions et futurs travaux

Nous avons élaboré un modèle de mélange en utilisant la distribution du chiffre d'affaires trimestriel relatif au cours d'une année comme donnée d'entrée pour expliquer (et corriger) les différences saisonnières entre la distribution du chiffre d'affaires obtenue à partir des données de la TVA et des données d'enquête. De plus, nous avons appliqué un algorithme ECM pour estimer ce modèle. À partir d'études par simulations, nous concluons que l'algorithme ECM fonctionne bien : dans les conditions de nos essais et s'il dispose de valeurs de départ raisonnables, il récupère les vraies valeurs des paramètres et le bon nombre de groupes. Quand l'algorithme est lancé avec de mauvaises valeurs de départ, il peut atteindre des maximums locaux, auquel cas les mesures d'ajustement sont moins optimales. Il est donc important de lancer l'algorithme avec plusieurs valeurs de départ, surtout pour les paramètres κ_g. Nous avons appliqué le modèle aux données réelles du secteur économique Placement professionnel et avons constaté que le modèle pouvait effectivement expliquer une grande partie des effets saisonniers. Bien que le modèle présenté ici ait été conçu aux fins de notre application particulière, nous pensons que la méthode est utile dans d'autres situations où les sources comportent des erreurs de mesure intra-annuelle à la fois aléatoires et structurelles.

Plusieurs étapes devront être réalisées avant que nous puissions utiliser le modèle dans un processus de production statistique. Premièrement, nous souhaitons appliquer le modèle à d'autres secteurs économiques et à des années plus récentes. Deuxièmement, nous voulons déterminer si la probabilité qu'une unité appartienne à un certain groupe (c.-à-d. les valeurs prédites par l'algorithme ECM actuel) peut être prédite par un nouveau modèle, au moyen de données de registre disponibles, comme la distribution relative de la TVA. Nous pourrions alors employer ce nouveau modèle pour prédire l'appartenance au groupe pour *toutes* les unités de la population et calculer par la suite les valeurs corrigées de TVA au niveau micro. Cela pourrait servir à calculer des indices de chiffre d'affaires ajustés pour les données du recensement, après quoi nous pourrions comparer l'étalonnage des séries mensuelles aux indices trimestriels du recensement ajustés par rapport aux indices trimestriels originaux du recensement. Enfin, nous aimerions savoir lequel des deux modèles de mélange, celui fondé sur les valeurs de chiffre d'affaires relatif ou absolu, est le plus utile pour décrire les tendances saisonnières.

Bibliographie

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, London, UK: Chapman & Hall.

Bikker, R.P., J. Daalmans et N. Mushkudiani (2013), « Benchmarking Large Accounting Frameworks: a Generalised Multivariate Model », *Economic Systems Research*, 25, p. 390-408.

Daalmans, J. (2018), « Kwartaalinpassing KICR DRT: pro rata versus Denton », rapport interne (en néerlandais), La Haye : Statistics Netherlands.

Van Delden, A. et S. Scholtus (2017), « Correspondence between survey and administrative data on quarterly turnover », document de travail 2017-3, La Haye : Statistics Netherlands. Disponible à l'adresse : https://www.cbs.nl/en-gb/background/2017/07/correspondence-between-survey-and-admin-data-on-quarterly-turnover.

Van Delden, A. et S. Scholtus (2019), « Analysing response differences between sample survey and VAT turnover », document de travail, La Haye : Statistics Netherlands. Disponible à l'adresse : https://www.cbs.nl/en-gb/background/2019/22/analysing-response-differences-in-vat.

Van Delden, A., J. Pannekoek, R. Banning et A. de Boer (2016), « Analysing correspondence between administrative and survey data », *Statistical Journal of the IAOS*, 32, p. 569-584.

Van Delden, A., S. Scholtus et N. Ostlund (2020), « Modélisation des erreurs de mesure afin d'assurer la cohérence entre les taux de croissance du chiffre d'affaires mensuels et trimestriels », *Recueil du Symposium 2018 de Statistique Canada*. Disponible à l'adresse : https://www.statcan.gc.ca/eng/conferences/symposium2018/program.

Denton, F.T. (1971), « Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization », *Journal of the American Statistical Association*, 66, p. 99-102.

Di Zio, M. et U. Guarnera (2013), « A Contamination Model for Selective Editing », *Journal of Official Statistics*, 29, p. 539-555.

Lewis, D. et J. Woods (2013), Issues to consider when turning to the use of administrative data: the UK experience. Document présenté à la conférence New Techniques and Technologies for Statistics (NTTS) de 2013. Disponible à l'adresse : https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_142.pdf

McLachlan, G.J. et D. Peel (2000), *Finite Mixture Models*, New York : John Wiley & Sons.

Țîru, A.M., Pop, I. et B. Oancea (2019), Error detection and data imputation methods for administrative data sources used in business statistics. Document présenté à l'atelier du Système statistique européen sur les données administratives concernant les entreprises, l'agriculture et la pêche à Bucarest, en Roumanie, le 17 et le 18 octobre 2019.