Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Modélisation des erreurs de mesure afin d'assurer la cohérence entre les taux de croissance du chiffre d'affaires mensuels et trimestriels

Arnout van Delden, Sander Scholtus et Nicole Ostlund[1]

Résumé

Pour un certain nombre de secteurs économiques, le Bureau central de la statistique (CBS) des Pays-Bas produit des taux de croissance du chiffre d'affaires des entreprises : des chiffres mensuels fondés sur une enquête-échantillon et des chiffres trimestriels principalement fondés sur des données administratives. Le CBS cherche à étalonner les taux de croissance mensuels sur les taux trimestriels afin de produire des données cohérentes. Les résultats préliminaires de l'étalonnage ont révélé que les taux de croissance administratifs trimestriels du chiffre d'affaires se sont avérés relativement importants au quatrième trimestre de l'année comparativement aux données d'enquête, tandis que c'était le contraire au premier trimestre. Cet effet est probablement causé par des tendances trimestrielles dans les erreurs de mesure, attribuables par exemple aux processus administratifs au sein des entreprises. Nous présentons une méthodologie, fondée sur un modèle de régression par mélange, qui cherche à détecter automatiquement ces erreurs de mesure.

Mots-clés : Modèles par mélange; erreurs de mesure; erreurs de déclaration; données fiscales; tendances saisonnières.

1. Introduction

Pour un certain nombre de secteurs économiques, le Bureau central de la statistique (CBS) des Pays-Bas produit deux séries chronologiques sur le chiffre d'affaires : une série mensuelle fondée sur le chiffre d'affaires d'après une enquête-échantillon et une série trimestrielle fondée sur des données du recensement. Les données du recensement se composent d'une combinaison de données relatives à la taxe sur la valeur ajoutée (TVA) pour les petites entreprises simples et de données d'enquête pour les entreprises plus complexes. Les petites entreprises simples sont appelées unités X de niveau non supérieur et les entreprises plus complexes sont appelées entreprises X de niveau supérieur.

La série chronologique mensuelle est utilisée pour publier des données liées aux statistiques à court terme et sert aux diffusions préliminaires des comptes nationaux trimestriels. La somme des estimations trimestrielles fondées sur les données du recensement est utilisée pour étalonner les résultats des statistiques structurelles annuelles sur les entreprises, qui servent à leur tour aux diffusions tardives des comptes nationaux annuels. De cette façon, les différences entre les deux séries chronologiques contribuent aux différences entre la diffusion hâtive et la diffusion tardive des chiffres des comptes nationaux. Pour améliorer la qualité de nos données, nous cherchons à étalonner la série chronologique mensuelle sur la série trimestrielle, à l'aide de la méthode de Denton (Bikker et coll., 2013; Denton, 1971).

Le CBS cherche à étalonner les deux séries à partir de 2015. Cependant, les résultats préliminaires de l'étalonnage des données sur le commerce de détail de 2015 ont révélé que les taux de croissance d'une année à l'autre du chiffre d'affaires trimestriel issu de l'enquête ont été rajustés 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). Selon le trimestre, ces rajustements atteignaient presque ou dépassaient les marges de 95 p. cent pour les taux de croissance d'une année à l'autre du commerce de détail de 0,7 points de pourcentage (Scholtus et de Wolf, 2011). D'après une analyse préliminaire, Van Delden et Scholtus (2017) ont révélé que le taux de croissance original d'un trimestre à l'autre au troisième trimestre de 2015

[1]Arnout van Delden, Statistics Netherlands, P.O. Box 24500 2490 HA, La Haye, Pays-Bas (a.vandelden@cbs.nl); Sander Scholtus, Statistics Netherlands, P.O. Box 24500 2490 HA, La Haye, Pays-Bas (s.scholtus@cbs.nl); Nicole Ostlund, Statistics Netherlands, P.O. Box 24500 2490 HA, La Haye, Pays-Bas

était de 8,3 points de pourcentage, tandis qu'il s'établissait à 7,9 sans l'effet saisonnier estimé. Cette différence correspondait à un chiffre d'affaires de plus de 100 millions d'euros. Le CBS estimait que cet effet était trop important. Ce résultat a inspiré la première question qu'aborde le présent document : dans quelle mesure existe-t-il des différences saisonnières systématiques dans la déclaration des données d'enquête et des données fiscales?

L'application de méthodes automatiques d'étalonnage exige que les erreurs de mesure importantes et systématiques des deux séries aient été corrigées. La correction des erreurs de mesure dans les statistiques des entreprises fait souvent appel à une combinaison de modifications automatiques et manuelles. La modification manuelle se limite souvent à un nombre réduit d'enregistrements ayant le plus d'influence. Un deuxième objectif du présent document consiste à déterminer si les différences saisonnières observées découlent d'erreurs de mesure importantes dans un ensemble limité d'unités influentes ou sont attribuables à des erreurs systématiques dans un plus grand ensemble d'unités. Cela permettra de déterminer si la correction des erreurs peut être effectuée au moyen d'une modification manuelle ou en appliquant une méthode de correction générique.

2. Données empiriques

Nous avons comparé les chiffres d'affaires d'après les données d'enquête et la TVA des unités X de niveau non supérieur sur une base trimestrielle, en utilisant les données de 2014, de 2015 et de 2016 des secteurs économiques de la fabrication, de la construction, du commerce de détail et du placement en emploi. La fabrication, la construction et le commerce de détail sont des secteurs pour lesquels une enquête mensuelle existe. Jusqu'à tout récemment, les données du placement en emploi étaient produites sur une base trimestrielle et elles étaient fondées sur une enquête-échantillon trimestrielle. Aujourd'hui, ces données sont entièrement fondées sur la TVA. Nous avons intégré le placement en emploi à l'étude parce que les résultats préliminaires ont révélé que ce secteur pourrait présenter des effets saisonniers clairs permettant de comprendre les effets dans d'autres secteurs. Toutes nos analyses reposent sur des microdonnées qui sont classées par activité économique selon la Nomenclature statistique des activités économiques dans les communautés européennes (NACE). Le terme « secteur économique » correspond approximativement au premier chiffre du code de la NACE, tandis que le terme « industries » se rapporte à des codes plus détaillés de la NACE. Les effets saisonniers ont été analysés séparément pour chaque secteur économique, plutôt que pour chaque industrie, car les effets étaient trop subtils pour permettre de produire des estimations exactes compte tenu de la quantité de données disponibles à l'échelle de l'industrie.

La TVA et les microdonnées d'enquête qui ont été utilisées pour produire les données des deux séries chronologiques ont été couplées au niveau des unités statistiques, les entreprises, au moyen d'un numéro d'identification d'entreprise unique. Dans ces données couplées, quatre catégories d'unités ont été omises :

1. les unités qui étaient susceptibles de présenter « un millier d'erreurs » (voir la section 2.3 dans Van Delden et Scholtus, 2017);
2. les unités qui n'étaient pas présentes dans les deux ensembles de données pour l'ensemble des quatre trimestres d'une année;
3. les unités qui n'ont pas déclaré leur chiffre d'affaires pour l'ensemble des quatre trimestres de l'année;
4. les industries pour lesquelles les estimations du chiffre d'affaires ou les estimations de la variation fondées sur la TVA ne sont pas considérées fiables en raison de différences entre la TVA et l'enquête en ce qui concerne la définition du chiffre d'affaires.

Nous désignons l'ensemble d'unités finales les unités « sélectionnées ». Nous avons appliqué ces quatre sélections pour assurer que les effets saisonniers que nous relevons ne sont pas attribuables à d'autres facteurs. Van Delden et Scholtus (2017) ont révélé que les effets saisonniers n'étaient pas très sensibles à ces sélections. Le tableau 2-1 présente des chiffres de base sur la population X de niveau non supérieur.

Tableau 2-1
Chiffres de base sur la population X de niveau non supérieur par secteur économique : chiffre d'affaires total (T en 10^9 euros) d'après la TVA, taille de la population (N en 10^3 entreprises) et nombre total d'unités sélectionnées (n entreprises)

Année	Fabrication			Construction			Commerce de détail			Placement en emploi		
	T	N	n	T	N	n	T	N	n	T	N	n
2014	21,5	56,6	2 296	12,3	143,3	863	12,9	110,4	2 070	3,6	12,2	1 290
2015	22,4	58,5	2 187	13,1	149,7	740	13,5	115,1	1 590	4,0	12,5	936
2016	23,3	60,3	2 271	14,3	156,5	735	14,1	117,8	1 627	4,4	12,8	1 086

3. Y a-t-il des différences saisonnières dans la déclaration?

3.1 Méthodologie

Van Delden et Scholtus (2017) ont démontré que la relation entre les chiffres d'affaires d'après l'enquête trimestrielle et la TVA peut être bien décrite au moyen d'un modèle linéaire simple, dans lequel la pente varie selon le trimestre de l'année en conjugaison avec une ordonnée à l'origine commune. Nous avons appliqué une analyse de régression intégrant le chiffre d'affaires d'après la TVA comme variable indépendante et le chiffre d'affaires d'après l'enquête-échantillon comme variable dépendante. Nous sommes conscients que les deux sources peuvent comporter des erreurs de mesure et que les erreurs liées à la variable indépendante peuvent donner lieu à une sous-estimation des pentes de l'analyse de régression. À la section 4.2, nous décrivons les résultats obtenus pour un modèle élargi, comprenant un groupe d'unités ne présentant pas d'effet trimestriel, c.-à-d. présentant une pente annuelle. La pente de ces unités était très proche de 1, ce qui laisse croire que l'effet de la sous-estimation des pentes était presque négligeable. Il convient de souligner que, après l'exclusion des industries pour lesquelles il existe des différences entre la TVA et l'enquête en ce qui concerne la définition du chiffre d'affaires (c.-à-d., la quatrième catégorie mentionnée ci-dessus), on s'attendrait à ce que la vraie pente soit de 1 en l'absence d'erreurs de mesure aléatoires.

Nous avons par conséquent appliqué le modèle linéaire suivant pour une année donnée. Soit x_i^q le chiffre d'affaires d'après la TVA pour le trimestre q de l'entreprise i, et supposons que y_i^q soit son chiffre d'affaires d'après l'enquête-échantillon. De plus, supposons que α soit l'ordonnée à l'origine commune, que $\beta^{q=1}$ soit la pente pour le trimestre 1 et supposons que $d\beta^{q=q*}$ représente la différence dans la pente entre le trimestre $q = q^*$ et le trimestre 1. Enfin, supposons que $\delta_{q*}^q \in \{0,1\}$ est une variable nominale qui indique si $q = q^*$, avec $q^* \in \{2,3,4\}$. Nous avons utilisé le modèle de base suivant :

$$y_i^q = \alpha + \left(\beta^{q=1} + d\beta^{q=2}\delta_2^q + d\beta^{q=3}\delta_3^q + d\beta^{q=4}\delta_4^q\right)x_i^q + \varepsilon_i^q \qquad (1)$$

Dans cette équation, ε_i^q est un terme d'écart. Nous avons supposé que ε_i^q présente une distribution normale avec une moyenne de 0 et que sa variance varie en fonction des poids ω_i^q des unités, selon $\tilde{\sigma}^2/\omega_i^q$. Ces poids tiennent compte de l'hétéroscédasticité des données.

Nous avons élargi le modèle (1) afin de tenir compte de la présence de valeurs aberrantes dans les données. À cette fin, nous avons utilisé un modèle par mélange fini comparable à celui de Di Zio et Guarnera (2013). Nous avons supposé que les données ont été générées à partir d'un mélange de deux ensembles d'unités : un ensemble présentant une faible variance de l'erreur et un autre ensemble présentant une plus grande variance de l'erreur. Nous appliquerons ce modèle à d'autres groupes d'unités à la section 4. Ce modèle par mélange à deux groupes (M2) était donné par :

$$y_i^q = \alpha + \left(\beta^{q=1} + d\beta^{q=2}\delta_2^q + d\beta^{q=3}\delta_3^q + d\beta^{q=4}\delta_4^q\right)x_i^q + \varepsilon_i^q + z_i e_i^q \qquad (2)$$

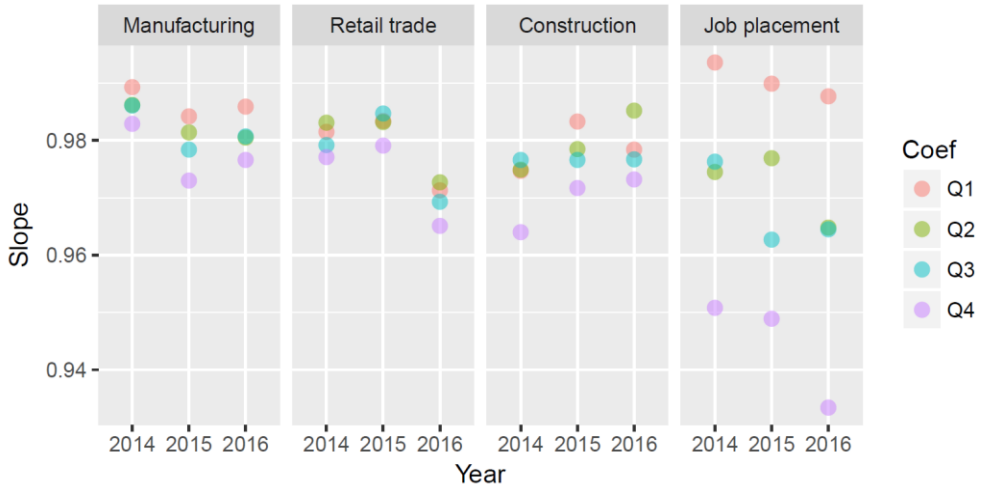
où $z_i \in \{0,1\}$ est un indicateur non observé avec $P(z_i = 1) = \pi$, et e_i^q est un autre terme d'écart présentant une distribution normale avec une moyenne de 0 et une variance de $(\vartheta - 1)\tilde{\sigma}^2/\omega_i^q$ qui touche seulement les unités avec $z_i = 1$. L'espérance conditionnelle de z_i compte tenu des données observées pour l'unité i est désignée par τ_i. On

peut l'interpréter comme une probabilité d'appartenir à un groupe. On suppose que ε_i^q, z_i et e_i^q sont mutuellement indépendants. Dans ce modèle, la variance du terme d'écart pour une unité donnée est augmentée par un facteur ϑ lorsque $z_i = 1$. Il convient de souligner que nous avons supposé que les unités sont affectées au même groupe pour une année complète.

Nous avons utilisé l'ensemble des unités sélectionnées, tel que décrit ci-dessus, pour estimer la formule (2). Nous avons utilisé un estimateur pour la formule (2) qui comporte un poids d'étalonnage, ce poids étant défini comme le rapport de la taille de la population à la taille de l'ensemble des unités sélectionnées par strate d'échantillonnage. Une strate d'échantillonnage est donnée par la combinaison d'une industrie et d'une catégorie de taille d'entreprise à un chiffre. Les paramètres du modèle ont été estimés au moyen d'un algorithme de maximisation de l'espérance conditionnelle (ECM) semblable à celui de Di Zio et Guarnera (2013). On peut obtenir des précisions dans Van Delden et Scholtus (2017).

3.2 Résultats

Figure 3.2-1
Pentes estimées d'après les unités X de niveau non supérieur qui répondent à l'enquête et déclarent des données sur la TVA. Les trimestres sont numérotés à compter du premier trimestre de 2014.



Pour les quatre secteurs économiques et les trois années, la pente estimée au quatrième trimestre était plus faible que celle du premier trimestre (voir la figure 3.2-1). La taille absolue de $d\beta^{q=4}$ était plus grande pour le placement en emploi (intervalle : -0,054 à -0,041), suivi de la fabrication (intervalle : -0,006 à -0,011), de la construction (intervalle : -0,005 à -0,008) et du commerce de détail (intervalle : -0,004 à -0,006). Pour chaque coefficient d'effet de la pente ($d\beta^{q=q^*}$ dans l'équation (2)), nous avons calculé la valeur p de l'hypothèse que sa valeur est de 0. Les effets étaient forts pour la fabrication (valeurs p pour toutes les années $< 0{,}01$) et plus faibles pour la construction et le commerce de détail, alors qu'une partie des valeurs p se situaient entre 0,05 et 0,10. Pour tous les secteurs économiques et toutes les années, la valeur p du coefficient d'effet de la pente $d\beta^{q=q^*}$ était le plus faible pour le quatrième trimestre de l'année.

4. Les différences dans la déclaration sont-elles attribuables à un ensemble limité d'unités?

Dans une analyse préliminaire, nous avons calculé la contribution de chacune des unités aux pentes estimées du modèle par mélange à deux groupes. Nous avons trié les unités selon la valeur absolue de cette contribution et avons observé qu'un ensemble limité d'unités ne pouvait à lui seul expliquer les différences dans la pente trimestrielle. Nous avons voulu mieux comprendre quelles unités contribuent aux différences saisonnières dans la déclaration en élargissant le modèle par mélange. Cet élargissement est décrit à la section suivante.

4.1 Méthodologie

Nous avons mis à l'essai un certain nombre d'élargissements du modèle à deux groupes. Dans ces élargissements, nous avons autorisé davantage de groupes d'unités, chaque groupe ayant sa propre pente trimestrielle ou annuelle et sa propre variance. Nous avons modélisé le tout en introduisant une variable nominale z_{gi} qui prend la valeur de 1 lorsque l'unité i appartient au groupe g et de 0 autrement. Le symbole τ_{gi} représente l'espérance de z_{gi} compte tenu des données observées pour l'unité i. De plus, nous avons comparé trois structures liées à la matrice de variance-covariance des quatre écarts trimestriels pour la même unité (Σ) : diagonale, par bandes et libre. Dans le cas d'une structure diagonale, tous les éléments diagonaux ont la même valeur positive et tous les autres éléments ont la valeur de zéro. Il convient de souligner qu'on a supposé cette structure de variance-covariance pour le modèle à deux groupes ci-dessus. Dans le cas d'une structure par bandes, tous les éléments de Σ sur une diagonale inférieure à la même distance de la diagonale principale ont une valeur commune. Dans le cas d'une structure libre, la seule restriction est que Σ est une matrice semi-définie symétrique positive.

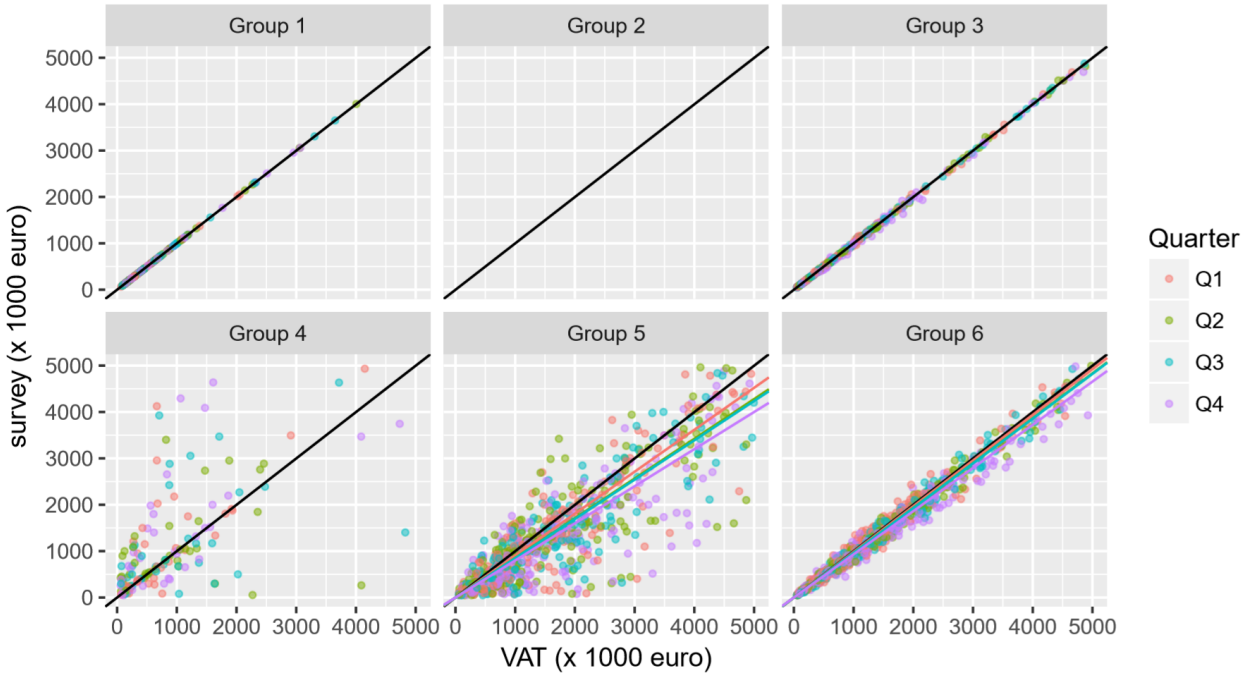
Un élargissement de l'algorithme ECM a été utilisé pour estimer ces modèles par mélange. Tous les modèles ont débuté par une série de valeurs initiales et la solution, et la meilleure valeur de vraisemblance a été sélectionnée. Pour comparer le rendement de différents modèles, nous avons calculé le critère d'information d'Akaike et le critère d'information bayésien (AIC et BIC), ainsi que ce que l'on appelle l'ICL-BIC (McLachlan et Peel, 2000). L'ICL-BIC ajoute un terme au BIC en fonction de l'entropie des probabilités d'affectation à un groupe τ_{gi} pour mesurer à quel point le modèle peut affecter chacune des unités à un groupe.

4.2 Résultats

Le modèle ayant le meilleur rendement dépendait du secteur économique et de l'année, mais dans l'ensemble, un modèle à six groupes avait le meilleur rendement. Ce modèle est désigné par M6. De plus, les structures par bandes et libre de la matrice de variance-covariance permettaient toujours d'obtenir de meilleurs résultats que la structure diagonale, et les différences dans les pentes estimées entre les structures par bandes et libre étaient généralement faibles.

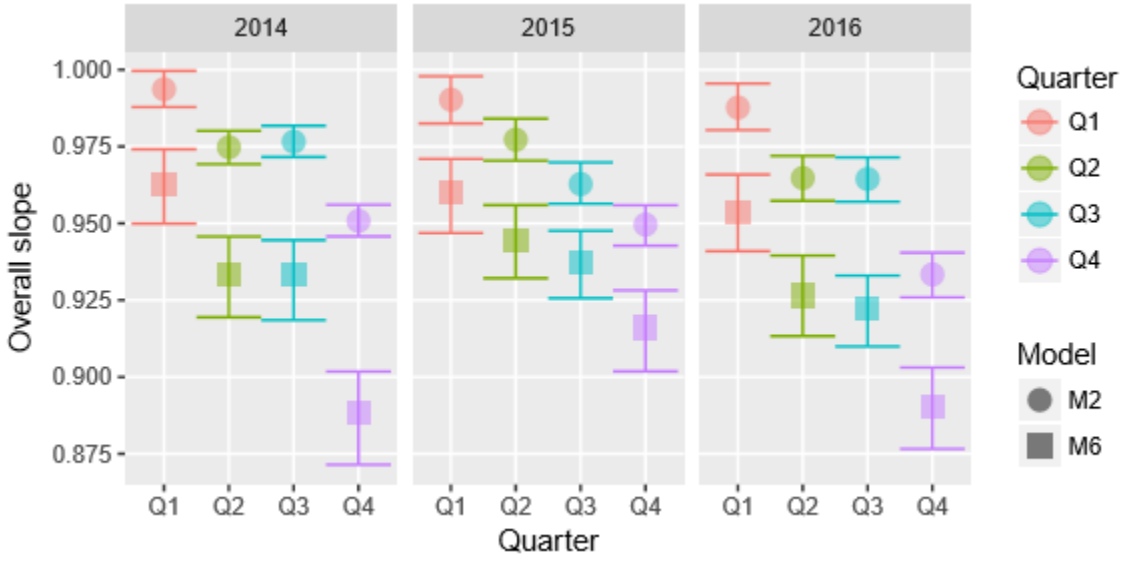
Pour le placement en emploi, le modèle M6 offrant le meilleur rendement avait une structure de variance-covariance libre. La relation entre les chiffres d'affaires d'après l'enquête trimestrielle et la TVA en 2016 pour les six groupes est représentée à la figure 4.2-1. Le premier groupe (4,8 % des unités) est formé des unités qui déclarent pratiquement les mêmes valeurs de chiffre d'affaires dans les deux sources. Le deuxième groupe (0,0 % des unités) est constitué des unités qui ont inclus par erreur un taux de TVA au chiffre d'affaires déclaré. Le troisième groupe (14,7 % des unités) se rapporte à un groupe qui présente une plus grande variance que le groupe 1 mais la même pente. Le groupe 4 (6,9 % des unités) représente des unités présentant des valeurs très aberrantes. Les groupes 5 (34,0 % des unités) et 6 (39,6 % des unités) représentent des unités présentant des effets saisonniers. Les effets trimestriels dans le groupe 5 sont plus marqués que dans le groupe 6, leurs pentes trimestrielles sont plus faibles et la variance est plus importante. La ligne noire dans la figure 4.2-1 indique la pente annuelle estimée commune des groupes 1 à 4, et les lignes colorées indiquent les pentes trimestrielles estimées des groupes 5 et 6.

Figure 4.2-1
Graphique de la relation entre les chiffres d'affaires d'après l'enquête trimestrielle et la TVA pour les six groupes dans le modèle M6 pour 2016



Pour le modèle M6, nous avons calculé les pentes trimestrielles moyennes pondérées en utilisant les probabilités d'appartenir à un groupe τ_{gi}; voir la figure 4.2-1. Nous avons observé que les pentes selon le modèle M6 étaient plus faibles que pour le modèle M2, mais les différences relatives entre les trimestres étaient semblables.

Figure 4.2-2
Pentes trimestrielles moyennes pondérées pour le modèle par mélange à deux groupes (M2) et le modèle par mélange à six groupes (M6). Les barres d'erreur du modèle indiquent les intervalles de confiance de 95 %, calculés à l'aide d'une procédure bootstrap



5. Conclusion et discussion

À l'aide d'un modèle par mélange à deux groupes simples, nous avons observé des effets saisonniers dans les quatre secteurs économiques et tout au long des trois années subséquentes. Cela suggère fortement qu'il existe effectivement des différences saisonnières systématiques dans la déclaration du chiffre d'affaire d'après l'enquête et la TVA. Les résultats du modèle par mélange élargi indiquaient qu'un grand groupe d'environ 75 % des unités peuvent contribuer à ces effets saisonniers. Selon le modèle M6, qui était bien ajusté aux données, ces différences trimestrielles dans la déclaration peuvent être attribuables à deux groupes d'unités différents dans la population : un groupe présentant des effets trimestriels plutôt importants, des pentes bien au-dessous de 1 et une grande variance, et un groupe d'unités présentant des effets trimestriels plus faibles, des pentes se rapprochant de 1 et une plus faible variance.

Pour la suite des choses, nous tenterons de comprendre les causes de ces tendances trimestrielles en lien avec le comportement de déclaration administrative qu'adoptent les entreprises. À cette fin, nous aimerions interroger les employés des bureaux administratifs et une sélection d'entreprises présentant des tendances particulières en matière de déclaration. À l'aide de ces renseignements, nous voulons connaître laquelle des deux séries et lesquelles des unités ont des tendances saisonnières déclarées qui présentent les moins d'erreurs de mesure. Nous utiliserons ces renseignements pour établir une approche permettant de corriger les effets saisonniers dans les données d'enquête ou les données sur la TVA, ou les deux, en vue de faciliter l'étalonnage dans l'avenir.

L'utilisation de modèles par mélange dans la détection des erreurs de mesure a déjà été proposée dans les statistiques officielles, par exemple par Di Zio et Guarnera (2013) et Guarnera et Varriale (2016). Nous avons étendu cette approche à la détection des effets de déclaration saisonniers dans deux sources. L'application aux données d'enquête et aux données sur la TVA peut également s'avérer pertinente pour d'autres pays qui utilisent, ou prévoient utiliser, la TVA comme source d'information sur le chiffre d'affaires. De façon plus générale, d'autres sources de données administratives infra-annuelles pourraient également subir ces effets et pourraient bénéficier de cette approche.

Remerciements

Nous remercions Jeroen Pannekoek pour son examen critique d'une version préliminaire du présent document. Les opinions exprimées dans le présent document sont celles des auteurs et ne témoignent pas nécessairement des politiques du Bureau central de la statistique des Pays-Bas.

Bibliographie

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

Delden, A. van, et S. Scholtus (2017), « Correspondence between survey and administrative data on quarterly turnover », *CBS Discussion paper 2017-3*.

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.

Guarnera, U., et R. Varriale (2016), « Estimation from Contaminated Multi-Source Data Based on Latent Class Models », *Statistical Journal of the IAOS*, 32, p. 537–544.

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

Scholtus, S., et P.P. de Wolf (2011), « Sampling design for the short-term statistics », rapport non publié, Den Haag, The Netherlands: Statistics Netherlands.