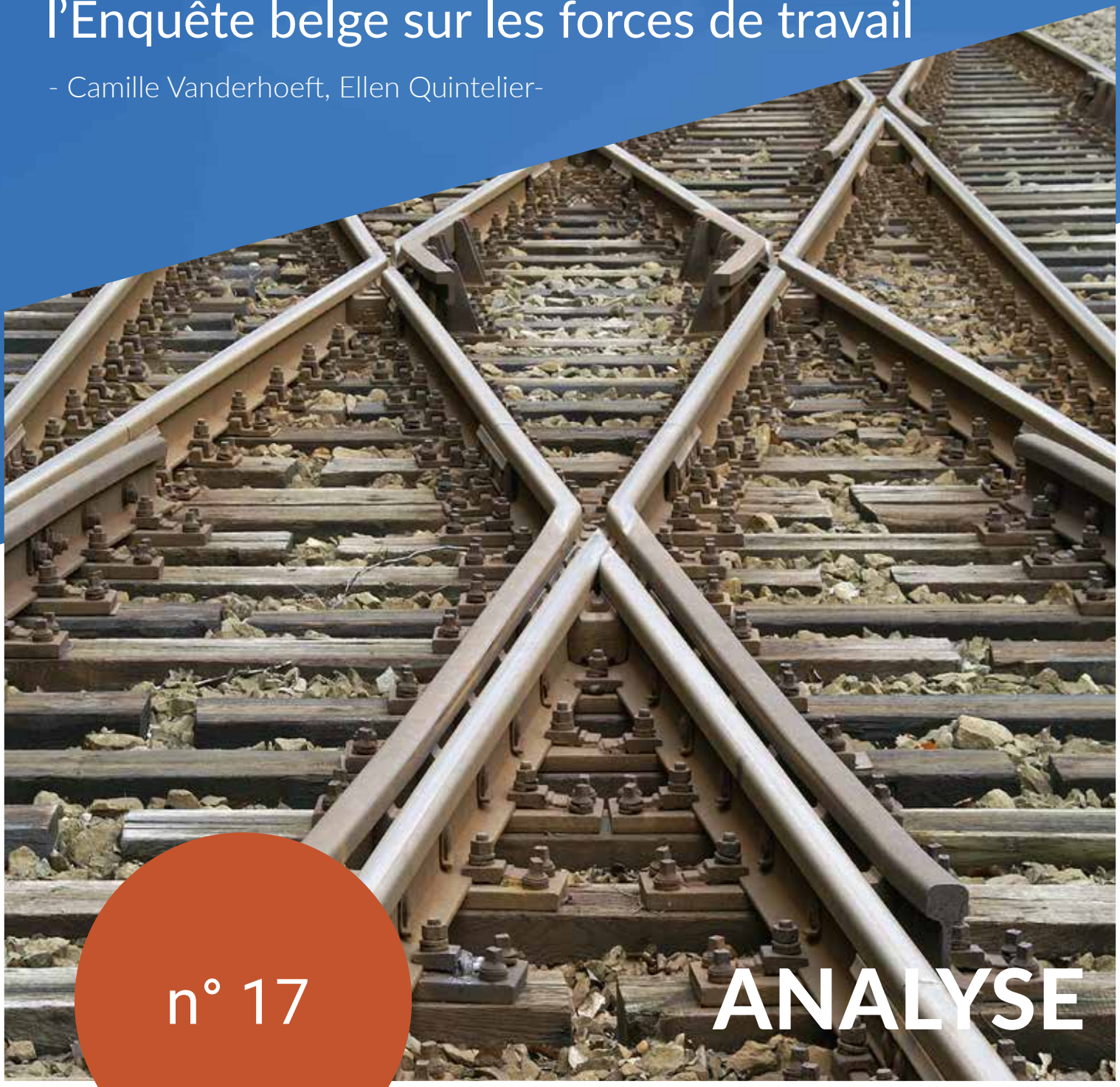
STATBEL
La Belgique en chiffres

L'estimation des transitions BIT sur le marché du travail sur la base des données du panel de l'Enquête belge sur les forces de travail

- Camille Vanderhoeft, Ellen Quintelier-



n° 17

ANALYSE

09.2021

economie

SPF Economie, P.M.E., Classes moyennes et Energie

.be

L'estimation des transitions BIT sur le marché du travail sur la base des données du panel de l'Enquête belge sur les forces de travail

Camille Vanderhoeft[1], Ellen Quintelier[2]

[1] Méthodologue à Statbel (Direction générale Statistique – Statistics Belgium)

[2] Statisticienne à Statbel (Direction générale Statistique – Statistics Belgium)

L'ESTIMATION DES TRANSITIONS BIT SUR LE MARCHE DU TRAVAIL SUR LA BASE DES DONNEES DU PANEL DE L'ENQUETE BELGE SUR LES FORCES DE TRAVAIL

Combien de personnes occupées travaillent encore après un trimestre et combien sont au chômage ou inactives ?

Quelle évolution est observée au niveau du statut sur le marché du travail après un an ?

Certains groupes de population restent-ils plus souvent au chômage ?

Quelle est la méthodologie utilisée par Statbel pour estimer les transitions ?

Camille Vanderhoeft, Ellen Quintelier

RESUME

Depuis 2017, l'Enquête sur les forces de travail en Belgique est une enquête par panel. Elle permet d'interroger les personnes concernant leur statut sur le marché du travail pendant une période de 18 mois, et de déterminer si les personnes occupées travaillent toujours trois mois ou un an plus tard, ou si elles sont alors au chômage ou inactives. Le même raisonnement est appliqué aux chômeurs et aux inactifs. Notre objectif consiste donc à chiffrer les neuf transitions possibles entre les trois statuts sur le marché du travail (chômeur, occupé et inactif). En d'autres termes, nous voulons estimer les matrices 3x3 de transition du marché du travail.

L'estimation de ces matrices de transition n'est cependant pas aussi simple qu'il n'y paraît : nous voulons en effet que les totaux des lignes et des colonnes des matrices de transition soient cohérents par rapport aux chiffres trimestriels ou annuels qui peuvent être calculés à partir des échantillons trimestriels ou annuels, et dont les résultats sont publiés comme indicateurs officiels. Nous expliquerons cette méthode en détail dans cette analyse.

Nous commencerons l'analyse en expliquant la structure des données et les transitions possibles au chapitre 1. Au chapitre 2, nous détaillerons la méthodologie, qui est entièrement basée sur le calage. Nous sommes partis de la méthode utilisée par Eurostat, l'office statistique européen, que nous avons développée un peu plus afin de fournir un large éventail de ventilations en fonction de diverses variables contextuelles, sans perdre (trop) de cohérence par rapport aux indicateurs officiels. Le résultat de ces développements méthodologiques est un modèle de calage entièrement intégré. Enfin, nous aborderons au chapitre 3 la publication des matrices de transition estimées et illustrerons les difficultés qui peuvent survenir lorsque de telles matrices doivent être estimées pour de plus petits groupes de population. L'exemple sur les chômeurs de courte et de longue durée au paragraphe 3.4 permet de montrer de quelle manière les transitions dans des sous-populations spécifiques peuvent être étudiées.

Dans les conclusions, nous résumons les principaux résultats de nos développements méthodologiques.

Cette analyse se concentre sur la méthodologie, et non sur l'interprétation, l'explication socio-économique ou l'utilisation des résultats. Une discussion concise des résultats est disponible en marge de chaque publication de nouvelles estimations sur le site web de Statbel, l'office belge de statistique, par exemple sous forme de communiqués de presse.

TABLE DES MATIÈRES

LISTE D'ABRÉVIATIONS

EFT	Enquête sur les forces de travail
BIT	Bureau international du Travail
AGE	Classe d'âge
BQ	Trimestre de départ, trimestre initial
EDU	Niveau d'enseignement
EQ	Trimestre de fin
IPF	Ajustage de précision proportionnel itératif
LS	Echantillon longitudinal
NAT	Classe de nationalité
NC	Cohérence numérique
NC-C	Cohérence numérique selon la méthode classique
NC-E	Cohérence numérique selon la méthode d'Eurostat
REG	Région de résidence
RG	Groupe de rotation
SEX	Sexe, genre
W	Wave (vague)
yyyyTt	Trimestre t en année yyyy

LISTE DES SCHÉMAS ET TABLEAUX

1 Introduction

Depuis l'introduction d'un panel pour l'Enquête belge sur les forces de travail (EFT), nous pouvons non seulement publier des chiffres annuels et trimestriels, mais aussi estimer des transitions dans le statut BIT sur le marché du travail, c.-à-d. « occupé", « chômeur" ou « inactif" – (en abrégé : le statut BIT[3]) d'un trimestre ou d'une année à l'autre. Il est ainsi possible d'analyser comment le statut BIT des répondants évolue au cours d'un certain nombre de trimestres. Le statut BIT est de loin la variable la plus importante de l'Enquête sur les forces de travail. Cette variable indique si une personne travaille, est au chômage ou est inactive durant la semaine de référence, conformément aux définitions en vigueur au niveau international[4]. Les matrices de transition, qui comprennent les nombres (estimés) de personnes passant d'un statut BIT à l'autre, peuvent donner une indication de la mesure dans laquelle les chômeurs trouvent du travail, les personnes occupées continuent de travailler, etc. Après trois ans de collecte de données dans un panel, une série limitée de données est disponible et permet de montrer les transitions et les tendances au niveau des transitions.

En 2017, Statbel, l'office belge de statistique, a introduit un panel rotatif pour l'EFT : chaque trimestre, un nouvel échantillon ou groupe de rotation (RG) est tiré et utilisé sur le terrain. Le panel est rotatif en ce sens qu'environ un quart de l'*échantillon trimestriel* est remplacé par une nouvelle sélection chaque trimestre. Les répondants d'un même RG sont interrogés pendant deux trimestres consécutifs, puis ne le sont plus pendant deux trimestres et le sont à nouveau pendant deux trimestres (Termote & Depickere, 2018). Nous parlons ici d'un scénario 2(2)2. L'interrogation de chaque RG est donc répartie sur six trimestres. Lors de la première participation d'un répondant, nous parlons de la première vague ou wave (W1) ; la deuxième fois, de deuxième vague (W2), etc. Le schéma 1 de Termote & Depickere (2018) l'illustre clairement. Les résultats trimestriels, basés sur l'échantillon trimestriel, ont toujours été basés sur les données de quatre vagues (pour quatre RG différents) depuis le premier trimestre ; voir schéma 2 de Termote & Depickere (2018). Les résultats trimestriels de 2017 sont basés sur des échantillons trimestriels qui, dans une certaine mesure, sont composés différemment ; voir schéma 4 de Termote & Depickere (2018).

Le schéma 4 de Termote & Depickere (2018) indique en outre clairement qu'un panel a déjà été mis en place au troisième trimestre 2016. Les cinq premiers RG ont cependant été interrogés selon des scénarios qui diffèrent du scénario 2(2)2. La période allant du troisième trimestre 2016 au quatrième trimestre 2017 est une phase transitoire, qui était nécessaire pour permettre la transition de l'enquête continue à une enquête par panel, en tenant compte de diverses exigences.

Sur la base de ces données de panel, nous allons calculer trois types de transitions : les *transitions trimestrielles* (transitions entre trimestres consécutifs), les *transitions annuelles par trimestre* (transitions entre les mêmes trimestres de deux années consécutives, également appelées *transitions annuelles d'un trimestre spécifique*) et les *transitions annuelles* (transitions entre années consécutives). Ces transitions sont expliquées plus en détail ci-dessous.

1.1 Transitions trimestrielles : transitions entre trimestres successifs

Les transitions trimestrielles sont basées sur le chevauchement d'échantillons trimestriels de deux trimestres consécutifs – nous parlons de *trimestre de départ* (BQ) et de *trimestre de fin* (EQ) –, par exemple, le troisième et le quatrième trimestre de 2019 (2019T3 et 2019T4, respectivement), ou le quatrième trimestre de 2019 et le premier trimestre de 2020 (2019T4 et 2020T1, respectivement). Ce chevauchement, que nous appellerons *échantillon longitudinal* (LS) pour l'estimation des transitions trimestrielles, satisfait à une exigence d'Eurostat concernant le panel pour l'EFT[5] : le chevauchement théorique des échantillons doit être d'au moins 50 % entre deux échantillons trimestriels consécutifs. Le panel belge répond à cette exigence : le chevauchement des échantillons trimestriels pour 2019T3 et 2019T4, par exemple, se compose de RG13, avec des observations pour chaque répondant en W3 et W4, et de RG17, avec des observations pour chaque répondant en W1 et W2. Deux des quatre RG (soit 50 % des échantillons initiaux) dans chacun des deux échantillons trimestriels donnent ainsi lieu à des observations dans deux trimestres consécutifs.

[3]Avec cette terminologie, nous indiquons que nous utilisons les définitions du Bureau international du Travail (BIT) : les concepts tels que « occupé », « chômeur » et « inactif » dans cette analyse doivent toujours être interprétés conformément aux définitions du BIT.

[4] Voir https://statbel.fgov.be/fr/themes/emploi-formation/marche-du-travail/faq.

[5] Règlement 2019/1700 établissant un cadre commun pour des statistiques européennes relatives aux personnes et aux ménages (https://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX:32019R1700).

Notez que chaque LS pour l'estimation des transitions trimestrielles contient des répondants d'un RG bien défini avec des observations en W1 et W2, et des répondants d'un autre RG bien défini avec des observations en W3 et W4. Ceci est une conséquence immédiate de l'application du scénario 2(2)2.

Les résultats relatifs aux transitions trimestrielles ont été publiés par Eurostat sur la base de sa méthodologie jusqu'à la fin de 2020. Statbel a maintenant développé sa propre méthodologie, inspirée de celle d'Eurostat, pour calculer lui-même les transitions. À partir de 2021, Statbel produit et publie lui-même les matrices de transition, et les anciennes estimations produites et publiées par Eurostat sont remplacées par les estimations de Statbel.

1.2 Transitions annuelles par trimestre : transitions entre les mêmes trimestres de deux années consécutives

Les transitions annuelles par trimestre, ou *transitions annuelles d'un trimestre spécifique*, sont basées sur le chevauchement d'échantillons trimestriels pour le même trimestre de deux années consécutives, par exemple 2018T2 et 2019T2, ou 2019T3 et 2020T3 ; ici aussi, nous parlons de BQ et d'EQ. Ce chevauchement, que nous appellerons *échantillon longitudinal* (LS) pour l'estimation des transitions annuelles par trimestre, satisfait également à une exigence d'Eurostat concernant le panel pour l'EFT : le chevauchement théorique des échantillons doit être d'au moins 20 % entre les échantillons trimestriels d'un même trimestre de deux années consécutives. Le panel belge répond à cette exigence : le chevauchement des échantillons trimestriels pour 2018T2 et 2019T2, par exemple, se compose de RG11, avec des observations pour chaque répondant en W2 et W4, et de RG12, avec des observations pour chaque répondant en W1 et W3 ; deux des quatre RG (soit 50 % des échantillons initiaux) dans chacun des deux échantillons trimestriels donnent ainsi lieu à des observations au BQ et au EQ. Avec ce chevauchement « théorique" de 50 %, le panel belge dépasse donc largement l'exigence des 20 %.

Notez que chaque LS pour l'estimation des transitions annuelles d'un trimestre spécifique contient des répondants d'un RG bien défini avec des observations en W1 et W3, et des répondants d'un autre RG bien défini avec des observations en W2 et W4. Ceci est également une conséquence immédiate de l'application du scénario 2(2)2.

Les résultats sur les transitions annuelles d'un trimestre spécifique ne sont pas publiés par Eurostat car pour certains pays, l'échantillon longitudinal (LS) est trop petit, étant donné le chevauchement théorique demandé de 20 % seulement. Puisqu'un chevauchement théorique de 50 % est atteint dans l'EFT belge, Statbel pourra produire et publier une partie des résultats pour ces transitions annuelles.

1.3 Transitions annuelles : transitions entre années consécutives

Si nous prenons la moyenne de quatre transitions annuelles d'un trimestre spécifique, nous obtenons les *transitions annuelles* (globales). Les résultats sur les transitions annuelles sont publiés par Eurostat sur la base de sa propre méthodologie. Entre-temps, Statbel produit et publie désormais également des chiffres en utilisant la méthodologie décrite dans cette analyse.

L'EFT est sujette à la non-réponse et, en tant qu'enquête par panel, à l'attrition (en dépit du caractère obligatoire de l'enquête). En moyenne, 74,4 % des personnes sélectionnées répondent positivement à la première interrogation (chiffres pour les enquêtes organisées en 2019). Parmi les répondants de la première vague, 87,2 % participent encore à la deuxième enquête, 90,1 % d'entre eux à la troisième et 93,8 % d'entre eux à la quatrième. En moyenne, il reste donc encore 54,8 % de l'échantillon initial dans la quatrième vague. Toutes les personnes qui ne répondent pas l'enquête ne refusent pas forcément d'y participer : dans certains cas, l'adresse est erronée, les intéressés ont déménagé, les enquêteurs ne parviennent pas à contacter le ménage, etc. Le chevauchement entre les échantillons trimestriels est en pratique plus faible que les 50 % théoriques discutés ci-dessus. En outre, le chevauchement est légèrement plus faible pour les transitions annuelles que pour les transitions trimestrielles[6].

1.4 Matrices de transition

Les transitions trimestrielles entre BQ et EQ peuvent être présentées sous la forme d'une matrice de transition, c.-à-d. un tableau, comme dans le Schéma 1 ci-dessous, avec sur la diagonale (dans les cellules grises) le nombre d'individus qui ne

[6]Par exemple, la transition annuelle 2019T1-2020T1 compte 12.667 observations non pondérées ; la transition trimestrielle 2019T4-2020T1 compte 14.507 observations non pondérées.

changent pas de statut BIT entre BQ et EQ[7] (par exemple, ceux qui sont inactifs durant la semaine de référence du BQ et durant la semaine de référence du EQ) et dans les autres cellules, le nombre d'individus qui changent de statut BIT (par exemple, d'occupé durant la semaine de référence du BQ à inactif durant la semaine de référence du EQ). La matrice est complétée par des marges : (1) les totaux des lignes de la dernière colonne, représentant la distribution du statut BIT au BQ ; (2) les totaux des colonnes de la dernière ligne, représentant la distribution du statut BIT au EQ ; (3) le nombre total de personnes. Les nombres figurant dans une matrice de transition sont des nombres estimés d'individus dans une (sous-)population bien définie étudiée : au Schéma 1, nous utilisons la notation $\widehat{N}$ (c.-à-d. $\widehat{N}_{11}$, ...) pour les estimations des nombres réels de la (sous-)population N (c.-à-d. N_{11}, ...).

Schéma 1 : Présentation générale d'une matrice de transition estimée, avec marges

		Statut BIT au trimestre de fin (EQ)			Total au trimestre de départ (BQ)
		Chômeur	Occupé	Inactif	
Statut BIT au trimestre de départ (BQ)	Chômeur	$\widehat{N}_{11}$	$\widehat{N}_{12}$	$\widehat{N}_{13}$	$\widehat{N}_{1+}$
	Occupé	$\widehat{N}_{21}$	$\widehat{N}_{22}$	$\widehat{N}_{23}$	$\widehat{N}_{2+}$
	Inactif	$\widehat{N}_{31}$	$\widehat{N}_{32}$	$\widehat{N}_{33}$	$\widehat{N}_{3+}$
Total au trimestre de fin (EQ)		$\widehat{N}_{+1}$	$\widehat{N}_{+2}$	$\widehat{N}_{+3}$	$\widehat{N}_{++}$

1.5 Matrices de transition relative ou matrices des taux de transition

Les transitions relatives peuvent être obtenues en divisant, par ligne, les nombres dans les cellules par les totaux de ligne correspondants, et en multipliant par 100, c.-à-d. $\hat{p}_{j|i} = 100 \times \widehat{N}_{ij}/\widehat{N}_{i+}$; le pourcentage ainsi obtenu est une estimation du pourcentage $p_{j|i} = 100 \times N_{ij}/N_{i+}$ des individus dans le statut i (par ex. chômeur) au BQ, qui se retrouvent au EQ dans le statut j (par ex. occupé). La dernière ligne contient la distribution estimée en pourcentage du statut BIT au EQ : $\hat{p}_{+j} = 100 \times \widehat{N}_{+j}/\widehat{N}_{++}$. Une matrice de transition relative, dorénavant appelée *matrice des taux de transition*, est représentée schématiquement au Schéma 2.

Schéma 2: Présentation générale d'une matrice des taux de transition, avec marges

		Statut BIT au trimestre de fin (EQ)			Total au trimestre de départ (BQ)
		Chômeur	Occupé	Inactif	
Statut BIT au trimestre de départ (BQ)	Chômeur	$\hat{p}_{1\|1}$	$\hat{p}_{1\|2}$	$\hat{p}_{1\|3}$	100 %
	Occupé	$\hat{p}_{2\|1}$	$\hat{p}_{2\|2}$	$\hat{p}_{2\|3}$	100 %
	Inactif	$\hat{p}_{3\|1}$	$\hat{p}_{3\|2}$	$\hat{p}_{3\|3}$	100 %
Total au trimestre de fin (EQ)		$\hat{p}_{+1}$	$\hat{p}_{+2}$	$\hat{p}_{+3}$	100 %

Dans les fichiers Excel sur le site web de Statbel (voir chapitre 3), la matrice des taux de transition correspondante sera également publiée pour chaque matrice de transition.

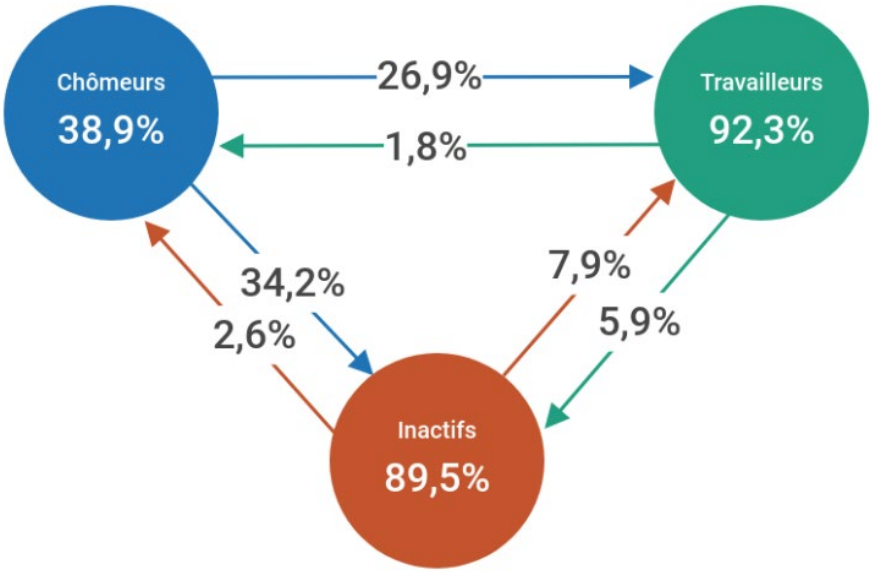
Une matrice de transition telle que présentée au Schéma 1 peut être transformée non seulement en une matrice des taux de transition telle que présentée au Schéma 2, en calculant les pourcentages de lignes $\hat{p}_{j|i}$, mais aussi en une matrice alternative des taux de transition en calculant les pourcentages de colonnes, c.-à-d. $\hat{q}_{i|j} = 100 \times \widehat{N}_{ij}/\widehat{N}_{+j}$. Ces pourcentages de colonnes $\hat{q}_{i|j}$ indiquent quel pourcentage d'individus dans le statut j au EQ se trouvait dans le statut i au BQ, tandis que

[7] Pour certains répondants, le statut BIT peut changer plusieurs fois entre l'observation au BQ et l'observation au EQ. Cependant, seuls les statuts des semaines de référence (au BQ et au EQ) sont enregistrés ; les statuts intermédiaires ne sont pas notés. Ainsi, il est possible qu'un changement de statut d'un répondant passe inaperçu, mais que le statut enregistré au BQ et au EQ soit le même ; un tel répondant contribue aux nombres sur la diagonale de la matrice de transition.

les pourcentages de lignes $\hat{p}_{j|i}$ indiquent quel pourcentage d'individus dans le statut i au BQ se trouvera dans le statut j au EQ. L'interprétation des résultats dans une matrice de transition peut conduire à l'un ou aux deux types de pourcentages (équivalents), selon l'approche de l'utilisateur. Pour l'explication méthodologique dans cette analyse, il importe peu quelles transitions relatives sont utilisées ; dans ce texte, le terme « matrice de taux de transition" fait toujours référence aux pourcentages de lignes $\hat{p}_{j|i}$, et nous ne présentons que des pourcentages de lignes dans les tableaux.

Au par. 1.3, il est indiqué qu'une matrice de transition annuelle est obtenue en tant que moyenne de quatre matrices de transition annuelle d'un trimestre spécifique (ce point est discuté plus en détail d'un point de vue méthodologique au par. 2.9). Chacune des quatre matrices de transition annuelle d'un trimestre spécifique est accompagnée d'une matrice des taux de transition annuelle d'un trimestre spécifique. Il est important de noter que la matrice des taux de transition annuelle associée à la matrice de transition annuelle ne peut pas être calculée comme une moyenne des quatre matrices de taux de transition annuelle d'un trimestre spécifique, mais plutôt directement à partir de la matrice de transition annuelle.

Les taux de transition peuvent être visualisés dans un diagramme, par exemple pour les transitions annuelles entre 2019 et 2020 :



Ce diagramme montre que, selon les estimations :

- parmi les chômeurs en 2019, 38,9 % sont toujours au chômage en 2020, 26,9 % sont occupés en 2020 et 34,2 % sont inactifs en 2020 ;
- parmi les personnes occupées en 2019, 92,3 % sont toujours occupées en 2020, 1,8 % sont au chômage en 2020 et 5,9 % sont inactives en 2020 ;
- parmi les inactifs en 2019, 89,5 % sont toujours inactifs en 2020, 2,6 % sont au chômage en 2020 et 7,9 % sont occupés en 2020.

1.6 Matrices de transition non pondérées ou matrices de la taille de l'échantillon

Enfin, chaque matrice de transition estimée publiée sera également accompagnée d'une « matrice de transition non pondérée" associée, qui n'est rien d'autre que la matrice reprenant les nombres de répondants dans le LS sur laquelle est basée la matrice de transition : nous parlons donc également d'une *matrice de la taille de l'échantillon*. Le Schéma 3 présente de manière schématique une telle matrice de la taille de l'échantillon.

Schéma 3 : Présentation générale d'une matrice de la taille de l'échantillon, avec marges

		Statut BIT au trimestre de fin (EQ)			Total au trimestre de départ (BQ)
		Chômeur	**Occupé**	**Inactif**	
Statut BIT au trimestre de départ (BQ)	**Chômeur**	n_{11}	n_{12}	n_{13}	n_{1+}
	Occupé	n_{21}	n_{22}	n_{23}	n_{2+}
	Inactif	n_{31}	n_{32}	n_{33}	n_{3+}
Total au trimestre de fin (EQ)		n_{+1}	n_{+2}	n_{+3}	n_{++}

Une matrice de la taille de l'échantillon représente la distribution non pondérée du LS en fonction du statut BIT au BQ et au EQ. La matrice de transition représente la distribution pondérée du LS selon le statut BIT au BQ et au EQ : chaque estimation $\widehat{N}_{ij}$ dans la matrice de transition est la somme des poids de calage longitudinaux pour les n_{ij} répondants correspondants dans le LS. La matrice de la taille de l'échantillon fournit des indications sur la fiabilité de la matrice (de taux) de transition : plus le nombre n_{ij} est élevé, plus les estimations $\widehat{N}_{ij}$ et $\hat{p}_{i|j}$ sont précises.

Dans les fichiers Excel sur le site web de Statbel (voir chapitre 3), la matrice de la taille de l'échantillon correspondante sera également publiée pour chaque matrice de transition.

Notez que la matrice de la taille de l'échantillon associée à une matrice de transition annuelle est simplement la somme des matrices de la taille de l'échantillon associées aux quatre matrices de transition annuelle d'un trimestre spécifique.

Le calcul des poids de calage longitudinaux est le sujet central de cette analyse : voir chapitre 2.

2 Méthodologie

La méthodologie développée par Statbel sera illustrée dans ce chapitre à l'aide de la paire de trimestres consécutifs 2018T3 et 2018T4, pour lesquels les *transitions trimestrielles* seront estimées. La méthodologie est exactement la même pour l'estimation des *transitions annuelles d'un trimestre spécifique*, basée sur des paires de mêmes trimestres d'années consécutives.

Notre analyse méthodologique diffère de celle d'Eurostat (Eurostat, 2015a et Eurostat, 2015b), étant donné l'utilisation rigoureuse d'une méthodologie de calage générale, telle que développée par Deville et Särndal (1992). Elle inclut un logiciel général approprié : Statbel utilise le SAS®-macro CALMAR2 (Le Guennec et Sautory, 2002 et Sautory, 1993). Cela permet d'étendre systématiquement des modèles plus simples (comme ceux introduits par Eurostat) afin d'élargir les objectifs sans rendre les calculs plus difficiles à réaliser.

La méthodologie utilisée par Statbel pour estimer les transitions trimestrielles et les transitions annuelles d'un trimestre spécifique est abordée en détail aux paragraphes 2.1 à 2.7. Au par. 2.8 nous comparons cette méthodologie avec celle introduite par Eurostat (2015b).

Au par. 2.9, nous abordons l'estimation des *transitions annuelles (globales)*, en tant que moyennes simples des transitions annuelles d'un trimestre spécifique.

2.1 Calage et variables de calage

Le calage est le processus de correction de poids initiaux pour les unités d'un échantillon, de sorte que pour certaines variables, la distribution finale pondérée de l'échantillon est identique à une distribution de référence (éventuellement estimée). L'échantillon en question dans cette analyse est l'*échantillon longitudinal* (voir par. 2.2.3), et les unités de cet échantillon sont des répondants individuels. Nous appelons les variables en question les *variables de calage*. Les poids finaux corrigés sont appelés *poids de calage longitudinaux*. Certains principes et propriétés (pratiques) des techniques de calage sont expliqués en annexe B.

Une variable de calage est donc une variable pour laquelle la distribution dans l'échantillon redressé devra être la même qu'une distribution de référence (éventuellement estimée). Les variables de calage potentielles abordées dans cette analyse sont :

- STAT1, le statut BIT au BQ, et STAT2, le statut BIT au EQ, avec les valeurs (classes) :
  - *Manquant* pour les répondants de la classe d'âge 0-14
  - *Chômeur, Occupé* ou *Inactif* pour les répondants de la classe d'âge 15+ ;
- SEX, le genre (sexe) – qui est supposé ne pas changer entre BQ et EQ – avec les valeurs *Homme* et *Femme* ;
- AGE1, la classe d'âge au BQ, et AGE2, la classe d'âge au EQ, avec valeurs (classes) 0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74 et 75+ ; si nécessaire, nous travaillons avec un autre regroupement des classes d'âge, par ex. AGE1 et AGE2 avec les classes 0-14, 15-34, 35-54, 55-74 et 75+, ou AGE1 et AGE2 avec les classes 0-14, 15-29, 30-74 et 75+ ;
- REG1, la région de résidence au BQ, et REG2, la région de résidence au EQ, avec les valeurs *BRU* (Région de Bruxelles-Capitale), *VLA* (Région flamande) et *WAL* (Région wallonne) ;
- NAT1, la classe de nationalité au BQ, et NAT2, la classe de nationalité au EQ, avec les valeurs (classes) *BE* (belge), *EU* (nationalité EU28, sans la Belgique) et *non EU* (nationalité non EU28) ; si nécessaire, nous travaillons avec un regroupement supplémentaire, par exemple NAT1 et NAT2 avec les classes *BE* et *non BE* ;
- EDU1, le plus haut niveau d'instruction atteint au BQ, et EDU2, le plus haut niveau d'instruction atteint au EQ, avec les valeurs (classes) *Faible* (aucun diplôme, ou au plus un diplôme de l'enseignement secondaire inférieur), *Moyen* (diplôme de l'enseignement secondaire supérieur) et *Élevé* (au moins un diplôme de l'enseignement supérieur).

Les répondants d'un échantillon à redresser sont exclus si une ou plusieurs de ces variables ont une valeur indéterminée (*manquante*). Notez que les variables de calage peuvent être décomposées en variables contextuelles (sexe, âge, région de résidence, nationalité et niveau d'instruction) et en variables d'étude (statut BIT).

Changer la valeur d'une variable de calage entre BQ et EQ pour certains répondants ne pose aucun problème. Pour le sexe (SEX), nous n'avons pas trouvé de tels répondants dans la pratique, mais cette possibilité existe aussi pour cette variable ; dans ce cas, nous considérerions les variables SEX1 et SEX2. La classe d'âge (AGE) change inévitablement pour un nombre

important de répondants entre BQ et EQ. La région de résidence (REG), la nationalité (NAT) et le plus haut niveau d'instruction (EDU) changent dans une moindre mesure entre BQ et EQ. Le changement de statut BIT entre BQ et EQ est le sujet de cette étude ; les variables STAT1 et STAT2 jouent un rôle particulier en tant que variables de calage.

2.2 Les échantillons

Trois échantillons (de répondants[8]) entrent en ligne de compte pour l'estimation des transitions entre n'importe quelle paire de trimestres, c.-à-d. pour l'estimation tant de transitions trimestrielles que de transitions annuelles d'un trimestre spécifique. Nous discutons de ces trois échantillons dans les trois paragraphes qui suivent.

2.2.1 L'échantillon trimestriel de départ

Il s'agit de l'échantillon des répondants au BQ. Cet échantillon contient des répondants de quatre RG. Pour l'estimation des statistiques trimestrielles (pour les variables de base), l'échantillon BQ a été redressé sur les distributions de population de diverses variables contextuelles. On obtient ainsi un poids de calage w_i^{BQ} pour chaque répondant i de l'échantillon BQ.

Pour 2018T3, l'échantillon BQ comprend les répondants des RG 8, 9, 12 et 13 ; voir schéma 1 de Termote & Depickere (2018). Dans cette analyse, cet échantillon est limité aux répondants de la classe d'âge 15-74 ans.

L'échantillon BQ fournit des distributions estimées de la population (pour la classe d'âge 15-74 ans) qui seront utilisées comme distributions de référence dans le calage de l'échantillon longitudinal (LS) ; ceci est discuté plus en détail au par. 2.3.

2.2.2 L'échantillon trimestriel final

Il s'agit de l'échantillon des répondants au EQ. Cet échantillon contient des répondants de quatre RG. Pour l'estimation des statistiques trimestrielles (pour les variables de base), l'échantillon EQ a été redressé sur les distributions de population de diverses variables contextuelles. On obtient ainsi un poids de calage w_i^{EQ} pour chaque répondant i de l'échantillon EQ.

Pour 2018T4, l'échantillon EQ comprend les répondants des RG 9, 10, 13 et 14 ; voir schéma 1 de Termote & Depickere (2018). Dans cette analyse, cet échantillon est limité aux répondants de la classe d'âge 15-74 ans.

L'échantillon EQ fournit des distributions estimées de la population (pour la classe d'âge 15-74 ans) qui seront utilisées comme distributions de référence dans le calage de l'échantillon longitudinal (LS) ; ceci est discuté plus en détail au par. 2.3.

2.2.3 L'échantillon longitudinal (LS)

L'*échantillon longitudinal* (LS) est le chevauchement (ou intersection) des échantillons BQ et EQ, c.-à-d. la collection de tous les répondants qui ont été observés à la fois au BQ et au EQ. Dans les fichiers de données contenant les échantillons BQ et EQ, les répondants ont été identifiés par un numéro de répondant unique, qui ne change pas d'une vague à l'autre. Le fichier de données contenant le LS peut donc être facilement construit en reliant les fichiers de données contenant les échantillons BQ et EQ par ce numéro de répondant. Pour la paire 2018T3-2018T4, ce chevauchement est limité aux répondants de deux RG, à savoir RG9 et RG13 ; pour RG9, nous disposons des observations des vagues 3 et 4, pour RG13, des vagues 1 et 2 (voir schéma 1 dans Termote & Depickere (2018)). Schéma :

Trimestre de départ	Trimestre de fin	RG dans le chevauchement	1er RG	2e RG
			Observations des vagues...	
2018T3	2018T4	9 et 13	3 et 4	1 et 2

Le Tableau B 1 en annexe A montre toutes les paires possibles de trimestres consécutifs depuis le lancement du panel au 2016T3, avec les RG du LS, et pour chacun de ces RG, les vagues consécutives dans lesquelles les répondants du LS ont été interrogés. Le tableau a été dérivé des schémas 1 et 4 de Termote & Depickere (2018). Le Tableau B 1 montre une composition standard des LS à partir de la paire 2017T4-2018T1 : deux RG, l'un contenant les répondants avec des observations dans les vagues 3 et 4, et l'autre contenant les répondants avec des observations dans les vagues 1 et 2. Pour les paires précédentes, c.-à-d. 2016T3-2016T4 à 2017T3-2017T4, à l'exception de la paire 2017T1-2017T2, la composition du LS diffère de la composition standard : parfois trois RG sont considérés et/ou les observations des répondants peuvent également être

[8] À partir de ce point du texte, tous les échantillons sont des échantillons de répondants ; nous ne le répéterons pas explicitement à chaque fois.

disponibles dans les vagues 2 et 3 ; ceci découle des scénarios différents pour les RG tirés lors de la phase de démarrage du panel EFT (voir schéma 4 dans Termote & Depickere (2018)). Ces compositions divergentes pourraient potentiellement avoir un impact sur les transitions trimestrielles estimées ; voir Termote & Depickere (2018).

Le Tableau B 2 en annexe A montre la composition des LS pour les paires de mêmes trimestres dans des années consécutives ; le tableau a été dérivé des schémas 1 et 4 de Termote & Depickere (2018). Dans ce tableau également, nous constatons une composition standard à partir de 2017T1-2018T1, et des compositions divergentes pour 2016T3-2017T3 et 2016T4-2017T4, qui sont dues à des scénarios différents pour les RG tirés dans la phase de démarrage du panel EFT (voir schéma 4 dans Termote & Depickere (2018)).

Dans cette analyse, le LS est toujours limité aux répondants qui ont entre 15 et 74 ans (c.-à-d. qui n'ont pas encore atteint l'âge de 75 ans) dans les deux échantillons trimestriels. Cela permet d'éviter un quatrième statut BIT, à savoir *manquant* pour les moins de 14 ans, dans les matrices de transition.

Pour les répondants i dans un LS, les deux poids de calage w_i^{BQ} et w_i^{EQ} sont disponibles ; c'est le poids de calage w_i^{EQ} de l'EQ qui interviendra comme poids initial dans le calage du LS et sera donc corrigé pour aboutir au poids de calage (longitudinal) final pour l'estimation des transitions.

2.3 Les distributions de référence

Les échantillons BQ et EQ sont des échantillons de référence, à partir desquels les estimations des distributions de référence du statut BIT sont calculées pour le calage du LS. Pour ce faire, le poids de calage est utilisé : w_i^{BQ} pour le calcul des distributions de référence à partir de l'échantillon BQ et w_i^{EQ} pour le calcul des distributions de référence à partir de l'échantillon EQ. De cette manière, la distribution estimée de STAT1 (limitée à la tranche d'âge 15-74 ans) au BQ par exemple peut être estimée : pour chaque valeur b de STAT1, la somme des poids de calage des répondants pour lesquels STAT1 = b est une estimation du nombre de personnes dans la population (limitée à la tranche d'âge 15-74 ans) avec STAT1 = b, et cette estimation est utilisée comme référence dans le calage du LS (si STAT1 est une variable de calage).

Pour BQ 2018T3, la distribution de la population estimée du statut BIT (STAT1)[9] qui sera une des variables de calage les plus importantes, se présente comme suit :

2018T3	Chômeur	Occupé	Inactif	*Total*
Absolu (nombre de personnes)	300.215,88	4.785.248,91	3.325.058,21	8.410.523,00
Relatif (% de personnes)	3,57 %	56,90 %	39,56 %	100,00 %

Pour EQ 2018T4, la distribution de la population estimée du statut BIT (STAT2) qui sera une des variables de calage les plus importantes, se présente comme suit :

2018T4	Chômeur	Occupé	Inactif	*Total*
Absolu (nombre de personnes)	290.402,66	4.804.732,96	3.335.187,38	8.430.323,00
Relatif (% de personnes)	3,44 %	56,99 %	39,56 %	100,00 %

Les distributions de référence du statut BIT seront également calculées par sexe, par région, etc. et utilisées dans le calage des LS. Par ailleurs, des distributions estimées de variables contextuelles (sexe, région, catégorie d'âge, etc.) – sans intervention du statut BIT – sont également utilisées.

2.4 Modèles de calage de base, pour des matrices de transition globales et ventilées selon le sexe

L'objectif principal du calage d'un LS est de rendre les marges des matrices de transition du statut BIT estimées cohérentes avec les distributions connues du statut BIT au BQ et au EQ. Dans ce paragraphe, nous construisons progressivement des

[9] Statbel ne publie aucun chiffre (trimestriel ou annuel) pour la population âgée de 15 à 74 ans ; mais ces chiffres sont disponibles sur le site web d'Eurostat. Davantage d'informations sur les chiffres publiés sont disponibles au chapitre 3.

modèles de calage appropriés à cette fin et discutons des difficultés rencontrées. Dans les paragraphes suivants 2.5, 2.6 et 2.7, ces modèles de calage sont étendus pour atteindre des objectifs supplémentaires.

2.4.1 Calage sur les distributions globales du statut BIT

Dans ce sous-paragraphe, notre objectif est d'adapter la matrice de transition globale initiale de la paire de trimestres 2018T3 et 2018T4 aux distributions globales estimées du statut BIT au BQ (2018T3) et au EQ (2018Q4).

Le LS pour la paire de trimestres successifs 2018T3-2018T4 contient 13.510 répondants (dans la classe d'âge 15-74 ans pour les deux trimestres), pour lesquels le statut BIT est connu au 2018T3 (STAT1) et 2018T4 (STAT2). La matrice de transition de l'échantillon non pondéré est présentée au Tableau 1. Il convient de noter que plus de 93 % des répondants ne changent pas de statut BIT.

Tableau 1 Échantillon longitudinal non pondéré 2018T3-2018T4, selon le statut BIT au BQ et au EQ

Statut BIT 2018T3	**Statut BIT 2018T4** Chômeur	Occupé	Inactif	Total
Chômeur	224	120	130	474
Occupé	57	7.102	296	7.455
Inactif	93	239	5.249	5.581
Total	374	7.461	5.675	13.510

En utilisant les poids de calage w_i^{EQ} de 2018T4 pour l'ensemble des 13.510 répondants i du LS, nous obtenons la matrice de transition 3x3 pondérée initiale, avec des marges (c.-à-d. les totaux des lignes et des colonnes, intitulés "Total"), dans le Tableau 2.

Tableau 2 Matrice de transition initiale 2018T3-2018T4, et distributions de référence

Statut BIT 2018T3	**Statut BIT 2018T4** Chômeur	Occupé	Inactif	Total	Distribution réf. 2018T3 (a)	(b)	(c)
Chômeur	75.566,19	40.185,88	42.546,79	158.298,86	*300.215,88*	*300.922,65*	*300.215,88*
Occupé	20.888,66	2.304.490,45	93.986,41	2.419.365,52	*4.785.248,91*	*4.796.514,31*	*4.785.248,91*
Inactif	40.808,23	85.134,34	1.466.478,06	1.592.420,62	*3.325.058,21*	*3.332.886,05*	*3.344.858,21*
Total	137.263,08	2.429.810,66	1.603.011,26	4.170.085,00	***8.410.523,00***	***8.430.323,00***	***8.430.323,00***
Distribution réf. 2018T4	*290.402,66*	*4.804.732,96*	*3.335.187,38*	***8.430.323,00***			

Le total global des poids w_i^{EQ} pour les répondants du LS ne s'élève qu'à 4.170.085,00, ce qui est très peu par rapport aux chiffres estimés de la population totale (pour le groupe d'âge 15-74 ans) de 8.410.523,00 au 2018T3 et de 8.430.323,00 au 2018T4. Cela s'explique par le fait que le LS ne couvre que deux des quatre RG de chaque trimestre, comme nous l'avons vu précédemment.

L'adaptation de la matrice de transition du Tableau 2 aux distributions de référence des BQ et EQ, que nous retrouvons également dans le Tableau 2, dans la colonne intitulée "Distribution réf. 2018T3 (a)" et dans la ligne intitulée "Distribution réf. 2018T4", est un calage du LS selon un modèle ayant la structure linéaire suivante :

$$\text{STAT1} + \text{STAT2} \qquad \text{(LS-1)}$$

Ce calage ne peut pas être effectué immédiatement, en raison de l'incohérence numérique entre les distributions de référence : les totaux sont différents. Selon la théorie du calage, cela signifie que les équations de calage du modèle LS-1 sont incohérentes d'un point de vue numérique. Cette incohérence peut être supprimée de deux manières (au moins) :

- La *méthode classique* : la distribution de référence "Distribution réf. 2018T3 (a)" du BQ est multipliée par le facteur 8.430.323,00/8.410.523,00 $\cong$ 1,002354. Nous trouvons la distribution corrigée dans la colonne intitulée "Distribution réf. 2018T3 (b)" du Tableau 2 ; le soulignement indique que le chiffre de la colonne (a) est adapté pour chaque statut BIT.

- La *méthode d'Eurostat*[10] : seul le chiffre du statut BIT *Inactif* de la distribution "Distribution réf. 2018T3 (a)" du BQ est changé de la différence 8.430.323,00 – 8.410.523,00 = 19.800,00 ; c.-à-d. que 3.325.058,21 est adapté et devient 3.325.058,21 + 19.800,00 = 3.344.858,21. Nous trouvons la distribution corrigée dans la colonne intitulée "Distribution réf. 2018T3 (c)" du Tableau 2 ; le soulignement indique que seul le chiffre du statut BIT *Inactif* est adapté.

Il convient de noter que le facteur 1,002354 dans la méthode classique, ainsi que le changement 19.800,00 dans la méthode d'Eurostat reflètent une croissance (probable) de la population. Ces méthodes fonctionnent aussi lorsque la population diminue.

Après avoir appliqué l'une ou l'autre des deux méthodes de réalisation de la cohérence numérique – on parle de méthodes ou de modèles NC –, le LS peut être redressé, d'après le modèle LS-1, sur une distribution de référence corrigée du statut BIT au BQ (2018T3) et la distribution de référence du statut BIT au EQ (2018T4). Le résultat des deux calages pour la paire 2018T3-2018T4 est présenté au Tableau 3. Notez que les marges de la matrice de transition sont effectivement égales aux distributions de référence (corrigées) correspondantes.

Le calage selon le modèle LS-1 peut être effectué à l'aide d'une méthode *iterative proportional fitting* (méthode IPF). C'est la méthode appliquée au sein d'Eurostat (2015b) (voir par. 2.8) ; voir aussi annexe B.5. Statbel applique la méthode plus universelle de Newton-Raphson, via la macro SAS® CALMAR2, principalement en vue d'extensions du modèle simple LS-1. Afin d'appliquer CALMAR2, la méthode ou fonction de calage doit être choisie (voir annexe B). Statbel opte pour la méthode exponentielle, parce qu'elle correspond à la méthode IPF. Afin de conserver cette correspondance entre les méthodes de Statbel et d'Eurostat, Statbel opte toujours pour la méthode exponentielle pour appliquer les modèles SL développés ci-après (via CALMAR2).

[10] Nous parlons de la "méthode d'Eurostat" car, selon le rapport d'Eurostat (2015a) de la Task Force Flow Statistics, cette méthode fait l'objet d'un consensus: "*The favoured approach prioritises consistency with the target quarter, i.e. guarantees that the total longitudinal population is identical to the one of the target quarter and the more recent figures of employed, unemployed, and inactive in that quarter are exactly met when adding up the levels of the transition matrix. For the initial quarter this would only be the case for employment and unemployment – inactivity would serve as a residual category, i.e. possible total population differences between the two quarters would be assigned to the inactive population in the initial quarter.*" et "*Similar weighting conditions enforcing consistency with five of the six marginal values are used by several countries producing flow statistics already.*"

Tableau 3 Estimation des transitions globales 2018T3-2018T4 avec le modèle LS-1, après correction globale de l'incohérence numérique

Après application de la méthode classique pour supprimer l'incohérence numérique							
Statut BIT	**Statut BIT 2018T4**				Distribution réf. 2018T3		
2018T3	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	151.285,33	71.737,93	77.899,39	300.922,65	*300.215,88*	*300.922,65*	*300.215,88*
Occupé	46.349,14	4.559.446,13	190.719,04	4.796.514,31	*4.785.248,91*	*4.796.514,31*	*4.785.248,91*
Inactif	92.768,19	173.548,90	3.066.568,96	3.332.886,05	*3.325.058,21*	*3.332.886,05*	*3.344.858,21*
Total	290.402,66	4.804.732,96	3.335.187,38	**8.430.323,00**	***8.410.523,00***	***8.430.323,00***	***8.430.323,00***
Distribution réf. 2018T4	*290.402,66*	*4.804.732,96*	*3.335.187,38*	***8.430.323,00***			
Après application de la méthode d'Eurostat pour supprimer l'incohérence numérique							
Statut BIT	**Statut BIT 2018T4**				Distribution réf. 2018T3		
2018T3	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	150.637,85	72.628,75	76.949,28	300.215,88	*300.215,88*	*300.922,65*	*300.215,88*
Occupé	45.528,92	4.553.865,52	185.854,46	4.785.248,91	*4.785.248,91*	*4.796.514,31*	*4.785.248,91*
Inactif	94.235,89	178.238,69	3.072.383,63	3.344.858,21	*3.325.058,21*	*3.332.886,05*	*3.344.858,21*
Total	290.402,66	4.804.732,96	3.335.187,38	**8.430.323,00**	***8.410.523,00***	***8.430.323,00***	***8.430.323,00***
Distribution réf. 2018T4	*290.402,66*	*4.804.732,96*	*3.335.187,38*	***8.430.323,00***			

Les cellules grisées indiquent des chiffres issus des échantillons BQ et EQ redressés initialement, qui sont reproduits exactement : la méthode d'Eurostat en reproduit davantage que la méthode classique. Si la méthode classique ne reproduit pas la distribution absolue du statut BIT au BQ 2018T3, elle reproduit la distribution relative du statut BIT au BQ 2018T3 : en effet, en termes relatifs, les distributions sous "Distribution réf. 2018T3 (a)" et "Distribution réf. 2018T3 (b)" sont exactement les mêmes. Avec la méthode d'Eurostat par contre, les chiffres absolus de chômeurs et de personnes occupées sont reproduits exactement mais pas la distribution relative du statut BIT au BQ 2018T3 : en effet, en termes relatifs, les distributions sous "Distribution réf. 2018T3 (a)" et "Distribution réf. 2018T3 (c)" sont différentes.

La méthode classique et celle d'Eurostat entrainent des matrices de transition estimées différentes, comme le montre le Tableau 3. En chiffres absolus, la différence est la plus importante (en valeur absolue) pour la transition *Inactif-Inactif* (5.814,67 = 3.072.383,63 – 3.066.568,96) ; en chiffres relatifs, les différences sont moins frappantes (voir Tableau 4) : la plus grande différence s'élève à 0,35 point de pourcentage pour la transition *Chômeur-Occupé*.

Tableau 4 Estimation des transitions relatives globales 2018T3-2018T4 avec le modèle LS-1, après correction globale de l'incohérence numérique

	Après la méthode classique				*Après la méthode d'Eurostat*			
Statut BIT 2018T3	**Statut BIT 2018T4**				**Statut BIT 2018T4**			
	Chômeur	Occupé	Inactif	Total	Chômeur	Occupé	Inactif	Total
Chômeur	50,27	23,84	25,89	100,00	50,18	24,19	25,63	100,00
Occupé	0,97	95,06	3,98	100,00	0,95	95,16	3,88	100,00
Inactif	2,78	5,21	92,01	100,00	2,82	5,33	91,85	100,00
Total	3,44	56,99	39,56	100,00	3,44	56,99	39,56	100,00
Distribution réf. 2018T4	3,44	56,99	39,56	100,00	3,44	56,99	39,56	100,00

2.4.2 Calage sur les distributions du statut BIT d'un sexe spécifique

Ensuite, notre objectif est de redresser les matrices de transition initiales pour les hommes et les femmes en fonction des distributions de référence du statut BIT selon le sexe. Formellement, cela revient à appliquer un modèle de calage pour le LS avec la structure linéaire suivante :

$$SEX*(STAT1 + STAT2) = SEX*STAT1 + SEX*STAT2 \qquad \text{(LS-2)}$$

Pour la paire 2018T3-2018T4, les matrices de transition initiales (utilisant de nouveau les poids w_i^{EQ}), ainsi que les distributions de référence initiales de la colonne intitulée “Distribution réf. 2018T3 (a)” et des lignes intitulées “Distribution réf. 2018T4”, sont présentées dans le Tableau 5. Nous aborderons ci-dessous les colonnes intitulées “Distribution réf. 2018T3 (b)” et “Distribution réf. 2018T3 (c)”, mais nous notons déjà que les sommes par sexe donnent les chiffres du Tableau 2, sauf pour la colonne intitulée “Distribution réf. 2018T3 (b)”.

Tableau 5 Matrices de transition initiales 2018T3-2018T4, et distributions de référence, par sexe

Hommes							
Statut BIT 2018T3	**Statut BIT 2018T4**				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	43.362,60	23.495,56	19.128,01	85.986,16	*170.610,00*	*171.013,58*	*170.610,00*
Occupé	11.590,60	1.230.669,29	46.394,42	1.288.654,31	*2.530.132,88*	*2.536.117,96*	*2.530.132,88*
Inactif	21.659,78	41.811,68	649.808,94	713.280,40	*1.494.108,13*	*1.497.642,47*	*1.504.031,13*
Total	76.612,98	1.295.976,53	715.331,37	2.087.920,88	***4.194.851,00***	***4.204.774,00***	***4.204.774,00***
Distribution réf. 2018T4	*165.517,27*	*2.547.516,81*	*1.491.739,92*	***4.204.774,00***			
Femmes							
Statut BIT 2018T3	**Statut BIT 2018T4**				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	32.203,59	16.690,32	23.418,78	72.312,69	*129.605,88*	*129.909,54*	*129.605,88*
Occupé	9.298,06	1.073.821,16	47.591,99	1.130.711,21	*2.255.116,03*	*2.260.399,59*	*2.255.116,03*
Inactif	19.148,45	43.322,65	816.669,12	879.140,22	*1.830.950,09*	*1.835.239,87*	*1.840.827,09*
Total	60.650,10	1.133.834,13	887.679,89	2.082.164,12	***4.215.672,00***	***4.225.549,00***	***4.225.549,00***
Distribution réf. 2018T4	*124.885,40*	*2.257.216,15*	*1.843.447,46*	***4.225.549,00***			

Encore une fois, les distributions de référence du BQ doivent être adaptées pour que le système d'équations de calage du modèle LS-2 soit numériquement cohérent. Par sexe, nous corrigeons la distribution de référence de 2018T3 sur la distribution de référence de 2018T4 :

- Avec la méthode classique : pour les hommes, la distribution sous “Distribution réf. 2018T3 (a)” est multipliée par le facteur 4.204.774,00/4.194.851,00 ≅ 1,002366, et nous trouvons la distribution sous “Distribution réf. 2018T3 (b)” ; pour les femmes, la distribution sous “Distribution réf. 2018T3 (a)” est multipliée par le facteur 4.225.549,00/4.215.672,00 ≅ 1,002343, et nous trouvons la distribution sous “Distribution réf. 2018T3 (b)”.
- Avec la méthode d’Eurostat : pour les hommes, le chiffre du statut BIT *Inactif* de la distribution sous “Distribution réf. 2018T3 (a)” est changé de la différence 4.204.774,00 – 4.194.851,00 = 9.923,00, et nous trouvons la distribution sous “Distribution réf. 2018T3 (c)” ; pour les femmes, le chiffre statut BIT *Inactif* de la distribution sous “Distribution réf. 2018T3 (a)” est changé de la différence 4.225.549,00 – 4.215.672,00 = 9.877,00, et nous trouvons la distribution sous “Distribution réf. 2018T3 (c)”.

Le modèle de calage LS-2 peut alors être appliqué. Cela peut se faire avec la méthode IPF pour les hommes et les femmes séparément. Statbel opte toutefois pour l'utilisation de CALMAR2, avec la méthode de calage exponentielle, comme déjà indiqué dans le sous-paragraphe précédent.

Les matrices de transition obtenues sont présentées dans le Tableau 6, et les matrices de transition relatives obtenues dans le Tableau 7. Comme lors du calage avec le modèle LS-1, nous constatons également que lors du calage avec le modèle LS-2, la méthode classique et celle d’Eurostat visant à obtenir une cohérence numérique ne montrent pas de très grandes différences dans les résultats. Les chiffres grisés sont les chiffres issus des échantillons BQ et EQ redressés initiaux qui sont reproduits de manière exacte ; voir également les remarques du Tableau 3.

Tableau 6 Estimation des transitions d'un sexe spécifique 2018T3-2018T4 avec le modèle LS-2, après corrections par sexe de l'incohérence numérique

Après application de la méthode classique pour obtenir la cohérence numérique

Hommes

Statut BIT 2018T3	**Statut BIT 2018T4** Chômeur	Occupé	Inactif	Total	Distribution réf. 2018T3 (a)	(b)	(c)
Chômeur	90.507,62	43.967,79	36.538,17	171.013,58	*170.610,00*	*171.013,58*	*170.610,00*
Occupé	25.397,17	2.417.684,53	93.036,26	2.536.117,96	*2.530.132,88*	*2.536.117,96*	*2.530.132,88*
Inactif	49.612,48	85.864,50	1.362.165,49	1.497.642,47	*1.494.108,13*	*1.497.642,47*	*1.504.031,13*
Total	165.517,27	2.547.516,81	1.491.739,92	**4.204.774,00**	***4.194.851,00***	***4.204.774,00***	***4.204.774,00***
Distribution réf. 2018T4	*165.517,27*	*2.547.516,81*	*1.491.739,92*	***4.204.774,00***			

Femmes

Statut BIT 2018T3	**Statut BIT 2018T4** Chômeur	Occupé	Inactif	Total	Distribution réf. 2018T3 (a)	(b)	(c)
Chômeur	61.033,08	28.125,93	40.750,53	129.909,54	*129.605,88*	*129.909,54*	*129.605,88*
Occupé	20.854,76	2.141.538,41	98.006,42	2.260.399,59	*2.255.116,03*	*2.260.399,59*	*2.255.116,03*
Inactif	42.997,56	87.551,81	1.704.690,50	1.835.239,87	*1.830.950,09*	*1.835.239,87*	*1.840.827,09*
Total	124.885,40	2.257.216,15	1.843.447,46	**4.225.549,00**	***4.215.672,00***	***4.225.549,00***	***4.225.549,00***
Distribution réf. 2018Q4	*124.885,40*	*2.257.216,15*	*1.843.447,46*	***4.225.549,00***			

Après application de la méthode d'Eurostat pour obtenir une cohérence numérique

Hommes

Statut BIT 2018T3	**Statut BIT 2018T4** Chômeur	Occupé	Inactif	Total	Distribution réf. 2018T3 (a)	(b)	(c)
Chômeur	90.185,46	44.455,67	35.968,86	170.610,00	*170.610,00*	*171.013,58*	*170.610,00*
Occupé	24.997,79	2.414.666,63	90.468,45	2.530.132,88	*2.530.132,88*	*2.536.117,96*	*2.530.132,88*
Inactif	50.334,01	88.394,51	1.365.302,60	1.504.031,13	*1.494.108,13*	*1.497.642,47*	*1.504.031,13*
Total	165.517,27	2.547.516,81	1.491.739,92	**4.204.774,00**	***4.194.851,00***	***4.204.774,00***	***4.204.774,00***
Distribution réf. 2018T4	*165.517,27*	*2.547.516,81*	*1.491.739,92*	***4.204.774,00***			

Femmes

Statut BIT 2018T3	**Statut BIT 2018T4** Chômeur	Occupé	Inactif	Total	Distribution réf. 2018T3 (a)	(b)	(c)
Chômeur	60.700,65	28.514,91	40.390,33	129.605,88	*129.605,88*	*129.909,54*	*129.605,88*
Occupé	20.433,80	2.138.981,63	95.700,60	2.255.116,03	*2.255.116,03*	*2.260.399,59*	*2.255.116,03*
Inactif	43.750,95	89.719,61	1.707.356,53	1.840.827,09	*1.830.950,09*	*1.835.239,87*	*1.840.827,09*
Total	124.885,40	2.257.216,15	1.843.447,46	4.225.549,00	***4.215.672,00***	***4.225.549,00***	***4.225.549,00***
Distribution réf. 2018T4	*124.885,40*	*2.257.216,15*	*1.843.447,46*	***4.225.549,00***			

Tableau 7 Estimation des taux de transition d'un sexe spécifique 2018T3-2018T4 avec le modèle LS-2, après corrections par sexe de l'incohérence numérique

Hommes								
	Après la méthode classique				*Après la méthode d'Eurostat*			
Statut BIT 2018T3	**Statut BIT 2018T4**				**Statut BIT 2018T4**			
	Chômeur	Occupé	Inactif	Total	Chômeur	Occupé	Inactif	Total
Chômeur	50,92	25,71	21,37	100,00	52,86	26,06	21,08	100,00
Occupé	1,00	95,33	3,67	100,00	0,99	95,44	3,58	100,00
Inactif	3,31	5,73	90,95	100,00	3,35	5,88	90,78	100,00
Total	3,94	60,59	35,48	100,00	3,94	60,59	35,48	100,00
Distribution réf. 2018T4	3,94	60,59	35,48	100,00	3,94	60,59	35,48	100,00
Femmes								
	Après la méthode classique				*Après la méthode d'Eurostat*			
Statut BIT 2018T3	**Statut BIT 2018T4**				**Statut BIT 2018T4**			
	Chômeur	Occupé	Inactif	Total	Chômeur	Occupé	Inactif	Total
Chômeur	46,98	21,65	31,37	100,00	46,83	22,00	31,16	100,00
Occupé	0,92	94,74	4,34	100,00	0,91	94,85	4,24	100,00
Inactif	2,34	4,77	92,89	100,00	2,38	4,87	92,75	100,00
Total	2,96	53,42	43,63	100,00	2,96	53,42	43,63	100,00
Distribution réf. 2018T4	2,96	53,42	43,63	100,00	2,96	53,42	43,63	100,00

Après le calage selon le modèle LS-2, les matrices de transition et les distributions de référence pour les hommes et les femmes peuvent être additionnées : le résultat est les matrices de transition globales et les distributions de référence présentées dans le Tableau 8. Nous constatons que :

(1) les matrices de transition 3x3 du Tableau 8 diffèrent relativement peu des matrices de transition 3x3 du Tableau 3, quelle que soit la méthode utilisée pour supprimer l'incohérence numérique ;

(2) les totaux des colonnes correspondants (des lignes intitulées "Total") de ces matrices sont exactement les mêmes dans le Tableau 3 et le Tableau 8. Ces totaux représentent également la distribution de référence globale pour le EQ 2018T4 ;

(3) les distributions de référence sous l'intitulé "Distribution réf. 2018T3 (a)" sont toutes identiques. Il s'agit de la distribution de référence globale initiale pour le BQ 2018T3 ;

(4) les distributions de référence sous l'intitulé "Distribution réf. 2018T3 (c)" sont toutes identiques. Il s'agit de la distribution de référence corrigée pour le BQ2018T3 obtenue avec la méthode d'Eurostat ;

(5) les totaux des lignes correspondants (des colonnes intitulées "Total") de ces matrices sont exactement les mêmes dans le Tableau 3 et le Tableau 8 pour la méthode d'Eurostat ;

(6) les distributions de référence sous l'intitulé "Distribution réf. 2018T3 (b)" diffèrent dans le Tableau 3 et le Tableau 8. Il s'agit de deux versions des distributions de référence corrigées pour le BQ2018T3 obtenues avec la méthode classique ;

(7) les totaux des lignes correspondants (des colonnes intitulées "Total") de ces matrices ne sont pas les mêmes dans le Tableau 3 et le Tableau 8 pour la méthode classique.

Les constatations des points (2) et (3) sont évidentes : additionner des nombres estimés, tant au BQ qu'au EQ, pour les hommes et les femmes séparément dans n'importe quelle sous-population (p.ex. les chômeurs) pour obtenir le nombre total estimé de personnes de cette sous-population. La constatation du point (4) est évidente pour cette même raison, mais également parce que l'adaptation selon la méthode d'Eurostat ne vaut que pour les inactifs, et parce que la somme des adaptations distinctes des hommes inactifs et des femmes inactives donne l'adaptation globale des inactifs. La constatation au point (5) est en lien direct avec celle du point (4). La cause du problème mentionné aux points (6) et (7) lié à la méthode classique est l'application de différents facteurs de correction pour les hommes et les femmes afin d'obtenir une cohérence numérique. Il s'agit d'un argument en faveur du choix de la méthode d'Eurostat.

Le fait que les différences, après application de la méthode classique, citées aux points (6) et (7), soient faibles découle du fait que le facteur de correction global (1,002354 ; voir paragraphe précédent) ne diffère que légèrement des facteurs de

correction par sexe (1,002366 pour les hommes et 1,002343 pour les femmes ; voir ci-dessus). Pour d'autres ventilations de la matrice de transition que celle selon le sexe, on peut s'attendre à de plus grandes différences.

Tableau 8 Estimation des transitions globales 2018T3-2018T4 avec le modèle LS-2, après correction par sexe pour l'incohérence numérique

Après application de la méthode classique pour supprimer l'incohérence numérique

Statut BIT 2018T3	**Statut BIT 2018T4**				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	151.540,70	72.093,71	77.288,70	300.923,12	*300.215,88*	*300.923,12*	*300.215,88*
Occupé	46.251,92	4.559.222,94	191.042,68	4.796.517,55	*4.785.248,91*	*4.796.517,55*	*4.785.248,91*
Inactif	92.610,04	173.416,30	3.066.855,99	3.332.882,33	*3.325.058,21*	*3.332.882,33*	*3.344.858,21*
Total	290.402,66	4.804.732,96	3.335.187,38	**8.430.323,00**	***8.410.523,00***	***8.430.323,00***	***8.430.323,00***
Distribution réf. 2018T4	*290.402,66*	*4.804.732,96*	*3.335.187,38*	***8.430.323,00***			

Après application de la méthode d'Eurostat pour supprimer l'incohérence numérique

Statut BIT 2018T3	**Statut BIT 2018T4**				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	150.886,11	72.970,58	76.359,19	300.215,88	*300.215,88*	*300.923,12*	*300.215,88*
Occupé	45.431,59	4.553.648,26	186.169,05	4.785.248,91	*4.785.248,91*	*4.796.517,55*	*4.785.248,91*
Inactif	94.084,96	178.114,12	3.072.659,13	3.344.858,21	*3.325.058,21*	*3.332.882,33*	*3.344.858,21*
Total	290.402,66	4.804.732,96	3.335.187,38	**8.430.323,00**	***8.410.523,00***	***8.430.323,00***	***8.430.323,00***
Distribution réf. 2018T4	*290.402,66*	*4.804.732,96*	*3.335.187,38*	***8.430.323,00***			

Notez que le modèle LS-2 peut aussi être formulé comme suit :

STAT1 + STAT2 + SEX*STAT1 + SEX*STAT2 (LS-2a)

Cela explique que le modèle LS-2 permet non seulement de redresser en fonction de deux distributions de référence du statut BIT par sexe, mais aussi implicitement sur deux distributions de référence globales du statut BIT. Avec le modèle LS-1, le calage se fait uniquement sur les deux distributions de référence globales du statut BIT. Les distributions de référence globales pour le BQ 2018T3 ne sont pas identiques pour les modèles LS-1 et LS-2 lorsque la méthode classique est appliquée pour corriger l'incohérence numérique. Comme démontré ci-dessus, ce n'est pas la conséquence directe de la différence entre les modèles LS-1 et LS-2, mais bien de la différence dans la correction classique préalable de l'incohérence numérique. Pour les deux modèles nous avons en effet réalisé la correction nécessaire *minimale*, à savoir

- une correction globale lorsque le modèle LS-1 – c.-à-d. STAT1 + STAT2 – est appliqué ;
- une correction par sexe lorsque le modèle LS-2 – c.-à-d. SEX*(STAT1 + STAT2) – est appliqué.

Afin de résoudre ce problème de modifications des distributions de référence globales pour le BQ 2018T3 en appliquant la méthode classique pour supprimer les incohérences numériques, nous pouvons également appliquer la correction par sexe si les matrices de transition sont redressées selon le modèle LS-1. Le calage selon le modèle LS-1, après application de la méthode classique par sexe, donne les estimations et les distributions de référence dans le Tableau 9 (volet supérieur). Les estimations des transitions dans le Tableau 9 correspondent très bien aux estimations du Tableau 3 ; les différences sont dues aux différences dans la méthode de correction – globale pour le Tableau 3, par sexe pour le Tableau 9 – ce qui se reflète principalement dans les différences de distribution de référence corrigée pour le BQ 2018T3. Notez qu'il n'y a aucune différence lorsque la méthode Eurostat est appliquée.

Tableau 9 Estimation des transitions globales 2018T3-2018T4 avec le modèle LS-1, après correction par sexe pour l'incohérence numérique

Après application de la méthode classique pour supprimer l'incohérence numérique

Statut BIT 2018T3	**Statut BIT 2018T4**				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	151.285,55	71.737,77	77.899,79	300.923,12	*300.215,88*	*300.923,12*	*300.215,88*
Occupé	46.349,33	4.559.447,69	190.720,53	4.796.517,55	*4.785.248,91*	*4.796.517,55*	*4.785.248,91*
Inactif	92.767,78	173.547,49	3.066.567,06	3.332.882,33	*3.325.058,21*	*3.332.882,33*	*3.344.858,21*
Total	290.402,66	4.804.732,96	3.335.187,38	**8.430.323,00**	***8.410.523,00***	***8.430.323,00***	***8.430.323,00***
Distribution réf. 2018T4	*290.402,66*	*4.804.732,96*	*3.335.187,38*	***8.430.323,00***			

Après application de la méthode d'Eurostat pour supprimer l'incohérence numérique

Statut BIT 2018T3	**Statut BIT 2018T4**				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	150.637,85	72.628,75	76.949,28	300.215,88	*300.215,88*	*300.923,12*	*300.215,88*
Occupé	45.528,92	4.553.865,52	185.854,46	4.785.248,91	*4.785.248,91*	*4.796.517,55*	*4.785.248,91*
Inactif	94.235,89	178.238,69	3.072.383,63	3.344.858,21	*3.325.058,21*	*3.332.882,33*	*3.344.858,21*
Total	290.402,66	4.804.732,96	3.335.187,38	**8.430.323,00**	***8.410.523,00***	***8.430.323,00***	***8.430.323,00***
Distribution réf. 2018T4	*290.402,66*	*4.804.732,96*	*3.335.187,38*	***8.430.323,00***			

2.5 Méthodes NC comme modèles de calage

2.5.1 Introduction

Parmi les méthodes NC introduites au par. 2.4, à savoir les méthodes permettant d'obtenir la cohérence numérique entre les distributions de référence de BQ et EQ, nous distinguons deux classes de méthodes : les méthodes classiques, que nous appellerons les méthodes NC-C, et les méthodes d'Eurostat, que nous appellerons les méthodes NC-E. Nous avons montré que les méthodes NC sont nécessaires avant que les modèles de calage puissent être appliqués aux LS – appelés modèles (de calage) LS.

Dans ce paragraphe, nous examinerons les méthodes NC plus en détail. Nous les examinons sous l'angle de l'élargissement ou l'affinement ultérieurs des modèles de calage LS : ces modèles nécessiteront des méthodes NC plus détaillées.

Il semble possible et pratique de formuler et d'employer les méthodes NC comme modèles de calage ; nous parlons alors de modèles (de calage) NC et nous faisons, le cas échéant, la distinction entre les modèles NC-C et NC-E. Cela permet d'une part de réaliser de manière efficace et universelle les calculs pour obtenir les cohérences numériques souhaitées et nécessaires entre les distributions du statut BIT des échantillons BQ et EQ (tous deux redressés) afin que le statisticien (pragmatique) qui souhaite appliquer un modèle LS spécifique puisse rapidement assurer les cohérences numériques nécessaires. Cela permet e.a. aussi de passer de manière flexible d'un modèle LS à un autre, afin de comparer divers modèles LS et de choisir un modèle définitif. D'autre part, cela nous permet aussi de présenter une description claire et formelle des méthodes ou des modèles, une comparaison des modèles et le choix d'un modèle définitif.

Les méthodes NC-C peuvent évidemment être formulées comme des modèles de calage ; il s'agit alors d'un calage de l'échantillon redressé BQ sur les distributions estimées qui sont entièrement déterminées à partir de l'échantillon redressé EQ. La formulation des méthodes NC-E comme modèles de calage est moins évidente. Il s'agit ici d'un calage de l'échantillon redressé BQ sur des informations agrégées combinées provenant à la fois de l'échantillon redressé EQ et de l'échantillon redressé BQ. Les modèles NC-E qui en résultent ont par ailleurs une particularité : les totaux de calage peuvent être négatifs.

Dans les paragraphes suivants, nous présenterons les méthodes NC déjà appliquées plus haut comme des modèles de calage NC via leur structure linéaire. Nous expliquerons aussi les extensions possibles de ces modèles NC, ainsi que la possibilité

d'obtenir des totaux de calage négatifs pour la classe des modèles NC-E et la méthode de calage choisie en fonction de ces résultats. Ce dernier point sera illustré par un exemple. La formulation mathématique des modèles NC-C et NC-E est traitée dans l'annexe C.

Pour les principes généraux et la terminologie des modèles de calage, nous faisons référence à l'annexe B.

2.5.2 Modèles NC-C et NC-E : variantes

Au par. 2.4, nous avons déjà appliqué les méthodes ou modèles NC suivants pour corriger les incohérences numériques :

- Pour adapter la distribution globale du statut BIT du BQ à celle du EQ avec la méthode classique :

  1 (NC-C-1)

- Pour adapter les distributions par sexe du statut BIT du BQ à celle du EQ avec la méthode classique :

  SEX (NC-C-2)

- Pour adapter la distribution globale du statut BIT du BQ à celle du EQ avec la méthode Eurostat :

  STAT1 (NC-E-1)

- Pour adapter les distributions par sexe du statut BIT du BQ à celle du EQ avec la méthode Eurostat :

  SEX * STAT1 (NC-E-2)

Le modèle NC-C-1 produit exactement un facteur de correction global, qui est appliqué au poids de calage de chaque répondant de l'échantillon BQ. Pour le modèle NC-E-1, nous pouvons dire par analogie qu'il en résulte également un facteur de correction global, qui n'est toutefois appliqué qu'au poids de calage de chaque répondant inactif de l'échantillon BQ ; le facteur de correction pour tous les répondants chômeurs ou occupés de l'échantillon BQ est exactement 1.

Le modèle NC-C-2 produit exactement deux facteurs de correction, à savoir un pour chaque sexe ; un facteur est appliqué au poids de calage de chaque répondant masculin de l'échantillon BQ, l'autre au poids de calage de chaque répondant féminin de l'échantillon BQ. Pour le modèle NC-E-2, nous pouvons affirmer de manière analogue qu'il en résulte deux facteurs de correction, à savoir un pour chaque sexe ; un facteur est appliqué au poids de calage de chaque répondant inactif masculin de l'échantillon BQ, l'autre au poids de calage de chaque répondante inactive de l'échantillon BQ ; le facteur de correction pour tous les répondants chômeurs et occupés, hommes et femmes, de l'échantillon BQ est exactement 1.

Dans l'annexe C, nous démontrons mathématiquement que les méthodes non seulement NC-C, mais aussi NC-E peuvent être considérées comme modèles de calage ; ce qui explique notamment la notation ci-dessus via la structure linéaire. Cela permet d'affiner et d'appliquer efficacement – à condition de développer un logiciel approprié – les modèles NC en incluant davantage de variables contextuelles. Cela est nécessaire en vue de l'extension (voir par. 2.6) des modèles de calage LS-1 et LS-2 lorsque des matrices de transition doivent être ventilées selon d'autres variables contextuelles que (seulement) le sexe. Les modèles suivants offrent également la possibilité de corriger les incohérences numériques entre les distributions de référence :

SEX * REG1 * EDU1 * AGE1 * NAT1 (NC-C-3)

(SEX * REG1 * EDU1 * AGE1 * NAT1) * STAT1 (NC-E-3)

Notez que les modèles NC-C-1 jusque 3 et les modèles NC-E-1 jusque 3 sont tous de type "post-stratification". Cela signifie qu'en général, les facteurs de correction (uniques) peuvent être calculés séparément par cellule à l'intersection de toutes les variables concernées, ce qui ne nécessite pas de logiciel sophistiqué. Pour NC-E-3, il faut, tout comme pour NC-E-1 et NC-E-2, que les facteurs de correction pour les chômeurs et les actifs soient toujours égaux à 1.

Le modèle NC-C-3 exige qu'au moins un répondant ait été trouvé dans l'échantillon BQ pour chaque cellule au croisement des variables de calage concernées et qu'une estimation du chiffre de la population puisse être calculé sur base de l'échantillon EQ. Dans le cas contraire, le modèle NC-C-3 doit être "simplifié", ce qui est, par exemple, possible en :

(i) regroupant une ou plusieurs variables ;

(ii) renonçant au croisement complet des variables de calage ;

(iii) supprimant une ou plusieurs variables de calage ;

(iv) utilisant une combinaison de (i), (ii) et/ou (iii).

Certaines variantes de NC-C-3 sont par exemple :

SEX * REG1 * EDU1 * AGE1 * NAT1 (NC-C-3a)

SEX + REG1 + EDU1 + AGE1 + NAT1 (NC-C-3b)

SEX * AGE1 + REG1 + NAT1 * EDU1 (NC-C-3c)

SEX * AGE1 + REG1 + NAT1 * EDU1 (NC-C-3d)

où AGE1 et AGE1 sont des regroupements de AGE1 et où NAT1 est un regroupement de NAT1 (voir par. 2.1).

Une simplification éventuellement nécessaire du modèle NC-E-3 peut être obtenue selon le même raisonnement, à condition que STAT1 reste inchangé, que la simplification ait lieu entre les parenthèses et que, par conséquent, STAT1 apparaisse toujours au croisement avec chaque terme retenu entre les parenthèses :

(SEX * REG1 * EDU1 * AGE1 * NAT1) * STAT1 (NC-E-3a)

(SEX + REG1 + EDU1 + AGE1 + NAT1) * STAT1 (NC-E-3b)

(SEX * AGE1 + REG1 + NAT1 * EDU1) * STAT1 (NC-E-3c)

(SEX * AGE1 + REG1 + NAT1 * EDU1) * STAT1 (NC-E-3d)

Statbel a développé les macros SAS® nécessaires afin de pouvoir appliquer facilement, avec CALMAR2, de tels modèles ; voir annexe B.6.

Le choix du modèle NC-C ou NC-E dépendra en fin de compte du choix final du modèle LS : voir par. 2.6.

Enfin, nous remarquons que tous les modèles NC sont formulés en termes de variables contextuelles et de statut BIT au BQ, ce qui résulte du fait que l'échantillon BQ est redressé et que, dès lors, les poids de calage w_i^{BQ} des répondants de l'échantillon BQ sont donc corrigés.

2.5.3 Totaux de calages négatifs éventuels et choix de la méthode de calage

2.5.3.1 Modèles NC-C

Les modèles NC-C entraînent (normalement) toujours des totaux de calage positifs. Ces totaux reflètent en effet la distribution des variables contextuelles – marginales et/ou conjointes – reprises dans le modèle de calage, et sont (dans cette analyse) les sommes des poids de calage positifs w_i^{EQ} pour les répondants i dans l'échantillon EQ.[11]

Les variantes NC-C-1 jusque 3 et NC-C-3a des modèles NC-C sont, comme indiqué précédemment, de type "post-stratification". Cela signifie que le choix de la méthode de calage (linéaire, exponentielle, …) n'a pas d'impact sur la solution des facteurs de correction dans le système d'équations de calage, et que ces modèles sont entièrement déterminés par leur structure linéaire. Avec l'exigence que pour les répondants de l'échantillon à redresser (dans ce cas dans l'échantillon BQ de répondants) avec les mêmes valeurs pour toutes les variables de calage – c.-à-d. une "cellule" dans le croisement complet des variables de calage – on obtient un même facteur de correction, ce facteur de correction peut être calculé séparément pour chaque cellule, à savoir comme le rapport entre le total de calage correspondant et la somme des poids initiaux des unités de la cellule dans l'échantillon à redresser. (Par ailleurs, aucun logiciel sophistiqué n'est nécessaire pour cela).

Pour les variantes des modèles NC-C qui ne sont pas du type post-stratification, p.ex. NC-C-3b jusque d, une méthode de calage doit être choisie. Nous ne nous étendons pas sur ce point, car ce problème est bien connu dans la théorie du calage, et parce que nous choisissons finalement de ne pas utiliser un modèle NC-C, mais un modèle NC-E pour résoudre les problèmes d'incohérence numérique avant d'appliquer un modèle LS au LS.

[11] Dans un cas pratique rare, certains poids de calage w_i^{EQ} peuvent être nuls. Des totaux de calage nuls sont donc possibles. Nous ignorons cette possibilité dans notre analyse.

2.5.3.2 Modèles NC-E

En annexe C, au sous-paragraphe C.3, nous justifions le choix de Statbel d'une méthode linéaire pour l'application de modèles NC-E, étant donné qu'ils peuvent entraîner des totaux de calage négatifs. Nous illustrons cela au sous-paragraphe 2.5.4 suivant.

Cela implique que même pour les modèles NC-E du type post-stratification, comme NC-E-1 jusque 3 et NC-E-3a, la méthode de calage doit être choisie avec soin (p.ex. pour pouvoir appliquer CALMAR2). En effet, des totaux de calage négatifs ne peuvent être utilisés que si un ou plusieurs facteur(s) de correction négatif(s) résulte(nt) de l'application d'un modèle NC-E. Cela exclut, p.ex., la méthode exponentielle, car elle n'autorise pas de facteurs de correction négatifs ; la méthode logit pourrait être utilisée, à condition d'appliquer une limite inférieure négative pour les facteurs de correction ; etc. Finalement, nous choisissons toujours la méthode linéaire lorsque nous appliquons des modèles NC-E, donc aussi pour des modèles comme NC-E-3b jusque d qui ne sont pas du type post-stratification. La méthode linéaire est toujours applicable lorsque des totaux de calage sont négatifs (et dans ce cas, il est garanti que l'on trouvera également des facteurs de correction négatifs). Notez que même si tous les totaux de calage sont positifs, la méthode linéaire peut donner lieu à des facteurs de correction négatifs. Dans le contexte de l'obtention d'une cohérence numérique à l'aide de modèles NC, cela ne pose aucun problème, car ces résultats sous-jacents ne doivent pas être interprétés et publiés, ce qui est par contre le cas des résultats – les matrices de transition – de l'application des modèles LS.

Comme dans la section précédente, il suffit donc de représenter les modèles NC-E par leur structure linéaire, en raison du choix implicite de la méthode de calage linéaire.

2.5.4 Exemple : modèle NC-E avec totaux de calage négatifs

Imaginons que, pour la paire de trimestres 2018T3-2018T4, nous voulions appliquer le modèle de post-stratification NC-E-3a afin d'obtenir une cohérence numérique entre BQ et EQ. Le Tableau 10 montre l'ensemble des 13 cellules $sreanb$ dans le croisement[12] des six variables de calage SEX, REG1, EDU1, AGE1, NAT1 et STAT1 (indexées avec, respectivement, s, r, e, a, n et b), pour lesquelles un total de calage négatif $\tilde{T}^{BQ}_{sreanb}$ est obtenu (voir annexe C.1 pour les notations). Naturellement, cela n'est possible que pour $b = 3$ (inactifs). Pour la première ligne du Tableau 10, s'applique ce qui suit :

$$\begin{aligned}\tilde{T}^{BQ}_{srean3} &= T^{BQ}_{srean3} + \left(T^{EQ}_{srean} - T^{BQ}_{srean}\right)\\ &= 2307{,}35 + (12355{,}88 - 16446{,}84)\\ &= 2307{,}35 - 4090{,}96\\ &= -1783{,}61\end{aligned}$$

Le facteur de correction négatif en découle "manuellement" :

$$g_{srean3} = \tilde{T}^{BQ}_{srean3} \Big/ T^{BQ}_{srean3} = -1783{,}61 \big/ 2307{,}35 = -0{,}77301$$

avec lequel les poids de calage w_i^{BQ}, pour tous $i \in srean3$, sont multipliés. Les mêmes facteurs de correction négatifs (pour un total de 94 répondants dans les 13 cellules $srean3$) sont également obtenus en appliquant CALMAR2, avec la méthode linéaire.

[12] Le croisement complet de SEX, REG1, EDU1, AGE1, NAT1 et STAT1 contient au maximum 648 cellules non vides. Dans la pratique, pour la paire de trimestres 2018T3-2018T4, il n'y a que 585 cellules non vides. Parfois, c'est (plutôt) structurel : dans la tranche d'âge 65-74 ans ($a = 6$), il n'y a généralement pas de chômeurs ($b = 1$) ; parfois, c'est une coïncidence (due à des échantillons relativement petits), p.ex. la cellule (1,1,2,15-24,2,1) est vide : il n'y a pas de répondants masculins sans emploi, moyennement qualifiés, âgés de 15 à 24 ans et non belges à Bruxelles.

Tableau 10 Totaux de calage négatifs sous le modèle NC-E-3, entraînant des facteurs de correction négatifs

SEX (s)	REG1 (r)	EDU1 (e)	AGE1 (a)	NAT1 (n)	STAT1 (b)	T^{BQ}_{srean3}	T^{BQ}_{srean}	T^{EQ}_{srean}	$\tilde{T}^{BQ}_{srean3}$	g_{srean3}
1	1	2	35-44	1	3	2307,35	16446,84	12355,88	-1783,61	-0,77301
1	1	3	35-44	1	3	966,19	22975,53	20015,91	-1993,42	-2,06317
1	2	3	35-44	1	3	4451,36	167041,99	159310,30	-3280,33	-0,73693
1	2	3	45-54	2	3	1139,72	16865,93	15383,81	-342,40	-0,30042
1	2	3	55-64	2	3	514,89	8689,63	5580,80	-2593,93	-5,03780
1	3	1	25-34	2	3	1134,26	5865,31	3257,67	-1473,38	-1,29898
1	3	2	45-54	2	3	1703,68	16738,28	9267,72	-5766,89	-3,38496
1	3	3	25-34	2	3	1452,33	9950,56	8455,55	-42,68	-0,02939
1	3	3	45-54	1	3	2974,05	77765,55	72537,33	-2254,17	-0,75795
2	1	2	45-54	1	3	1460,67	15947,48	13249,76	-1237,05	-0,84691
2	1	2	45-54	2	3	604,60	5927,60	4395,32	-927,68	-1,53438
2	1	3	35-44	1	3	1876,05	24685,11	20755,26	-2053,80	-1,09475
2	2	3	35-44	1	3	6220,97	210030,97	197009,83	-6800,18	-1,09311

Bien que NC-E-3a soit un modèle de post-stratification, en raison des totaux de calage négatifs, la méthode de calage dans CALMAR2 ne peut pas être choisie arbitrairement. Par exemple, si nous choisissons la méthode du *raking ratio*, CALMAR2 converge effectivement, mais les équations de calage ne sont pas satisfaites pour les 13 cellules du Tableau 10 : pour ces cellules, CALMAR2 finit par rendre le facteur de correction g_{srean3} nul (pour toutes les autres cellules, le facteur de correction est correct).

Si l'on applique le modèle NC-E-3b, c.-à-d. (SEX + REG1 + EDU1 + AGE1 + NAT1) * STAT1, pour 2018T3-2018T4, il n'y a pas de totaux de calage négatifs : les totaux négatifs $\tilde{T}^{BQ}_{srean3}$ pour les 13 cellules du Tableau 10 sont alors repris dans les totaux positifs $\tilde{T}^{BQ}_{s3}$, $\tilde{T}^{BQ}_{r3}$, $\tilde{T}^{BQ}_{e3}$, $\tilde{T}^{BQ}_{a3}$ et $\tilde{T}^{BQ}_{n3}$. CALMAR2 ne produit alors, même avec la méthode linéaire, aucun facteur de correction négatif g_{srean3}. Si nous utilisons la méthode du *raking ratio*, CALMAR2 converge toujours ; les facteurs de correction pour $b = 3$ diffèrent relativement peu de ceux qui résultent de la méthode linéaire ; pour $b = 1$ et 2 tous les g_{sreanb} sont exactement égaux à 1, comme attendu pour les deux méthodes.

Rien n'exclut que pour d'autres paires de trimestres, certains totaux de calage pour un modèle NC-E, qui n'est pas de type post-stratification, soient tout de même négatifs. C'est pourquoi nous optons toujours pour la méthode linéaire lors de l'application d'un modèle NC-E.

2.6 Modèles de calage de base, avec ventilation selon plusieurs variables contextuelles

2.6.1 Etat d'avancement

Au par. 2.4, nous avons montré comment les modèles de calage peuvent être construits en tant que matrices de transition uniquement pour la population belge totale et par sexe doivent être cohérents avec les chiffres trimestriels du statut BIT : cela a conduit au modèle LS-2 pour le calage des LS. Au par. 2.5, nous avons abordé le problème sous-jacent de l'incohérence entre les chiffres du BQ et du EQ pour le statut BIT. La conclusion de ces deux paragraphes est la suivante :

- si (LS-1) STAT1 + STAT2 est appliqué dans le but d'estimer une matrice de transition globale, dont les marges sont cohérentes avec la distribution du statut BIT au BQ et au EQ, alors les échantillons redressés du BQ et de l'EQ peuvent être rendus cohérents via un modèle (NC-C-1) 1 ou (NC-E-1) STAT1 (au minimum) ;
- si (LS-2) SEX*(STAT1 + STAT2) est appliqué dans le but d'estimer des matrices de transition par sexe, dont les marges sont cohérentes avec les distributions du statut BIT au BQ et au EQ, alors les échantillons redressés du BQ et du EQ peuvent être rendus cohérents via un modèle (NC-C-2) SEX ou NC-E-2) SEX*STAT1 (au minimum).

Il convient de noter qu'il est également possible de combiner LS-1 avec un modèle (non minimal) NC-C-2 ou NC-E-2 : il y aura alors plus de cohérence numérique entre les échantillons trimestriels que ce qui n'est nécessaire (au minimum) pour rendre cohérente la matrice de transition globale avec la distribution du statut BIT du BQ et du EQ. La cohérence numérique obtenue

entre les échantillons trimestriels n'est alors pas pleinement exploitée, et la ventilation de la matrice de transition par sexe ne reflète pas la cohérence numérique par sexe entre les échantillons trimestriels en ce sens que les distributions d'un sexe spécifique du statut BIT ne sont pas reproduites. Nous pouvons affirmer que les modèles NC-C-2 et NC-E-2 sont *surdéterminés* pour appliquer LS-1.

Cependant, il n'est généralement pas possible de combiner le LS-2 avec NC-C-1 ou NC-E-1 : la cohérence numérique globale obtenue entre les échantillons trimestriels ne suffit pas pour appliquer LS-2 avec succès.

Une conclusion supplémentaire qui découle du par. 2.4 est que l'utilisation des modèles NC-E a des avantages par rapport à l'utilisation des modèles NC-C. C'est pourquoi, dans ce qui suit, nous ne travaillerons qu'avec des modèles NC-E. Les difficultés techniques spécifiques qui peuvent être rencontrées avec ces modèles sont abordées au par. 2.5.3.

2.6.2 Extension des objectifs et des modèles

Statbel s'est fixé comme objectif de publier des matrices de transition pour divers groupes de la population à partir de 2021 : pas seulement pour la population belge totale, mais également par sexe, par région, par niveau d'instruction, par tranche d'âge et par groupe de nationalité. Il est souhaitable que les marges de chaque matrice de transition publiée soient cohérentes avec les chiffres trimestriels du statut BIT publiés précédemment.

Pour atteindre cet objectif, la combinaison suivante de modèles de calage peut être appliquée :

$$(SEX + REGt + EDUt + AGEt + \underline{NATt}) * (STAT1 + STAT2) \qquad \text{(LS-3)}$$

$$(SEX * REG1 * EDU1 * AGE1 * \underline{NAT1}) * STAT1 \qquad \text{(NC-E-3a)}$$

Ici, REGt représente tant REG1 que REG2 (c.-à-d. la région de résidence du répondant au BQ et au EQ), parce que la résidence peut changer entre le BQ et le EQ. Idem pour EDUt, AGEt et NATt ; voir par. 2.1. La formulation pour LS-3 est une manière succincte de refléter la structure linéaire suivante :

$$(SEX + REG1 + EDU1 + AGE1 + \underline{NAT1}) * STAT1 + (SEX + REG2 + EDU2 + AGE2 + \underline{NAT2}) * STAT2$$

parce que naturellement les versions REG1, EDU1, AGE1 et NAT1 des variables contextuelles du BQ sont combinées à la version STAT1 du statut BIT du BQ et les versions REG2, EDU2, AGE2 et NAT2 des variables contextuelles du EQ à la version STAT2 pour le statut BIT au EQ. Notons que – du moins dans la pratique de Statbel – jusqu'à présent, SEX ne change de valeur pour aucun répondant entre le BQ et le EQ.

Le terme REG1*STAT1 dans le LS-3 implique que, pour chaque région, la matrice de transition est marginalement cohérente avec la distribution estimée corrigée du statut BIT du BQ obtenue via NC-E-3a, et le terme REG2*STAT2 dans le LS-3 implique que pour chaque région, la matrice de transition est marginalement cohérente avec la distribution estimée (inchangée) du statut BIT du EQ. Il en va de même pour les autres variables contextuelles.[13] Il s'agit clairement d'une extension du modèle LS-2. Le modèle LS-3 permet donc d'atteindre l'objectif. Le modèle NC-E-3a permet d'atteindre la cohérence numérique requise à cette fin entre les distributions estimées du statut BIT du BQ et du EQ.

Le Tableau 11 présente la matrice de transition par sexe après application des modèles de calage LS-3 et NC-E-3a. Ces résultats sont comparables à ceux du deuxième volet du Tableau 6, qui ont été obtenus après application des modèles de calage (LS-2) SEX*(STAT1 + STAT2) et (NC-E-2) SEX*STAT1. Les marges des matrices de transition sont, à l'exception des erreurs d'arrondi, identiques dans les deux tableaux. Les chiffres des transitions mêmes diffèrent, ce qui est dû à la différence entre les modèles LS-3 et LS-2. La plus grande différence absolue (en valeur absolue) est de 12.732,95 pour les hommes qui sont occupés tant au 2018T3 qu'au 2018T4. La plus grande différence relative (en valeur absolue) est de 11,29 % pour les hommes qui sont occupés au 2018T3 et inactifs au 2018T4. La plus grande différence entre les taux de transition estimés (en valeur absolue) s'élève à 1,9 point de pourcentage, pour les hommes qui étaient au chômage au 2018T3 et occupés au 2018T4.

Tableau 11 Estimation des transitions d'un sexe spécifique 2018T3-2018T4 avec le modèle LS-3, après corrections de l'incohérence numérique avec le modèle NC-E-3a

Hommes

[13] Dans certains cas, une petite perturbation du LS est nécessaire pour pouvoir appliquer un modèle LS, et donc pour obtenir les cohérences souhaitées. Ce point est discuté et illustré plus en détail dans l'annexe D.

Statut BIT 2018T3	Statut BIT 2018T4				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	89.572,22	47.673,26	33.364,51	170.610,00	*170.610,00*	*171.013,58*	*170.610,00*
Occupé	27.518,78	2.401.933,68	100.680,43	2.530.132,89	*2.530.132,88*	*2.536.117,96*	*2.530.132,88*
Inactif	48.426,27	97.909,88	1.357.694,98	1.504.031,13	*1.494.108,13*	*1.497.642,47*	*1.504.031,13*
Total	165.517,27	2.547.516,82	1.491.739,93	**4.204.774,02**	***4.194.851,00***	***4.204.774,00***	***4.204.774,00***
Distribution réf. 2018T4	*165.517,27*	*2.547.516,81*	*1.491.739,92*	***4.204.774,00***			

Femmes

Statut BIT 2018T3	Statut BIT 2018Q4				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	59.955,47	30.714,59	38.935,82	129.605,88	*129.605,88*	*129.909,54*	*129.605,88*
Occupé	20.804,83	2.134.385,85	99.925,36	2.255.116,03	*2.255.116,03*	*2.260.399,59*	*2.255.116,03*
Inactif	44.125,10	92.115,72	1.704.586,29	1.840.827,10	*1.830.950,09*	*1.835.239,87*	*1.840.827,09*
Total	124.885,40	2.257.216,16	1.843.447,47	**4.225.549,02**	***4.215.672,00***	***4.225.549,00***	***4.225.549,00***
Distribution réf. 2018T4	*124.885,40*	*2.257.216,15*	*1.843.447,46*	***4.225.549,00***			

Le modèle NC-E-3a (si applicable) garantit la cohérence des totaux de calage du modèle LS-3, de sorte que LS-3 soit applicable. Le modèle de post-stratification NC-E-3a peut être appliqué sans logiciel sophistiqué, mais il est en quelque sorte surdéterminé pour pouvoir appliquer LS-3. En effet, il ne garantit pas seulement la cohérence numérique entre les distributions marginales du statut BIT au BQ et au EQ par sexe, par région, etc. séparément, mais également pour chaque combinaison possible de cinq variables contextuelles. Une cohérence numérique aussi détaillée n'est nécessaire que si nous souhaitons appliquer la combinaison de modèles suivante :

(SEX * REG1 * EDU1 * AGE1 * NAT1) * STAT1 (NC-E-3a)

(SEX * REGt * EDUt * AGEt * NATt) * (STAT1 + STAT2) (LS-3a)

ce qui serait nécessaire si nous voulions produire une matrice de transition par combinaison de SEX, REG, EDU, AGE et NAT dont les marges soient cohérentes avec les chiffres trimestriels publiés précédemment pour le statut BIT. Cependant, pour le couple 2018T3-2018T4, le modèle LS-3a n'est pas applicable par nature parce que le LS est trop restreint ; en particulier :

- pour le croisement SEX * REG1 * EDU1 * AGE1 * NAT1 * STAT1, il y a 37 cellules vides dans l'échantillon LS, mais pas dans l'échantillon BQ ;
- pour le croisement SEX * REG2 * EDU2 * AGE2 * NAT2 * STAT2, il y a 45 cellules vides dans l'échantillon LS, mais pas dans l'échantillon EQ.

La combinaison de modèles suivante est suffisante pour atteindre l'objectif fixé initialement :

(SEX + REG1 + EDU1 + AGE1 + NAT1) * STAT1 (NC-E-3b)

(SEX + REGt + EDUt + AGEt + NATt) * (STAT1 + STAT2) (LS-3)

parce que pour le modèle LS-3, NC-E-3b est le modèle minimum pour atteindre la NC entre le BQ et le EQ. Cette combinaison de modèles conduit également au Tableau 11, ce qui montre par conséquent que le modèle NC-E-3a est surdéterminé pour le LS-3.

Nous ajouterons un terme supplémentaire au LS-3, afin d'arriver finalement à une combinaison finale de modèles qui sera appliquée pour obtenir les matrices de transition publiées. Ceci sera traité au par. 2.7 suivant.

2.7 Modèle de calage final, avec ajout de terme(s) structurel(s)

Bien que le modèle LS-3 atteigne pleinement les objectifs, nous ajoutons un autre terme SEX*AGE2*REG2, pour arriver au modèle LS-4 :

(SEX * REG1 * EDU1 * AGE1 * NAT1) * STAT1 (NC-E-3a)

SEX*AGE2*REG2 + (SEX + REGt + EDUt + AGEt + NATt) * (STAT1 + STAT2) (LS-4)

L'ajout du terme SEX*AGE2*REG2 ne change pas le choix du modèle NC-E (nous expliquons ci-dessous pourquoi le modèle NC-E-3a, et non NC-E-3b, a finalement été choisi). C'est finalement cette combinaison de modèles qui est appliquée par Statbel pour produire les matrices de transition.

Le Tableau 12 présente la matrice de transition par sexe après application des modèles de calage LS-4 et NC-E-3a.[14] Ces résultats sont comparables à ceux du Tableau 11, qui ont été obtenus après application des modèles de calage LS-3 et NC-E-3a. Les marges des matrices de transition sont, aux erreurs d'arrondi près, identiques dans les deux tableaux. Les chiffres des transitions eux-mêmes diffèrent relativement peu : la plus grande différence absolue (en valeur absolue) est de 932,35 pour les hommes qui sont inactifs tant au 2018T3 qu'au 2018T4. La plus grande différence relative (en valeur absolue) est de 1,25 % pour les hommes qui sont chômeurs au 2018T3 et inactifs au 2018T4. La plus grande différence entre les probabilités de transition estimées (en valeur absolue) s'élève à 0,3 point de pourcentage, pour les hommes qui étaient au chômage tant au 2018T3 qu'au 2018T4. Le terme SEX*AGE2*REG2 ajouté dans le LS-4 par rapport au LS-3 n'a donc qu'un effet limité sur les matrices de transition par sexe.

Tableau 12 Estimation des transitions d'un sexe spécifique 2018T3-2018T4 avec le modèle LS-4, après corrections de l'incohérence numérique avec le modèle NC-E-3a

Hommes							
Statut BIT 2018T3	**Statut BIT 2018T4**				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	90.155,76	47.506,15	32.948,09	170.610,00	*170.610,00*	*171.013,58*	*170.610,00*
Occupé	27.368,15	2.402.600,26	100.164,51	2.530.132,92	*2.530.132,88*	*2.536.117,96*	*2.530.132,88*
Inactif	47.993,36	97.410,44	1.358.627,33	1.504.031,14	*1.494.108,13*	*1.497.642,47*	*1.504.031,13*
Total	165.517,27	2.547.516,85	1.491.739,93	**4.204.774,05**	***4.194.851,00***	***4.204.774,00***	***4.204.774,00***
Distribution réf. 2018T4	*165.517,27*	*2.547.516,81*	*1.491.739,92*	***4.204.774,00***			
Femmes							
Statut BIT 2018T3	**Statut BIT 2018T4**				Distribution réf. 2018T3		
	Chômeur	Occupé	Inactif	Total	(a)	(b)	(c)
Chômeur	60.299,26	30.824,17	38.482,45	129.605,89	*129.605,88*	*129.909,54*	*129.605,88*
Occupé	20.841,99	2.133.679,02	100.595,03	2.255.116,05	*2.255.116,03*	*2.260.399,59*	*2.255.116,03*
Inactif	43.744,14	92.712,98	1.704.370,01	1.840.827,13	*1.830.950,09*	*1.835.239,87*	*1.840.827,09*
Total	124.885,40	2.257.216,17	1.843.447,49	**4.225.549,06**	***4.215.672,00***	***4.225.549,00***	***4.225.549,00***
Distribution réf. 2018T4	*124.885,40*	*2.257.216,15*	*1.843.447,46*	***4.225.549,00***			

Pourquoi alors ce terme supplémentaire ? La signification du terme SEX*AGE2*REG2 est que la composition ou la structure selon les variables SEX, AGE2 et REG2 de la population des 15-74 ans, telle qu'estimée au EQ, est introduite dans le LS redressé. Bien sûr, le terme (SEX + REG2 + EDU2 + AGE2 + NAT2) * STAT2, le fait déjà dans une certaine mesure, puisqu'il implique, entre autres, les termes SEX, AGE2 et REG2, ce qui introduit les distributions marginales de ces variables dans le LS. Le terme SEX*AGE2*REG2 y ajoute encore les distributions jointes (deux par deux, et pour les trois variables ensemble).

[14] Comme le modèle LS-3, le modèle LS-4 (pour 2018T3-2018T4) nécessite une perturbation du LS. Voir note de bas de page 13 et annexe D.

Le terme supplémentaire – que nous appellerons *terme structurel* – s'inspire des techniques décrites par Eurostat (2015b) ; nous y reviendrons au par. 2.8.

Après application de n'importe quelle combinaison de modèles, de nombreuses ventilations des matrices de transition peuvent être effectuées. Par exemple, une ventilation par province (niveau NUTS2 ; variable PROV2 par analogie à REG2) peut être effectuée même si seule la région (niveau NUTS1 ; variable REG2) figure dans la combinaison de modèles (par exemple, LS-4 avec NC-E-3a), ou même si la région ne figure pas dans la combinaison de modèles (par exemple, LS-2 avec NC-E-2). Autre exemple : après avoir appliqué la combinaison de modèles LS-4 avec NC-E-3a – le modèle final actuellement utilisé par Statbel pour les publications – des matrices de transition peuvent être créées pour toute combinaison de SEX et AGE2. Les marges de ces matrices de transition ne seront en général pas cohérentes avec les chiffres trimestriels, mais le terme structurel peut rendre les marges plus cohérentes avec les chiffres trimestriels. La manière idéale de rendre les marges dans les cellules de croisement SEX×AGE2 cohérentes avec les chiffres trimestriels est d'étendre le modèle LS-4 à SEX*AGE2*REG2 + (SEX + REGt + EDUt + AGEt + NATt + **SEX*AGEt**) * (STAT1 + STAT2) ; cependant, de telles extensions peuvent engendrer des problèmes dus à un LS trop réduit, comme l'illustre l'échec de l'application de la combinaison de modèles LS-3a avec NC-E-3a.

Bien entendu, SEX*AGE2*REG2 n'est pas le seul terme structurel potentiellement utile ; d'autres exemples sont SEX*AGE2, SEX*REG2, AGE2*REG2, REG2*NAT2, PROV2, SEX*PROV2, PROV2*AGE2*NAT2, etc. Le choix final de Statbel pour le terme structurel SEX*AGE2*REG2 est plutôt arbitraire ; une étude plus approfondie pourrait éventuellement conduire à des termes structurels “plus optimaux”.

Un autre effet possible d'un terme structurel est la réduction de la variance, c.-à-d. une augmentation de la précision des taux de transition estimés. Cet aspect mérite d'être étudié plus en profondeur, et est lié à la détermination de termes structurels “optimaux”, et par conséquent des modèles LS “optimaux” pour la production des matrices de transition.

Enfin, nous notons que Statbel n'a pas choisi, avec le LS-4, le modèle NC-E-3b, plus simple mais suffisant, mais plutôt le modèle NC-E-3a, plus détaillé, pour estimer et publier les matrices de transition, car au moment où les matrices de transition ont été produites et publiées, le logiciel (en SAS®, avec CALMAR2 et des macros génériques de Statbel) pour appliquer des modèles tels que NC-E-3b n'était pas encore au point. Les matrices de transition du Tableau 12 sont disponibles sur le site web de Statbel dans le fichier Excel téléchargeable EFT_TRANSITION_FR_QQ_P.xlsx, dans la feuille de calcul 2018T3-T4, lignes 16 à 28. Le même fichier Excel contient de nombreuses autres matrices de transition : d’autres paires de trimestres et d’autres ventilations. Un fichier Excel analogue EFT_TRANSITION_FR_JQ_P.xlsx contient les transitions annuelles d’un trimestre spécifique, qui sont aussi estimées avec la combinaison de modèles NC-E-3a et LS-4.

2.8 Approche par étapes d'Eurostat, et comparaison avec la méthode de Statbel

Dans ce paragraphe, nous abordons la méthode d'Eurostat (Eurostat, 2015b). Nous le faisons de manière concise en utilisant la terminologie utilisée ci-dessus, afin de faire apparaître les similitudes/différences avec le modèle final de Statbel.

En utilisant la notation de cette analyse (voir annexe C.1 et par. 2.5.4), Eurostat (2015b) utilise les variables contextuelles

- SEX (“sex”), dont nous indiquons les classes avec l'indice s, et
- AGE2 (“10-year age group”, avec les classes 15-24, 25-34, ... 65-74), dont les classes sont désignées par l'indice $\bar{a}$ (nous avons précédemment utilisé l'indice a pour les classes de la variable AGE1),

et les variables d'étude

- STAT1 (statut BIT au BQ), dont nous indiquons les classes avec l'indice b, et
- STAT1 (statut BIT au EQ), dont nous indiquons les classes avec l'indice $\bar{b}$.

Pour une paire de trimestres donnée, toutes les personnes i dans le LS ont un poids initial w_i^{EQ}. A partir de cela, Eurostat (2015b) calcule pour chacune des 12 combinaisons $s\bar{a}$ de SEX et AGE2 une matrice de transition initiale. Chacune de ces matrices de transition est corrigée sur la distribution estimée du statut BIT au EQ (STAT2) pour la sous-population $s\bar{a}$. Dans notre approche basée sur les modèles, la première correction correspond à un calage de l’ensemble du LS selon le modèle de post-stratification

(SEX * AGE2) * STAT2 (LS-0a)

Nous pouvons noter les facteurs de correction comme $c_{s\bar{a}\bar{b}}$; le résultat de cette première correction est un nouveau poids $c_{s\bar{a}\bar{b}}w_i^{EQ}$ pour chaque personne i dans le LS. Il en résulte une matrice de transition adaptée pour chaque combinaison $s\bar{a}$, basée sur les nouveaux poids $c_{s\bar{a}\bar{b}}w_i^{EQ}$.

- Une deuxième correction du LS est préparée par Eurostat (2015b) comme suit :
- Les matrices de transition adaptées sont sommées pour chaque sexe s sur les classes d'âge $\bar{a}$, ce qui donne des matrices de transition d'un sexe spécifique adaptées.
- Pour chaque sexe s et pour l'ensemble de la classe d'âge 15-74, la distribution estimée du statut BIT au BQ (STAT1) est adaptée afin d'être cohérente avec la distribution du statut BIT au EQ (STAT2). Cela est réalisé en adaptant (par sexe s) le nombre estimé d'inactifs au BQ, de sorte que (par sexe s) les sommes des nombres estimés de personnes occupées, de chômeurs et d'inactifs soient égales au BQ et au EQ. Dans notre approche basée sur des modèles, cette adaptation correspond à un calage de l'échantillon BQ des 15-74 ans, qui ont tous un poids w_i^{BQ} selon le modèle NC-E

SEX * STAT1 (NC-E-0)

On obtient ainsi un nouveau poids $g_{sb}w_i^{BQ}$ pour chaque personne i de l'échantillon BQ. Remarquez que les facteurs de correction g_{sb} sont égaux à 1 pour $b = 1$ (personnes occupées) et $b = 2$ (chômeurs). Remarquez également que le modèle NC-E-0 est le même que le modèle NC-E-2 (voir par. 2.5.2)

Ensuite, Eurostat (2015b) utilise la méthode IPF pour corriger les deux matrices de transition d'un sexe spécifique ajustées aux distributions correspondantes (par sexe s) du statut BIT au BQ et EQ. Dans notre approche basée sur les modèles, la deuxième correction correspond à un calage de l'ensemble du LS selon le modèle

SEX * (STAT1 + STAT2) (LS-0b)

Nous notons les facteurs de correction comme $\acute{c}_{sb\bar{b}}$; le résultat de cette deuxième correction est le poids final $\acute{c}_{sb\bar{b}}c_{s\bar{a}\bar{b}}w_i^{EQ}$ pour chaque personne i dans le LS. Remarquez que le modèle LS-0b est le même que le modèle LS-2 (voir par. 2.4.2)

Sachant que dans la méthode d'Eurostat, le traitement de l'échantillon BQ selon le modèle NC-E-0 peut aisément précéder les deux corrections successives des matrices de transition, nous pouvons succinctement conclure que dans la méthode de Statbel, la méthode d'Eurostat correspond à

- l'application du modèle NC-E-0 (ou NC-E-2) à l'échantillon BQ, avec les poids initiaux w_i^{BQ} ;
- puis appliquer le modèle LS-0a au LS, avec les poids initiaux w_i^{EQ}, ce qui donne les poids $c_{s\bar{a}\bar{b}}w_i^{EQ}$;
- puis appliquer le modèle LS-0b (ou LS-2) au LS, avec les nouveaux poids initiaux $c_{s\bar{a}\bar{b}}w_i^{EQ}$, ce qui donne les poids finaux $\acute{c}_{sb\bar{b}}c_{s\bar{a}\bar{b}}w_i^{EQ}$.

Le résultat est l'échantillon LS avec les poids de calage $\acute{c}_{sb\bar{b}}c_{s\bar{a}\bar{b}}w_i^{EQ}$, avec lequel les matrices de transition finales peuvent être calculées.

A noter que les calculs pour l'application des modèles NC-E-0, LS-0a et LS-0b ne nécessitent pas de logiciel sophistiqué : NC-E-0 et LS-0a sont des modèles de post-stratification pour lesquels un facteur de correction par cellule peut être calculé au croisement des variables concernées, et LS-0b peut être résolu avec IPF (comme d'habitude par valeur de SEX dans deux dimensions conformément aux termes STAT1 et STAT2, ou simultanément pour les hommes et les femmes dans deux dimensions correspondant aux termes SEX*STAT1 et SEX*STAT2).

De cette façon, le LS est redressé en deux étapes. Il n'y a en fait aucune raison à cela, de sorte qu'une alternative valable à la méthode d'Eurostat peut être formulée comme suit :

- l'application du modèle NC-E-0 (ou NC-E-2) à l'échantillon BQ, avec les poids initiaux w_i^{BQ} ;
- puis en appliquant le modèle LS-0c au LS, avec les poids initiaux w_i^{EQ}, où LC-0c atteint les objectifs de LS-0a et LS-0b simultanément :

(SEX * AGE2) * STAT2 + SEX * (STAT1 + STAT2) (LS-0c)

Le modèle LS-0c transforme les poids initiaux w_i^{EQ} en poids de calage $\bar{c}_{s\bar{a}b\bar{b}}w_i^{EQ}$, par exemple.

Un inconvénient de l'application par étapes des modèles LS-0a et LS-0b est que la cohérence obtenue avec LS-0a est généralement annihilée par l'application de LS-0b. Cela signifie qu'avec les facteurs de correction $\acute{c}_{sb\bar{b}}c_{s\bar{a}\bar{b}}$, les équations de calage du modèle LS-0a ne sont généralement pas satisfaites. L'avantage du modèle LS-0c, avec les facteurs de correction $\bar{c}_{s\bar{a}b\bar{b}}$, est que toutes les équations de calage, tant celles qui découlent du modèle LS-0a que celles qui découlent du modèle LS-0b, sont satisfaites simultanément. Les facteurs de correction $\acute{c}_{sb\bar{b}}c_{s\bar{a}\bar{b}}$ et $\bar{c}_{s\bar{a}b\bar{b}}$ ne sont donc généralement pas égaux.

Un inconvénient technique du modèle LS-0c est qu'aucune technique de post-stratification ne peut être utilisée pour appliquer ce modèle. En revanche, la méthode IPF peut en principe être appliquée, mais pas (comme d'habitude) en deux dimensions, mais en trois dimensions – conformément aux trois termes SEX*AGE2*STAT2, SEX*STAT1 et SEX*STAT2 – ce qui nécessite une mise en œuvre plus poussée de la méthode IPF. Cet inconvénient peut être facilement contourné en utilisant une macro générique telle que CALMAR2, qui est écrite en code SAS®. Il n'est pas possible de choisir l'algorithme IPF dans CALMAR2 : tous les modèles de calage qui peuvent être appliqués par CALMAR2 sont résolus par un algorithme numérique très universel, basé sur la méthode Newton-Raphson.

Pour appliquer le modèle LS-0c via CALMAR2, il faut choisir la méthode de calage. Étant donné que l'algorithme IPF conduit à la même solution de modèles de calage que l'algorithme universel dans CALMAR2 lorsque la méthode exponentielle (c.-à-d. une fonction de calage multiplicative) est choisie, il est évident d'utiliser la méthode exponentielle. Cela a motivé la décision de résoudre tous les modèles LS dans cette analyse en utilisant la méthode exponentielle. Par ce choix, nous nous conformons à la méthode d'Eurostat.

Enfin, nous pouvons comparer LS-0c avec le modèle final LS-4 de Statbel. Dans un premier temps, nous ignorons les variables de LS-4 qui n'apparaissent pas dans le modèle LS-0c, c.-à-d. que nous réduisons LS-4 au modèle plus simple suivant :

SEX*AGE2 + (SEX + AGEt) * (STAT1 + STAT2) (LS-4a)

Pour comprendre les similitudes et les différences entre LS-4a et LS-0c, nous réécrivons[15] LS-0c comme suit :

[SEX*AGE2 + SEX*AGE2*STAT2] + [AGE2*STAT2 + SEX*(STAT1 + STAT2)] (LS-0c)

ou comme :

[SEX*AGE2*STAT2] + [AGE2*STAT2 + SEX*(STAT1 + STAT2)] (LS-0c)

La différence entre la deuxième partie [AGE2*STAT2 + SEX*(STAT1 + STAT2)] de LS-0c et la deuxième partie (SEX + AGEt) * (STAT1 + STAT2) de LS-4a est un terme AGE1*STAT1. L'ajout de ce terme AGE1*STAT1 dans LS-0c élargirait donc l'objectif initial de la note explicative d'Eurostat de manière évidente : les matrices de transition seraient cohérentes avec les résultats du BQ et du EQ non seulement pour les hommes et les femmes, mais aussi par groupe d'âge.

La différence entre la première partie [SEX*AGE2 + SEX*AGE2*STAT2] de LS-0c et la première partie SEX*AGE2 de LS-4a est – formellement – le terme SEX*AGE2*STAT2, mais puisque SEX*AGE2 est impliqué par SEX*AGE2*STAT2, la différence réelle est constituée (1°) du terme STAT2, (2°) des interactions doubles SEX×STAT2 et AGE2×STAT2 et (3°) de l'interaction triple SEX×AGE2×STAT2. De plus, les termes STAT2, SEX×STAT2 et AGE2×STAT2 sont impliqués par la deuxième partie de LS-4a, ou par la deuxième partie étendue de LS-0c, telle que suggérée dans le paragraphe précédent.

Une méthode alternative pour comparer LS-4a et LS-0c par rapport à leurs premières parties respectives consiste à postuler que le terme structurel SEX*AGE2 de LS-4a est étendu au terme structurel SEX*AGE2*STAT2 de LS-0c. Cela semble être une "grande" extension, mais, si l'on tient également compte de la deuxième partie de LS-0c, nous constatons que certaines composantes de SEX*AGE2*STAT2 font déjà partie de la deuxième partie de l'extension suggérée de LS-0c : en effet, les termes STAT2, SEX×STAT2 et AGE2×STAT2 sont impliqués par les deux parties. En d'autres termes, il y a un "chevauchement" entre la première et la deuxième partie de (la version étendue de) LS-0c.

Notez que LS-4a, contrairement à LS-0c, est un peu plus facile à interpréter : la deuxième partie permet d'atteindre les objectifs de cohérence concernant les différentes distributions des variables d'étude STAT1 et STAT2, la première partie se concentre uniquement sur la structure de l'échantillon EQ redressé (c.-à-d. une structure estimée de la population) via les variables contextuelles. Et l'extension de LS-4a à LS-4 est donc un peu plus transparente : extension de la deuxième partie

[15] La réécriture ou la reformulation des structures linéaires des modèles de calage est basée sur la nature hiérarchique de ces structures : un terme tel que par exemple A*B*C (pour des variables catégorielles A, B et C) implique toujours les termes A*B, A*C, B*C, A, B, C et 1 ; voir annexe B.4.1.

pour obtenir plus de cohérence, extension de la première partie pour intégrer davantage de structure de la population (estimée) des 15-74 ans au EQ dans le LS redressé.

2.9 Estimation des transitions annuelles

Aux paragraphes 2.1 à 2.7, des modèles de calage ont été développés pour estimer les transitions trimestrielles et les transitions annuelles d'un trimestre spécifique. Les deux types de transitions peuvent être estimés à l'aide de la même méthodologie, puisqu'ils impliquent tous deux le calage d'un LS, qui est l'intersection de deux échantillons trimestriels. Sauf s'ils sont influencés par la phase de démarrage du panel, ces LS ont tous la même structure : ils sont composés de répondants de deux RG, pour chacun desquels exactement une transition est observée. On peut le déduire du Tableau B 1 pour les transitions trimestrielles et du Tableau B 2 pour les transitions annuelles d'un trimestre spécifique.

Si nous devions établir – en utilisant le schéma 1 de Termote & Depickere (2018) – un tableau similaire pour les transitions annuelles globales, en se basant sur des échantillons annuels de deux années consécutives (par exemple 2018 et 2019), et pour lesquels le LS est l'intersection de ces échantillons annuels, ce tableau se présenterait comme suit :

Année de départ	Année de fin	RG dans chevauchement	1er RG	2e RG	3e RG	4e RG	5e RG
			Observations des vagues...				
2018	2019	10, 11, 12, 13, 14	2 et 4	1 et 3 2 et 4	1 et 3 2 et 4	1 et 3 2 et 4	1 et 3

Le LS pour 2018-2019 impliquerait cinq RG (consécutifs), qui ne fournissent pas tous des données de la même manière : pour chaque répondant du premier et du cinquième RG, exactement une transition a été observée (entre W2 et W4, ou entre W1 et W3), mais pour chaque répondant des trois autres RG, deux transitions ont été observées (entre W1 et W3, ainsi qu'entre W2 et W4). Ces deux transitions, c.-à-d. ces deux "observations", pour le même répondant ne peuvent être considérées comme indépendantes, ce qui va à l'encontre de l'application courante des techniques de calage, qui supposent l'indépendance des "observations".

En raison de ce problème, les transitions annuelles sont simplement estimées comme une moyenne non pondérée des transitions annuelles estimées d'un trimestre spécifique.

Il convient de noter que les quatre transitions annuelles estimées d'un trimestre spécifique ne sont pas des statistiques indépendantes (par analogie à la non-indépendance des observations dans le LS pour l'estimation des transitions annuelles discutée ci-dessus). Cela ne pose pas de problème pour l'estimation des transitions annuelles en tant que moyenne. L'estimation de la variance pour les transitions annuelles devrait cependant être réalisée avec des techniques spécialement développées qui prennent en compte deux observations pour une grande partie des répondants, mais cela sort du cadre de cette analyse.

L'estimation des transitions annuelles en tant que moyenne non pondérée des transitions annuelles d'un trimestre spécifique est analogue à l'estimation des indicateurs clés annuels (tels que le nombre de chômeurs, le nombre de personnes occupées, etc.) en tant que moyennes non pondérées des indicateurs clés trimestriels. La méthode d'estimation de la variance pour les indicateurs clés annuels a également été adaptée au fait que la majorité des répondants contribuent à deux indicateurs clés trimestriels.

3 Chiffres publiés

A partir de 2021, Statbel publie chaque trimestre – c.-à-d. le trimestre *actuel* – les matrices de transition pour le trimestre actuel par rapport au trimestre précédent (soit les *transitions trimestrielles* les plus récentes, indiquées par QQ dans les fichiers Excel téléchargeables), et pour le trimestre actuel par rapport au même trimestre un an plus tôt (en d'autres termes les *transitions annuelles d'un trimestre spécifique* les plus récentes, indiquées par JQ). Chaque année, Statbel publiera les *transitions annuelles* (en d'autres termes, les moyennes des quatre plus récentes transitions annuelles d'un trimestre spécifique, indiquées par JJ). Chaque matrice de transition est accompagnée d'une matrice des taux de transition (ou pourcentages) et une matrice des tailles des échantillons de répondants (ou transitions non pondérées).

Pour chacun de ces trois types (QQ, JQ et JJ) de transitions, Statbel publie les transitions pour l'ensemble de la population des 15-74 ans, ainsi que des ventilations selon le sexe, la région, l'âge (moins de 30 ans vs au moins 30 ans), le niveau d'instruction (faible, moyen, élevé) et la nationalité (Belge vs non Belge). Ci-dessous, nous discutons pour chacun des trois types de transition les résultats et un certain nombre d'aspects qui doivent être pris en compte lors de l'utilisation et de l'interprétation.

Depuis janvier 2021, Statbel publie les transitions trimestrielles (type QQ) et les transitions annuelles d'un trimestre spécifique (type JQ) en même temps que les résultats trimestriels pour les indicateurs clés. Statbel fournit aussi les séries chronologiques de ces transitions : à partir de 2017T1-2017T2 pour le type QQ et à partir de 2017T1-2018T1 pour le type JQ. Depuis 2021 également, Statbel publie les transitions annuelles (type JJ). Elles seront à l'avenir publiées en même temps que les chiffres annuels, à savoir fin mars.

3.1 Transitions trimestrielles : transitions entre trimestres successifs

Nous expliquons brièvement ci-dessous quels chiffres sont publiés sur les transitions entre les trimestres successifs et comment ceux-ci peuvent être interprétés. Comme au chapitre 2, nous utilisons la paire de trimestres 2018T3-2018T4 pour l'illustrer. Les transitions trimestrielles publiées pour cette paire se trouvent dans la feuille de calcul 2018T3-T4 du fichier Excel téléchargeable EFT_TRANSITION_FR_QQ_P.xlsx ; les chiffres pour l'ensemble de la population des 15-74 ans sont compilés ci-dessous dans le Tableau 13. Les chiffres du volet A de ce tableau sont les estimations finales des transitions pour la paire de trimestres mentionnée. Notez qu'il s'agit des sommes arrondies des estimations par sexe, telles que présentées dans le Tableau 12, par exemple 90.155,76 + 60.299,26 = 150.455,02 dans le Tableau 12 devient 150.455 dans le Tableau 13.

Les transitions du volet A du Tableau 13 indiquent une estimation du nombre de personnes ayant effectué une transition particulière. Par exemple, nous constatons que 48.210 personnes sont passées d'occupé à chômeur. Ainsi, elles ont donc perdu l'emploi qu'elles avaient en 2018T3, mais recherchent activement un autre emploi en 2018T4. Le tableau montre également, par exemple, combien de personnes qui étaient auparavant au chômage ou inactives ont trouvé un emploi : 78.330 chômeurs et 190.123 inactifs au 2018T3 ont trouvé un emploi au 2018T4.

Les totaux pour la fin du trimestre 2018T4, c.-à-d. la distribution absolue estimée du statut BIT au 2018T4 (qui n'apparaît pas dans le Tableau 13), sont identiques à ceux qui résultent du calage de l'échantillon trimestriel 2018T4. Les totaux du trimestre de départ 2018T3, pour les chômeurs (300.216) et les personnes occupées (4.785.249), sont identiques à ceux qui résultent du calage de l'échantillon trimestriel de 2018T3 ; pour les inactifs, le total (3.344.858) diffère du résultat de ce calage (3.325.058), à la suite de la correction nécessaire pour l'incohérence numérique entre BQ et EQ.[16]

Le Tableau 13 montre également les pourcentages de transition dans le volet B. Il s'agit de pourcentages de lignes calculés à partir de la matrice de transition du volet A (voir le paragraphe 1.5 qui explique pourquoi le choix (arbitraire) des pourcentages de *lignes* a été fait dans cette analyse). Nous trouvons dans la diagonale du coin supérieur gauche au coin inférieur droit les pourcentages de personnes qui n'ont pas effectué de transition et dont le statut est donc stable. Ainsi, entre 2018T3 et 2018T4, 94,8 % des personnes restent occupées sur le marché du travail (soit 4.536.279 des 4.785.249 personnes). Les cellules qui ne sont pas sur la diagonale reprennent les chiffres des personnes qui ont effectué une transition : par exemple, 26,1 % des personnes passent de chômeur au 2018T3 à occupées au 2018T4 (soit 78.330 des 300.216

[16] Pour un aperçu des tableaux publiés par Eurostat, voir https://ec.europa.eu/eurostat/web/lfs/data/database ; les tableaux LFSQ_UGAN, LFSQ_EGAN et LFSQ_IGAN permettent par exemple de retrouver les totaux 300.216, 4.785.249 et 3.325.058 (pour 2018T3) en milliers, c.-à-d. 300,2 (x1000), 4.785,2 (x1000) et 3.325,1 (x1000).

personnes). De plus, parmi les personnes qui étaient inactives au 2018T3, 91,6 % sont toujours inactives, 2,7 % au chômage et 5,7 % au travail au 2018T4.

Tableau 13 Matrice de transition trimestrielle publiée pour 2018T3-2018T4, avec matrices associées de taux de transition et de taille d'échantillon – cf. publication EFT_TRANSITION_FR_QQ_P.xlsx

A. Transitions				
2018T4 / 2018T3	Chômeur T actuel	Occupé T actuel	Inactif T actuel	Total
Chômeur T précédent	150.455	78.330	71.431	300.216
Occupé T précédent	48.210	4.536.279	200.760	4.785.249
Inactif T précédent	91.738	190.123	3.062.997	3.344.858
B. Pourcentages de transition				
2018T4 / 2018T3	Chômeur T actuel	Occupé T actuel	Inactif T actuel	Total
Chômeur T précédent	50,1 %	26,1 %	23,8 %	100,0 %
Occupé T précédent	1,0 %	94,8 %	4,2 %	100,0 %
Inactif T précédent	2,7 %	5,7 %	91,6 %	100,0 %
C. Transitions non pondérées (tailles des échantillons de répondants)				
2018T4 / 2018T3	Chômeur T actuel	Occupé T actuel	Inactif T actuel	Total
Chômeur T précédent	224	120	130	474
Occupé T précédent	57	7.102	296	7.455
Inactif T précédent	93	239	5.249	5.581

Enfin, dans le volet C du Tableau 13, nous trouvons les nombres non pondérés de répondants pour chacune des neuf transitions possibles, en d'autres termes les tailles des échantillons de répondants. Notez que ces nombres sont également présentés dans le Tableau 1. En tant que tels, ils sont moins intéressants à interpréter directement, mais ils sont importants en tant qu'indication de la précision des chiffres du volet A et du volet B : plus le nombre non pondéré de répondants est faible, moins les estimations correspondantes des chiffres absolus des transitions du volet A et des chiffres relatifs des transitions du volet B sont précises et fiables.[17] C'est important pour l'interprétation des chiffres, et surtout quand on pense constater des fluctuations dans les séries chronologiques. Ils indiquent également que les ventilations selon diverses variables contextuelles pourraient devenir problématiques, comme l'illustrent les alinéas suivants pour les ventilations selon la nationalité et le groupe d'âge.

Ventilation selon la classe de nationalité

La ventilation des répondants par nationalité donne un exemple où les nombres de répondants diminuent rapidement. Comme indiqué au par. 2.1, l'intention initiale était de rendre les matrices de transition cohérentes avec les estimations trimestrielles du statut BIT pour trois classes de nationalité, à savoir *BE*, *EU* et *non EU*. Le Tableau 14 montre une faible représentation des non-Belges (EU et non EU combinés) dans le LS pour 2018T3-2018T4 : les six cellules non diagonales contiennent chacune moins de 30 répondants, ce qui rend les transitions et les taux de transition estimés dans ces six cellules peu fiables. Une distinction supplémentaire entre EU et non EU donnerait donc des résultats peu utiles.

[17] L'estimation de la variance pour les chiffres et pourcentages des transitions n'a pas encore été réalisée au moment d'écrire cette analyse.

Tableau 14 Distribution du sous-échantillon longitudinal 2018T3-2018T4 des non-Belges (EU et non EU combinés) par statut BIT aux trimestres de départ et de fin

2018T4 / 2018T3	Chômeur T actuel	Occupé T actuel	Inactif T actuel	Total
Chômeur T précédent	65	14	24	103
Occupé T précédent	8	777	28	813
Inactif T précédent	22	27	466	515

Ce qui est illustré ici pour 2018T3-2018T4 s'applique à la plupart, sinon à toutes les paires de trimestres (voir les différentes feuilles de calcul dans EFT_TRANSITION_FR_QQ_P.xlsx). C'est la raison pour laquelle nous n'avons finalement retenu dans les modèles de calage que la dichotomie *BE* versus *non-BE* (soit NAT1 et NAT2).

Ventilation selon la classe d'âge

Lorsque l'on ventile les matrices de transition par classe d'âge, un problème analogue à celui discuté ci-dessus pour la ventilation par classe de nationalité se pose. Comme indiqué au paragraphe 2.1, des classes d'âge de 10 ans, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74 (pour la population des 15-74 ans), ont été utilisées dans les calages pour l'estimation des transitions. Cependant, surtout pour les deux dernières classes, les nombres de répondants selon le statut BIT au BQ et au EQ sont si faibles – comme le montre le Tableau 15 – que cette ventilation n'a pas été retenue dans les publications.

Tableau 15 Distribution des sous-échantillons longitudinaux 2018T3-2018T4 des 55-64 ans et des 65-74 ans par statut BIT aux trimestres de départ et de fin

Classe d'âge 55-64				
2018T4 / 2018T3	Chômeur T actuel	Occupé T actuel	Inactif T actuel	Total
Chômeur T précédent	33	6	22	61
Occupé T précédent	4	1 416	72	1 492
Inactif T précédent	10	36	1137	1 183
Classe d'âge 65-74				
2018T4 / 2018T3	Chômeur T actuel	Occupé T actuel	Inactif T actuel	Total
Chômeur T précédent	0	0	1	1
Occupé T précédent	0	69	23	92
Inactif T précédent	1	17	2.164	2.182

Dans les publications, seule la ventilation en deux groupes d'âge, 15-29 ans versus 30-74 ans est utilisée ; une seule cellule des deux matrices de transition est alors basée sur moins de 30 répondants. Cela permet de contourner le problème de la publication éventuelle d'un (trop grand) nombre de chiffres non fiables, mais cela crée un problème d'incohérence entre les marges des matrices de transition et les estimations trimestrielles antérieures pour les distributions du statut BIT au BQ et au EQ.

Nous illustrons cela à l'aide de la matrice de transition 2018T3-2018T4 pour le groupe d'âge 15-29 ans : voir Tableau 16. Nous trouvons la matrice de transition estimée, avec les totaux des lignes dans la colonne intitulée "Total" dans EFT_TRANSITION_FR_QQ_P.xlsx dans la feuille de calcul 2018T3-T4 (plage A53:E56) ; dans la ligne intitulée "Total", nous avons également ajouté les totaux des colonnes de la matrice de transition. Ainsi, dans la ligne intitulée "Total", nous trouvons d'une part la distribution du statut BIT au 2018T4 pour les 15-29 ans, telle qu'estimée à partir du LS 2018T3-2018T4. Dans la ligne intitulée "Estimation 2018T4", en revanche, nous trouvons la distribution du statut BIT au 2018T4 (pour les 15-29 ans), telle qu'elle a été estimée au cours de ce trimestre sur la base de l'échantillon trimestriel complet (voir la note de bas de page 16 pour trouver ces chiffres sur le site d'Eurostat). Ces distributions ne sont pas égales car lors du calage de l'échantillon trimestriel de 2018T4, une classe d'âge a comme limite supérieure 29 (la variable de calage pour l'âge a comme classes 0-4, 5-9, ... 25-**29**, 30-34, ... 70-74, 75+), alors que lors du calage du LS 2018T3-2018T4 ce n'est pas le cas (la variable de calage pour l'âge a alors comme classes 15-24, 25-34, ... 65-74) ; à noter que, par contre, 15 est bien la limite inférieure d'une classe d'âge dans les deux calages. Les lignes intitulées "Différence" et "% de différence" quantifient l'écart entre les deux distributions du statut BIT pour les 15-29 ans au 2018T4. Un exercice similaire peut être réalisé pour la distribution du

statut BIT au 2018T3 pour les 15-29 ans. Le résultat se trouve dans les colonnes intitulées “Estimation 2018T3”, “Différence” et “% de différence”. Notez que nous avons délibérément omis les chiffres pour le “Inactif T précédent” car le nombre estimé de 1.004.312 devrait être corrigé pour être comparable au total donné de 998.745.[18]

Tableau 16 Matrice de transition publiée pour 2018T3-2018T4, pour la tranche d'âge 15-29 ans, et comparaison avec les distributions du statut BIT basées sur des calages trimestriels

Classe d’âge 15-29							
2018T4 / 2018T3	**Chômeur T actuel**	**Occupé T actuel**	**Inactif T actuel**	**Total**	**Estimation 2108T3**	**Diffé-rence**	**% de dif-férence**
Chômeur T précédent	40.128	42.782	33.796	116.706	***115.368***	1.338	1,16 %
Occupé T précédent	22.648	762.896	103.105	888.649	***905.197***	-16.548	-1,83 %
Inactif T précédent	29.455	80.629	888.660	998.745	**-**	-	-
Total	92.231	886.307	1.025.561	2.004.099			
Estimation 2108T4	***93.115***	***915.015***	***1.023.199***	**2.031.329**			
Différence	-844	-28.708	2.362	-27.230			
% de différence	-0.95 %	-3.14 %	0.23 %	-1.34 %			

Il doit être clair pour le lecteur que de telles incohérences entre les totaux des colonnes et des lignes des matrices de transition et les estimations trimestrielles correspondantes de la distribution du statut BIT peuvent toujours se produire – dans une plus ou moins grande mesure – si la sous-population pour laquelle la matrice de transition est déterminée ne correspond pas aux variables de calage dans le modèle d'estimation de la matrice de transition.

3.2 Transitions annuelles par trimestre : transitions entre les mêmes trimestres de deux années consécutives.

Statbel publie des transitions annuelles par trimestre (à savoir les transitions annuelles d’un trimestre spécifique) dans le fichier Excel téléchargeable EFT_TRANSITION_FR_JQ_P.xlsx. Dans le Tableau 17, nous reproduisons ci-dessous la matrice de transition globale, avec les matrices associées du taux de transition et de la taille de l'échantillon, pour les 15-74 ans, telle qu'elle figure aux lignes 11 à 14 de la feuille de calcul 2018T3-2019T3 dans le fichier Excel susmentionné.

Ce tableau doit bien sûr être lu de la même manière que le Tableau 13.

[18] Le chiffre corrigé qui est comparable à 998.745 est 983.534 = 2.004.099 – (115.368 + 905.197). Cette correction est tout à fait conforme à la correction de l'incohérence numérique qui doit être effectuée avant que le LS puisse être redressé.

Tableau 17 Matrice des transitions annuelles d'un trimestre spécifique publiée pour 2018T3-2019T3, avec matrices associées de taux de transition et de taille d'échantillon – cf. publication EFT_TRANSITION_FR_JQ_P.xlsx

A. Transitions				
2019T3 / 2018T3	**Chômeur T actuel**	**Occupé T actuel**	**Inactif T actuel**	**Total**
Chômeur T précédent	102.360	98.715	99.143	300.218
Occupé T précédent	65.146	4.453.152	266.952	4.785.249
Inactif T précédent	106.414	295.994	2.962.820	3.365.228
B. Pourcentages de transition				
2019T3 / 2018T3	**Chômeur T actuel**	**Occupé T actuel**	**Inactif T actuel**	**Total**
Chômeur T précédent	34,1 %	32,9 %	33,0 %	100,0 %
Occupé T précédent	1,4 %	93,1 %	5,6 %	100,0 %
Inactif T précédent	3,2 %	8,8 %	88,0 %	100,0 %
C. Transitions non pondérées (tailles des échantillons de répondants)				
2019T3 / 2018T3	**Chômeur T actuel**	**Occupé T actuel**	**Inactif T actuel**	**Total**
Chômeur T précédent	146	124	106	376
Occupé T précédent	85	6.183	396	6.664
Inactif T précédent	116	365	4.510	4.991

La comparaison de Tableau 13 et Tableau 17, dans lesquelles 2018T3 est chaque fois le trimestre de départ, montre que la dynamique sur une année est plus importante que sur un trimestre : alors que par exemple, d'un trimestre à l'autre, 50,1 % des personnes restent au chômage, sur une année, ce chiffre est de 34,1 % ; pour les personnes occupées (94,8 % contre 93,21 %) et les inactifs (91,6 % contre 88,0 %), la différence est plus faible, mais la dynamique reste plus importante pour les transitions annuelles.

Notez que les résultats du Tableau 13 sont basés sur un LS de 13.510 répondants, alors que les résultats du Tableau 17 sont basés sur un LS légèrement plus petit de 12.031 répondants. Cela s'explique bien sûr par une attrition plus importante sur un an que sur un trimestre.

Au par. 3.1, il a été illustré que les estimations des transitions trimestrielles de certaines sous-populations (par exemple, les non-Belges ou les 55-64 ans) sont basées sur un petit nombre de répondants, ce qui rend les estimations peu précises. Il en va bien sûr de même pour les estimations des transitions annuelles d'un trimestre spécifique.

En outre, le par. 3.1 a également illustré que les marges des matrices de transition trimestrielles ne reproduisent pas toujours les estimations trimestrielles pour les distributions du statut BIT aux trimestres de départ et de fin, par exemple, lorsque la sous-population des 15-29 ans est isolée. Ce problème se pose de la même manière pour les matrices des transitions annuelles d'un trimestre spécifique.

3.3 Transitions annuelles : transitions entre années consécutives

Après chaque année calendrier, Statbel publie enfin les transitions annuelles dans le fichier Excel téléchargeable EFT_TRANSITION_FR_JJ_P.xlsx. Actuellement, les transitions annuelles sont uniquement disponibles pour 2017-2018, 2018-2019 et 2019-2020. Pour estimer les transitions annuelles, nous calculons la moyenne non pondérée de quatre transitions annuelles d'un trimestre spécifique, comme expliqué au par. 2.9. Autrement dit, les matrices de transitions annuelles dans EFT_TRANSITION_FR_JJ_P.xlsx sont des moyennes non pondérées de matrices de transitions annuelles d'un trimestre spécifique dans EFT_TRANSITION_FR_JQ_P.xlsx. Le Tableau 18 montre la matrice de transition annuelle pour 2018-2019, avec les matrices associées du taux de transition et de la taille de l'échantillon. Il faut noter que la matrice du taux de transition annuelle n'est *pas* la moyenne de quatre matrices de taux de transition annuelle d'un trimestre spécifique. Elle est calculée directement à partir de la matrice de transition annuelle : voir par. 1.5.

Tableau 18 Matrice de transition annuelle publiée pour 2018-2019, avec les matrices associées du taux de transition et de la taille de l'échantillon – cf. publication EFT_TRANSITION_FR_JJ_P.xlsx

A. Transitions				
2018 \ 2019	**Chômeur A actuelle**	**Occupé A actuelle**	**Inactif A actuelle**	**Total**
Chômeur A précédente	109.689	90.340	100.714	300.743
Occupé A précédente	68.180	4.445.661	230.220	4.744.062
Inactif A précédente	95.776	283.498	3.028.909	3.408.183
B. Pourcentages de transition				
2018 \ 2019	**Chômeur A actuelle**	**Occupé A actuelle**	**Inactif A actuelle**	**Total**
Chômeur A précédente	36,5 %	30,0 %	33,5 %	100,0 %
Occupé A précédente	1,4 %	93,7 %	4,9 %	100,0 %
Inactif A précédente	2,8 %	8,3 %	88,9 %	100,0 %
C. Transitions non pondérées (tailles des échantillons de répondants)				
2018 \ 2019	**Chômeur A actuelle**	**Occupé A actuelle**	**Inactif A actuelle**	**Total**
Chômeur A précédente	602	465	496	1.563
Occupé A précédente	366	24.756	1.507	26.629
Inactif A précédente	423	1.418	18.565	20.406

Le même tendance apparait dans le volet B du Tableau 18 et dans le volet B du Tableau 17 en ce qui concerne la dynamique sur le marché du travail d'une année à l'autre (dans ce cas-ci de 2018 à 2019) : un tiers (33,5 %) des chômeurs de 2018 ont trouvé du travail l'année suivante et environ deux tiers non (36,5 % encore au chômage et 30,0 % inactif). Les volets C des deux tableaux montrent évidemment clairement que les (taux de) transitions annuelles sont plus précises que les (taux de) transitions annuelles d'un trimestre spécifique.

Le Tableau 19 ci-dessous montre que, par exemple pour la sous-population non-Belges, les transitions annuelles seront plus précises que les transitions trimestrielles (cf. Tableau 15) ou les transitions annuelles d'un trimestre spécifique pour cette sous-population. Naturellement, cela vaut aussi pour d'autres sous-populations. Les ventilations des matrices de transitions annuelles peuvent par conséquent être plus détaillées que les ventilations des matrices de transitions trimestrielles et de transitions annuelles d'un trimestre spécifique.

Tableau 19 Distribution du sous-échantillon longitudinal 2018-2019 des non-Belges (EU et non-EU) selon le statut BIT au trimestre de départ et de fin

2018 \ 2019	**Chômeur A actuelle**	**Occupé A actuelle**	**Inactif A actuelle**	**Total**
Chômeur A précédente	120	93	102	315
Occupé A précédente	69	2.414	149	2.632
Inactif A précédente	90	161	1.793	2.044

La même remarque qu'aux par. 3.1 et 3.2 s'applique à la cohérence entre les marges des matrices de transitions annuelles et les estimations annuelles pour les distributions de statut BIT au BQ et au EQ.

3.4 Une étude de cas : transitions du chômage de courte durée vs de longue durée

Lorsque le calage a été effectué pour un LS, les transitions de groupes de population arbitraires peuvent être analysées plus en détails. Dans ce paragraphe, nous prendrons l'exemple du chômage et comparerons le chômage de courte durée à celui de longue durée.

Dans ce qui précède, nous avons abondamment utilisé le LS 2018T3-2018T4 afin de présenter les méthodes de Statbel pour estimer les transitions trimestrielles. Le Tableau 13 montre la matrice de transition estimée. Dans ce paragraphe, l'accent est

mis uniquement sur les chômeurs au BQ 2018T3 et sur l'effet de la durée de leur chômage sur les probabilités de transition. Comme d'habitude, nous distinguons les chômeurs de *courte durée*, qui sont au chômage depuis maximum un an (au 2018T3), et ceux de *longue durée*, qui sont au chômage depuis minimum un an (dans 2018T3). La durée de chômage n'est pas connue pour un petit nombre de chômeurs au BQ dans le LS. Le Tableau 20 illustre les transitions estimées pour les chômeurs au BQ, répartis selon la durée de chômage. Il est à noter que les lignes intitulées "Tous les chômeurs" dans le Tableau 20 correspondent exactement aux lignes intitulées "Chômeur T précédent" dans le Tableau 13. Les totaux dans la dernière colonne du volet A du Tableau 20 (excepté le total global de 300.216), qui sont les sommes des poids de calage pour les répondants au chômage dans le LS, ne sont pas égaux aux estimations correspondantes qui pourraient être réalisées sur base de l'échantillon trimestriel de 2018T3, et ce pour la simple raison qu'aucune différence dans la durée de chômage n'est reprise dans les modèles de calage.

Le Tableau 20 permet de conclure que les chômeurs de longue durée au 2018T3 couraient un risque substantiellement plus élevé (65,3 %) d'être toujours au chômage le trimestre suivant par rapport aux chômeurs de courte durée (37,3 %). Parallèlement, les chômeurs de courte durée au 2018T3 avaient une probabilité substantiellement plus élevée (36,8 %) d'être occupés le trimestre suivant que les chômeurs de longue durée (13,1 %). Les probabilités de transition pour les chômeurs dont la durée de chômage est inconnue doivent être ignorées étant donné le nombre restreint de répondants sur lequel ces estimations sont basées. Nous les avons ajoutées afin d'être complets et d'expliquer la correspondance avec les résultats du Tableau 13.

Tableau 20 Transitions trimestrielles 2018T3-2018T4 pour les chômeurs au BQ, selon la durée de chômage

A. Transitions pour les chômeurs au BQ 2018T3				
2018T3 \ 2018T4	**Chômeur**	**Occupé**	**Inactif**	**Total**
Chômeur de courte durée	60.551	59.826	42.018	162.395
Chômeur de longue durée	88.667	17.775	29.413	135.854
Durée de chômage inconnue	1.237	729	-	1.967
Tous les chômeurs	150.455	78.330	71.431	300.216
B. Taux de transition pour les chômeurs au BQ 2018T3				
2018T3 \ 2018T4	**Chômeur**	**Occupé**	**Inactif**	**Total**
Chômeur de courte durée	37,3 %	36,8 %	25,9 %	100,0 %
Chômeur de longue durée	65,3 %	13,1 %	21,7 %	100,0 %
Durée de chômage inconnue	62,9 %	37,1 %	-	100,0 %
Tous les chômeurs	50,1 %	26,1 %	23,8 %	100,0 %
C. Transitions non pondérées pour les chômeurs au BQ 2018T3				
2018T3 \ 2018T4	**Chômeur**	**Occupé**	**Inactif**	**Total**
Chômeur de courte durée	86	93	69	248
Chômeur de longue durée	137	25	61	223
Durée de chômage inconnue	1	2	0	3
Tous les chômeurs	224	120	130	474

Le même exercice peut être réalisé pour les transitions annuelles. Le Tableau 21 montre les taux de transitions annuelles estimés depuis le lancement de l'enquête par panel en Belgique pour les chômeurs pendant l'année de départ, selon la durée de chômage. Il est à noter que la ligne intitulée "Tous les chômeurs" pour l'année de départ 2018 (2ième volet dans le Tableau 21) correspond exactement à la ligne intitulée "Chômeur A précédente" du volet B du Tableau 18 .

Tableau 21 Transitions annuelles 2017-2018, 2018-2019 et 2019-2020 pour les chômeurs pendant l'année de départ, selon la durée de chômage

2018 / 2017	Chômeur	Occupé	Inactif	Total
Chômeur de courte durée	33,3 %	41,3 %	25,4 %	100,0 %
Chômeur de longue durée	51,3 %	19,3 %	29,4 %	100,0 %
Durée de chômage inconnue	-	56,7 %	43,3 %	100,0 %
Tous les chômeurs	42,1 %	30,5 %	27,5 %	100,0 %
2019 / 2018	**Chômeur**	**Occupé**	**Inactif**	**Total**
Chômeur de courte durée	27,4 %	39,9 %	32,7 %	100,0 %
Chômeur de longue durée	45,8 %	19,9 %	34,3 %	100,0 %
Durée de chômage inconnue	-	51,7 %	37,8 %	100,0 %
Tous les chômeurs	36,5 %	30,0 %	33,5 %	100,0 %
2020 / 2019	**Chômeur**	**Occupé**	**Inactif**	**Total**
Chômeur de courte durée	27,9 %	36,9 %	35,3 %	100,0 %
Chômeur de longue durée	52,9 %	14,0 %	33,1 %	100,0 %
Durée de chômage inconnue	-	77,4 %	22,6 %	100,0 %
Tous les chômeurs	38,9 %	26,9 %	34,2 %	100,0 %

Le Tableau 21 confirme également les tendances escomptées : les chômeurs de longue durée ont une plus grande probabilité d'être au chômage et une plus faible probabilité de travailler l'année suivante que les chômeurs de courte durée.

CONCLUSIONS

Dans cette analyse, nous avons expliqué comment les matrices de transition sur le marché du travail belge sont produites par Statbel, et comment elles peuvent être interprétées. Nous avons montré que la méthode d'Eurostat (2015b) pour produire des matrices de transition peut être adaptée pour répondre à diverses exigences via un modèle unique :

- Cohérence avec les distributions globales (pour les 15-74 ans) du statut BIT au BQ et au EQ ;
- Cohérence avec les distributions du statut BIT au BQ et au EQ selon le sexe ;
- Cohérence avec les distributions du statut BIT au BQ et au EQ pour diverses autres sous-populations ;
- Dans une certaine mesure, transférer la structure de l'échantillon redressé EQ dans le LS.

Ces sous-populations peuvent être les cellules au croisement de deux ou plusieurs variables contextuelles telles que la région, le sexe, la classe d'âge, la classe de nationalité, le niveau d'instruction, ... Dans la pratique, bien sûr, la taille du LS sera un élément important pour déterminer les classes de ces variables et pour choisir les variables à inclure dans le modèle du calage.

Nous avons montré que les techniques de calage connues, telles qu'elles ont été introduites par Deville et Särndal (1992), peuvent être appliquées efficacement pour réaliser les calages pour la production de matrices de transition, avec ventilation selon plusieurs variables contextuelles. Les corrections préalables nécessaires des distributions du statut BIT au BQ (en ajustant les nombres estimés de personnes inactives), afin d'obtenir une cohérence entre les distributions du statut BIT au BQ et au EQ utilisées comme benchmarks dans le calage du LS, peuvent également être réalisées en utilisant les mêmes techniques de calage.

La macro SAS® CALMAR2 (Le Guennec et Sautory, 2002 ; Sautory, 1993), complétée par des macros supplémentaires qui permettent une construction efficace et flexible de l'input de CALMAR2, permet à Statbel d'appliquer facilement des modèles sophistiqués. Cela facilite grandement la comparaison de divers modèles de calage potentiels et enfin la détermination d'un modèle final. Bien sûr, la même chose peut être réalisée avec de nombreux autres logiciels dans lesquels les techniques de calage bien connues de Deville et Särndal (1992) sont mises en œuvre.

Les modèles de Statbel pour la production de matrices de transition sont basés sur le calage du LS, c.-à-d. sur la repondération des micro-données. Il s'agit d'une différence fondamentale par rapport aux techniques initiales proposées dans Eurostat (2015b), où les données agrégées sont corrigées à chaque fois. Avec le logiciel approprié, il est donc plus facile d'inclure davantage de variables contextuelles dans les modèles.

Une autre différence importante avec Eurostat (2015b) est que toutes les corrections souhaitées du LS peuvent être effectuées simultanément : à la fois les corrections visant à conserver les distributions du statut BIT au BQ et EQ, et les corrections visant à transférer, dans une certaine mesure, la structure de l'échantillon redressé EQ au LS.

À partir de 2021, Statbel applique ces méthodes intégrées à la production et à la publication d'estimations des transitions sur le marché du travail.

Enfin, nous indiquons que la solution de Statbel pour le calage du LS peut être une alternative à l'application de modèles économétriques (Kiiver et Espelage, 2016). Sous réserve que des méthodes appropriées d'estimation de la variance soient développées pour évaluer la précision des transitions estimées ainsi que pour comparer les transitions dans les sous-populations, l'estimation des matrices de transition pour diverses sous-populations peut soutenir l'analyse statistique (telle que la comparaison des sous-populations).

RÉFÉRENCES

Deville, J.-C., Särndal, C.-E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87, no. 814, 376–382.

Eurostat (2015a) Final report of the TF Flow Statistics, *Eurostat/F3/EMPL/17/15*.

Eurostat (2015b) Eurostat's method of initial flow estimates calculation, Annex 2 to the Final report of the TF Flow Statistics, *Eurostat/F3/EMPL/17/15*.

Kiiver, H., Espelage, F. (2016) The use of regression models in labour market flow statistics, *European Conference on Quality in Official Statistics* (Q2016).

https://ec.europa.eu/eurostat/documents/7894008/8033722/20170519-111211_Q2016_paper_flows_-_final.docx.pdf

Termote A., Depickere, A. (2018) Réforme de l'enquête sur les forces de travail en 2017, *Statbel Analyse*, nr. 4.

https://statbel.fgov.be/sites/default/files/Over_Statbel_FR/Analyse_eak_2017_fr_20181220.pdf

ANNEXES

A APERÇU DES ÉCHANTILLONS LONGITUDINAUX

Tableau B 1 Composition de l'échantillon longitudinal (LS) pour des paires de trimestres consécutifs

Trimestre de départ	Trimestre de fin	RG dans le chevauchement	1er RG	2e RG	3e RG
			Observations des vagues...		
2016T3	2016T4	1 et 2	1 et 2	1 et 2	-
2016T4	2017T1	1, 2 et 6	2 et 3	2 et 3	1 et 2
2017T1	2017T2	2 et 7	3 et 4	1 et 2	-
2017T2	2017T3	3, 4 et 8	2 et 3	2 et 3	1 et 2
2017T3	2017T4	5 et 9	2 et 3	1 et 2	-
2017T4	2018T1	6 et 10	3 et 4	1 et 2	-
2018T1	2018T2	7 et 11	3 et 4	1 et 2	-
2018T2	2018T3	8 et 12	3 et 4	1 et 2	-
2018T3	2018T4	9 et 13	3 et 4	1 et 2	-
2018T4	2019T1	10 et 14	3 et 4	1 et 2	-
2019T1	2019T2	11 et 15	3 et 4	1 et 2	-
2019T2	2019T3	12 et 16	3 et 4	1 et 2	-
2019T3	2019T4	13 et 17	3 et 4	1 et 2	-
2019T4	2020T1	14 et 18	3 et 4	1 et 2	-
2020T1	2020T2	15 et 19	3 et 4	1 et 2	-
2020T2	2020T3	16 et 20	3 et 4	1 et 2	-
2020T3	2020T4	17 et 21	3 et 4	1 et 2	-

Tableau B 2 Composition de l'échantillon longitudinal (LS) pour des paires de mêmes trimestres pendant des années consécutives

Trimestre de départ	Trimestre de fin	RG dans le chevauchement	1er RG	2e RG
			Observations des vagues...	
2016T3	2017T3	3	1 et 3	-
2016T4	2017T4	5 et 6	1 et 3	1 et 3
2017T1	2018T1	6 et 7	2 et 4	1 et 3
2017T2	2018T2	7 et 8	2 et 4	1 et 3
2017T3	2018T3	8 et 9	2 et 4	1 et 3
2017T4	2018T4	9 et 10	2 et 4	1 et 3
2018T1	2019T1	10 et 11	2 et 4	1 et 3
2018T2	2019T2	11 et 12	2 et 4	1 et 3
2018T3	2019T3	12 et 13	2 et 4	1 et 3
2018T4	2019T4	13 et 14	2 et 4	1 et 3
2019T1	2020T1	14 et 15	2 et 4	1 et 3
2019T2	2020T2	15 et 16	2 et 4	1 et 3
2019T3	2020T3	16 et 17	2 et 4	1 et 3
2019T4	2020T4	17 et 18	2 et 4	1 et 3

B PRINCIPES GENERAUX ET TERMINOLOGIE DU CALAGE

Dans cette annexe, nous vous donnons une brève introduction au calage. L'objectif n'est absolument pas d'être exhaustif. Le but est de mieux comprendre le texte principal (chapitre 2), notamment en expliquant la terminologie de la théorie du calage. Les **termes écrits en gras** ci-dessous sont des termes (statistiques) qui sont utilisés (fréquemment en général) dans le texte principal. Chaque fois que de tels termes sont mentionnés dans cette annexe, ils sont écrits en gras afin que le lecteur (du texte principal) puisse les retrouver facilement pour mieux en comprendre le sens. Les *termes écrits en italique* dans cette annexe n'apparaissent pas dans le texte principal : ils sont issus de la théorie du calage ou de la théorie sous-jacente de l'optimisation mathématique. Ces termes ne sont écrits qu'une seule fois en italique.

Vous trouverez de plus amples informations sur la théorie du calage et ses applications dans la littérature ; voir la sélection de publications au par. B.7 de cette annexe. Särndal (2007) et Devaud & Tillé (2019) reprennent de nombreuses autres références traitant de la théorie et des applications.

B.1 OBJECTIF

Le **calage** est l'étape dans le traitement de données où des **poids initiaux** des unités dans un ensemble de données sont adaptés à certaines **distributions de référence**. Le calage résulte directement en des **facteurs de correction** qui sont appliqués aux données à redresser afin d'obtenir des **estimations redressées** de certains **indicateurs**. Par conséquent, le **calage** est une méthode d'estimation.

B.2 DONNÉES DISPONIBLES

Les données à redresser peuvent être agrégées ou non. Dans le cas de données non agrégées, nous partons du principe que les données à redresser forment une liste d'observations individuelles dans laquelle chaque observation est constituée d'un ensemble (ou *vecteur*) de valeurs pour diverses variables dont nous distinguons trois types : les **variables de calage**, les **variables d'étude** et le **poids initial**. Les **indicateurs** dont il est question ci-dessus sont des (combinaisons de) paramètres de variables d'étude, comme les totaux et moyennes, les nombres et proportions, les ratios, ...

Le résultat de l'observation d'unités dans un échantillon aléatoire est l'exemple le plus fréquent de données non agrégées ; le **poids initial** est alors le plus souvent le **poids de sondage**, mais il peut aussi s'agir d'un poids de sondage adapté, comme un poids de sondage corrigé pour la non-réponse, ou un **poids de calage**, ...

Les sections B.3 et B.4 de la présente annexe B supposent implicitement que les données non agrégées doivent être redressées. Au par. B.5, nous reviendrons sur les données agrégées à redresser.

Les **distributions de référence** dont il est question ci-dessus sont calculées à partir d'un ou de plusieurs autres ensembles de données que celui à redresser. Le(s) ensemble(s) de données devant servir de base pour définir les distributions de référence peu(ven)t se présenter sous forme agrégée ou sous forme non agrégée. Dans les deux cas, il peut s'agir d'une population (étudiée ou cible) ou d'un échantillon (pondéré). Les distributions de référence peuvent être des distributions réelles ou des distributions estimées de la population.

B.3 PROBLÈME D'OPTIMISATION MATHÉMATIQUE

Le **calage** (de données non agrégées) peut toujours être formulé comme un *problème d'optimisation mathématique*, plus précisément un *problème de minimisation (mathématique)*. Dans le contexte du **calage**, un tel problème (que nous appellerons **problème de calage** et ultérieurement aussi **modèle de calage**) se compose d'un ensemble d'**équations de calage** et d'une *fonction cible* à minimiser. Les **facteurs de correction** cités ci-dessus sont les inconnues des **équations de calage** et les arguments dans la fonction cible. Résoudre un problème de calage revient à résoudre le problème de minimisation (mathématique), c.-à-d. à trouver la solution des équations de calage qui minimise la fonction cible. Nous supposons dans cette analyse qu'une solution peut être trouvée avec la *méthode des multiplicateurs de Lagrange*.

B.3.1 ÉQUATIONS DE CALAGE

Une **équation de calage** est une égalité mathématique dont le terme de gauche est une somme pondérée d'une **variable de calage** sur (un sous-ensemble de) l'ensemble de données à calibrer et le terme de droite est un nombre réel prédéfini provenant des **distributions de référence**. Chaque poids dans le terme de gauche est le produit d'un **poids initial** connu et d'un **facteur de correction** inconnu ; nous appelons ces produits les **poids de calage**. Le terme de gauche peut toujours être interprété comme une estimation d'un certain paramètre, le terme de droite comme une valeur de référence (une valeur réelle ou une valeur déjà estimée) pour ce paramètre. Les **équations de calage** définissent par conséquent quelles estimations, sur la base de l'ensemble de données à calibrer et en utilisant des **poids de calage,** doivent être égales à des valeurs de référence sûres.

Les **équations de calage** sont toujours linéaires dans les **facteurs de correction** (inconnus), ainsi que dans les **poids de calage** (inconnus).

B.3.2 MESURE DE DISTANCE, FONCTION DE CALAGE ET METHODE DE CALAGE

La fonction cible dans le problème de minimisation est une *mesure de distance globale*, à savoir la somme des *distances* $G_i(w_i, d_i)$ entre le poids initial d_i et le poids de calage w_i des observations individuelles i. Le statisticien qui souhaite effectuer un calage doit choisir les distances (à l'aide du logiciel qu'il désire utiliser). Si la fonction $G_i(.,d_i)$, pour chaque i, satisfait à certaines conditions (différentiabilité continue, convexité, ...), l'inversion de la dérivée de cette fonction mène à ce que l'on appelle une *fonction de calage* $F_i(.)$, qui est telle que $w_i = d_i F_i(\mathbf{x}_i^T \boldsymbol{\lambda})$. Dans ce cas-ci, $\mathbf{x}_i^T$ est le vecteur ligne des variables de calage pour l'observation i et $\boldsymbol{\lambda}$ est le vecteur colonne des *multiplicateurs de Lagrange*. À l'aide des fonctions de calage, les **équations de calage** peuvent être écrites en fonction des multiplicateurs de Lagrange. Pour résoudre le problème de minimisation, il faut donc résoudre le système d'équations en fonction des multiplicateurs de Lagrange. Si la solution $\boldsymbol{\lambda}$ existe et est trouvée, il est alors possible de calculer les **facteurs de correction** $F_i(\mathbf{x}_i^T \boldsymbol{\lambda})$ et par conséquent les **poids de calage** $w_i = d_i F_i(\mathbf{x}_i^T \boldsymbol{\lambda})$.

Choisir les fonctions de distance $G_i(.,d_i)$ revient au même que choisir les fonctions de calage $F_i(.)$. Un tel choix revient à sélectionner ce que l'on appelle la **méthode de calage**. Selon la nature des fonctions de calage choisies, nous parlons de **méthode de calage** linéaire, exponentielle, ...

Certaines distances quadratiques $G_i(w_i, d_i)$ mènent aux fonctions de calage linéaires $F_i(\mathbf{x}_i^T \boldsymbol{\lambda}) = 1 + q_i \mathbf{x}_i^T \boldsymbol{\lambda}$, avec $q_i = F'_i(0)$. Lors de ce choix de distances $G_i(w_i, d_i)$, nous parlons de sélectionner la **méthode (de calage) linéaire**. Dans le cas par exemple de $F_i(\mathbf{x}_i^T \boldsymbol{\lambda}) = exp(q_i \mathbf{x}_i^T \boldsymbol{\lambda})$, nous parlons de la **méthode exponentielle** (ou **multiplicative**). Nous renvoyons à la littérature (par. B.7) pour les fonctions de distance correspondant à ces fonctions de calage.

La **méthode linéaire** peut être complétée par des limites pour les **facteurs de correction** : $L \leq F_i(\mathbf{x}_i^T \boldsymbol{\lambda}) \leq U$, dont la *limite inférieure* L et la *limite supérieure* U doivent être choisies par le statisticien. Nous parlons alors de *méthode linéaire tronquée*. Si la **méthode exponentielle** est complétée d'une certaine manière par de telles limites, nous parlons alors de *méthode logit*. En choisissant de limiter les facteurs de correction de manière adéquate, il est possible d'éviter que les **facteurs de correction**, et donc aussi indirectement les **poids de calage**, soient négatifs ou aient des valeurs extrêmes.

D'autres choix de distances, de fonctions de calage ou de **méthodes de calage** sont discutés dans la littérature. Les développeurs de logiciel de calage font toujours une sélection (restreinte) des méthodes de calage qu'ils implémentent.

B.3.3 STRUCTURE LINÉAIRE

Ce qui précède dans cette annexe montre qu'un **problème ou modèle de calage** est entièrement défini par le choix des **variables de calage** et de la **méthode de calage**. Le choix des **variables de calage** définit les **équations de calage** et l'expression $\mathbf{x}_i^T \boldsymbol{\lambda}$, que nous appellerons la **structure linéaire** (du modèle de calage). Cette **structure linéaire** peut être représentée de manière formelle. En vue d'introduire une notation pratique pour la **structure linéaire**, nous supposons qu'un ensemble de données doit être redressé sur les distributions de trois variables (qualitatives) catégorielles A, B et C. Avec la structure linéaire (additive) A + B + C, nous recherchons le calage sur les *distributions marginales* des variables A, B et C. Avec la **structure linéaire** A + B*C, nous recherchons le calage sur la distribution marginale de la variable A et la *distribution conjointe* des variables B et C. Avec la **structure linéaire** A*C + B*C, nous recherchons le calage sur la distribution conjointe

des variables A et C et la distribution conjointe des variables B et C. Avec la **structure linéaire** A*B*C, nous recherchons le calage sur la distribution conjointe des variables A, B et C. Etc.

Le lecteur comprendra aisément qu'avec (seulement) trois variables catégorielles, il est possible de définir quelques autres structures linéaires. La notation peut évidemment être élargie à un plus grand nombre de variables catégorielles. L'applicabilité d'un **modèle de calage** avec une certaine **structure linéaire** dépend naturellement de la disponibilité des **distributions de référence** induites par la **structure linéaire**.

Dans une **structure linéaire** comme A*C + B*C, nous appelons A*C et B*C les **termes**. La **structure linéaire** A + B + C se compose de trois **termes**, à savoir A, B et C.

Étant donné que, dans le texte principal de cette analyse, les variables de calage sont toujours catégorielles ou qualitatives, nous ne discuterons pas ici des **structures linéaires** reprenant des **variables de calage** *quantitatives*.

Le choix d'une **structure linéaire** définit quelles variables x_j doivent être reprises dans les vecteurs $\mathbf{x}_i^T = \left(\cdots\ x_{ij}\ \cdots\right)$. En général, chaque **terme** dans la **structure linéaire** impliquera un ensemble de plusieurs variables x_j, chacune étant une variable 0-1 ou une variable indicatrice. Si A est par exemple un **terme** dans une **structure linéaire**, une variable 0-1 devra être construite pour chaque valeur a de A. Cette variable aura la valeur 1 pour les observations i pour lesquelles A = a, et la valeur 0 pour les observations i pour lesquelles A ≠ a. Si B*C est par exemple un **terme** dans une **structure linéaire**, une variable 0-1 devra être construite pour chaque combinaison bc de B et C (autrement dit chaque cellule bc au croisement des variables B et C). Cette variable aura la valeur 1 pour les observations i pour lesquelles B = b et C = c, et la valeur 0 pour les observations i pour lesquelles B ≠ b et/ou C ≠ c.

B.3.4 EXISTENCE ET UNICITE DES SOLUTIONS

Le choix adéquat des fonctions de distance $G_i(.,d_i)$ donne une solution unique à un **problème ou modèle de calage**, ce qui est principalement dû à la convexité nécessaire des fonctions de distance. Vous trouverez de plus amples détails à ce sujet dans la littérature (par. B.7).

L'existence d'une solution (dont le résultat final est des **facteurs de correction** et des **poids de calage**) est formellement due à l'existence d'une solution (éventuellement de plusieurs solutions) du système d'**équations de calage**, élargi par l'éventuelle limite des **facteurs de correction** découlant du choix de la **méthode de calage**. Nous énumérons ci-dessous quelques aspects qui peuvent donner lieu à l'existence ou non d'une solution du système élargi d'**équations de calage** :

(1) Les **distributions de référence** doivent être *cohérentes*. Cela signifie notamment toujours que toutes les **distributions de référence** doivent donner précisément le même chiffre total (de population). Plus généralement, cela signifie aussi que si différents **termes** de la **structure linéaire** définissent une même sous-population (comme l'union d'une ou plusieurs cellules résultant de ces termes), les **distributions de référence** correspondantes doivent s'additionner pour donner le même chiffre (de sous-population). Une incohérence peut survenir si les **distributions de référence** sont calculées à partir de différentes sources.

(2) Le système d'**équations de calage** ne peut pas être *surdéterminé*. Avec ceci, nous entendons par exemple (!) ce qui suit. Si un terme de la **structure linéaire** génère une certaine cellule à laquelle correspond un **total de calage** non nul, cette cellule doit alors être représentée dans l'ensemble de données à redresser. Nous pouvons dire formellement que, pour chaque **équation de calage** dont le terme de droite n'est pas nul, le terme de gauche ne peut pas être une "sommation vide".

(3) L'éventuel intervalle $[L,U]$ qui définit la limite des **facteurs de correction** ne peut pas être trop étroit. Il est possible que le système d'**équations de calage** donne des solutions sans la limite, mais que celles-ci ne répondent pas à la limite définie par $[L,U]$. Les logiciels de calage offrent rarement une option pour calculer automatiquement un intervalle (minimal) $[L,U]$ (*ReGenesees*, développé par ISTAT est une exception, voir Zardetto, 2015). En pratique, l'utilisateur commencera par choisir des valeurs "raisonnables" pour L et U et essaiera de résoudre le **problème de calage** avec ces seuils. Si une solution existe, l'intervalle $[L,U]$ peut éventuellement être réduit, en augmentant L et/ou en diminuant U. Si aucune solution n'existe, il faut élargir l'intervalle $[L,U]$ en diminuant L et/ou en augmentant U. L'utilisateur peut ensuite essayer de résoudre le **problème de calage** avec ces nouvelles limites. Chercher un intervalle $[L,U]$ "optimal" (c.-à-d. un intervalle dont le statisticien est satisfait) est donc un processus itératif (en pratique un processus *trial and error*).

B.4 PROPRIÉTÉS PRATIQUES

B.4.1 CARACTERE HIERARCHIQUE DE LA STRUCTURE LINEAIRE ET DISTRIBUTIVITE

Si un **terme** de la **structure linéaire** d'un **modèle de calage** implique la distribution conjointe des variables A, B et C par exemple, alors ce terme implique aussi la distribution conjointe de A et B, de A et C et de B et C ainsi que les distributions marginales de A, B et C. Les **distributions de référence** répondent automatiquement à cette caractéristique (si provenant de la même source). On ne peut dès lors pas redresser un modèle donné sur les distributions conjointes de deux variables A et B p.ex., sans redresser sur les distributions marginales de A et B. A l'inverse, il est bien sûr possible de redresser via un **modèle de calage** donné sur p.ex. les distributions marginales d'un certain nombre de variables mais pas sur certaines distributions conjointes de deux ou plus de ces variables. Cela a pour conséquence que la structure linéaire A*B*C peut être écrite de nombreuses manières différentes :

A*B*C

= A*B*C + A*B + A*C + B*C

= A*B*C + A*B + A*C + B*C + A + B + C

= A*B*C + A + B + C

= …

En outre, le **terme** 1 peut toujours être ajouté, comme par exemple dans A*B*C + 1, ou dans A + B*C + 1 ; ici, 1 représente une variable catégorielle qui ne prend qu'une seule valeur, ce qui se traduit d'ailleurs par une variable x_j de valeur 1 pour toutes les observations.

Enfin, nous pouvons également appliquer une règle de distributivité dans la formulation de **structures linéaires**, comme par exemple dans A*B + A*C = A*(B + C), ou dans A + B*(C*D + E) = A + B*C*D + B*E.

Ces conventions dans la notation des structures linéaires rendent faciles et efficaces la discussion et la représentation de divers **modèles** alternatifs, dont l'un peut être une extension ou une simplification de l'autre, ou dont certains **termes** sont communs mais pas d'autres.

Avec la règle de distributivité, on peut factoriser une structure linéaire, comme par exemple dans A*B + A*C = A*(B + C) dans laquelle le **terme** A est isolé, ou dans A*B*C + D*E*B*C = (A + D*E)*B*C dans laquelle le **terme** B*C est isolé. Cela peut conduire à ce que l'on appelle une *stratification* des **modèles de calage**, et à l'application d'un modèle de calage plus simple (B + C dans le 1er exemple ; A + D*E dans le 2ème exemple) dans chaque *strate de calage* (c.-à-d. la catégorie de A dans le 1er exemple ; la cellule dans le croisement B*C dans le 2ème exemple) séparément. Cela peut être utile pour redresser de grands ensembles de données (Vanderhoeft, 2001), pour faire dépendre la méthode de calage et/ou la limitation des **facteurs de correction** des strates de calage, et même pour faire dépendre la **structure linéaire** des strates de calage.

B.4.2 MODÈLES DE POST-STRATIFICATION

Si la **structure linéaire** est le croisement complet de toutes les **variables de calage** qu'elle contient, on dit alors que le modèle est de type post-stratification, ou on parle simplement d'un **modèle de post-stratification**. Voici des exemples : A*B*C, B*C, A, et aussi le modèle le plus simple avec la **structure linéaire** 1. "*La post-stratification est une technique de pondération simple, bien connue et largement utilisée*", comme indiqué et illustré dans Bethlehem (2008, *transl.*). Cette technique permet d'obtenir un **facteur de correction** et/ou un poids (final) unique pour chaque post-strate. Intégré dans la théorie de calage, cette solution spéciale n'est pas la seule solution possible offerte par le système des **équations de calage** mais cette solution est obtenue en utilisant n'importe quelle méthode de calage.

Notez que pour un **modèle de post-stratification**, le choix de la méthode de calage n'affecte pas le résultat (à condition que le choix de la limitation des **facteurs de correction** soit approprié, et que les **totaux de calage** soient positifs, ce qui est le cas dans les applications standard).

B.4.3 MEME FACTEUR DE CORRECTION POUR TOUTES LES OBSERVATIONS PRESENTANT LES MEMES VARIABLES DE CALAGE

Le par. B.3.2 de cette annexe nous permet de conclure que les **facteurs de correction** $F_i(\mathbf{x}_i^T\boldsymbol{\lambda})$ sont identiques pour toutes les observations i ayant le même vecteur de variables de calage $\mathbf{x}_i^T$. Ceci peut être considéré comme une généralisation de la propriété inhérente à la technique de post-stratification ; voir le paragraphe B.4.2 précédent.

B.4.4 FACTEURS DE CORRECTION POSITIFS ET POIDS DE CALAGE

Toutes les méthodes de calage ne donnent pas nécessairement des **facteurs de correction** uniquement positifs. La **méthode linéaire** peut parfois donner des **facteurs de correction** négatifs. La méthode exponentielle donne toujours des **facteurs de correction** uniquement positifs (si une solution existe). La méthode linéaire tronquée, avec le choix approprié de l'intervalle de limitation $[L, U]$, peut cependant produire des **facteurs de correction** uniquement positifs (si une solution existe).

Enfin, si nous supposons que les **poids initiaux** sont positifs, alors le signe des **facteurs de correction** détermine le signe des **poids de calage**.

B.5 DONNÉES AGRÉGÉES

Les données à redresser ne se présentent pas toujours sous la forme d'un set de données individuelles. Cela peut être le cas lorsque la personne qui doit effectuer le calage n'a pas le droit de traiter les ensembles de données contenant des données individuelles (par exemple, afin de protéger la vie privée des répondants individuels), ou lorsqu'un chercheur veut tester un exercice de calage sur un ensemble de statistiques publié sous forme agrégée.

Dans Eurostat (2015b), les données à redresser sont toujours présentées sous forme de tableaux d'estimations – plus précisément, sous forme de matrices de transition ; voir également le paragraphe 2.8 du chapitre 2 de cette analyse. Ceci est sans doute lié au choix de l'**algorithme** pour effectuer le calage, c.-à-d. l'*iterative proportional fitting* (IPF). Cet **algorithme** a en effet été développé dans le contexte de l'ajustement de données agrégées (par exemple, les tableaux de fréquence à 2 ou plusieurs dimensions) ; une application importante peut être trouvée dans Eurostat (2003).

B.6 LE LOGICIEL UTILISÉ PAR STATBEL

Les statisticiens et méthodologues de Statbel utilisent SAS® Enterprise Guide® pour le traitement et l'analyse des données. Le macro SAS® CALMAR2 (Sautory, 1993 ; LeGuennec & Sautory, 2002) est utilisé pour le calage.

L'input dans CALMAR2 s'effectue en deux parties : (1°) un ensemble de données contenant les données à redresser (généralement issues d'un échantillon de répondants), et (2°) un ensemble de données contenant les totaux de calage pour un modèle de calage déterminé. La préparation du premier ensemble de données est assez élémentaire, étant donné la structure simple de cet ensemble de données et la flexibilité qu'offre la macro CALMAR2 sur ce point. La préparation du second ensemble de données est plus complexe, d'abord parce que la structure de cet ensemble de données n'est pas standard, et ensuite parce que chaque modèle de calage implique un ensemble de données différent. En effet, cet ensemble de données définit en partie le modèle de calage qui sera appliqué. C'est pourquoi Statbel a développé des macros génériques supplémentaires qui créent le deuxième ensemble de données d'input (avec les totaux de calage) pour CALMAR2 à partir d'un ou plusieurs ensembles de données structurés simplement, et qui complètent également le premier ensemble de données d'input (avec les données à redresser) par des variables que CALMAR2 utilise dans le calage. CountsK est l'une de ces macros génériques, qui, par application répétée, calcule et ajoute au deuxième ensemble de données d'input les totaux de calage correspondant aux différents termes de la structure linéaire du modèle de calage, et ajoute une variable de calage appropriée au premier ensemble de données d'input à chaque application. La macro CountsK traite séparément chaque terme A*B*C*... qui dépend d'une ou plusieurs variables de calage catégorielles A, B, C, ... Par exemple :

%CountsK(frame=*Frame1*, wei=*WEI*, sample=*Dataset1*, varlst=*A B C*, lev=*E*, term=*ABC*, …)

calcule, conformément au terme A*B*C, qui résulte de la liste de variables "*A B C*", dans un modèle de calage, les totaux de calage issus de l'ensemble de données *Frame1*, en utilisant une variable de pondération *WEI* dans cet ensemble de données, et enregistre ces totaux dans un enregistrement de l'ensemble de données *Margins_E*. Cet enregistrement reçoit l'identifiant cv_E_ABC, à la suite des valeurs des arguments lev= et term= lorsque CountsK est mobilisé ; "cv" est l'abréviation de *calibration variable* (variable de calage). En outre, la macro CountsK ajoutera la variable de calage cv_E_ABC au *Dataset1*, qui contient les données à redresser ; cv_E_ABC est une numérotation des cellules dans le croisement A*B*C. La macro CountsK vérifie également que chaque cellule de *Frame1* qui est non vide dans le croisement A*B*C est représentée dans *Dataset1*. Des arguments supplémentaires de CountsK contrôlent l'initialisation de l'ensemble de données Marges_E, l'ajout d'un

enregistrement et la fermeture de Marges_E, ainsi que la production d'un aperçu ou d'un rapport de l'application (répétée) de CountsK.

Pour traiter les termes de calage impliquant une variable quantitative et une ou plusieurs variables catégorielles, Statbel a développé la macro TotalsK ; enfin, il y a également la macro Contrast1 qui permet de traiter les limites ou les équations de contraste. Les macros TotalsK et Contrast1 ne sont pas utilisées pour le calage des échantillons longitudinaux. Le traitement complet des macros CountsK, TotalsK et Contrast1 dépasse bien sûr le cadre de cette analyse.

Nous remarquons que *Frame1*, à partir duquel les totaux de calage sont calculés, peut être soit une population, soit un échantillon, et les deux peuvent se présenter sous forme agrégée ou non agrégée. Les pondérations adéquates pour le calcul des totaux de calage sont stockées dans la variable *WEI*. Enfin, il est également important que *Frame1* et *Dataset1* contiennent les variables A, B, C, ... ; en outre, *Frame1* et *Dataset1* peuvent avoir une structure très simple.

L'ensemble des macros CALMAR2, CountsK, ... permet à Statbel de tester de manière très flexible une large gamme de modèles de calage, pour finalement évoluer vers un modèle final.

B.7 REFERENCES SUR LA THEORIE DE CALAGE

Bethlehem, J. (2008) Wegen als correctie voor non-respons, *Statistische Methoden* nr. 08005, CBS, Voorburg/Heerlen. https://www.cbs.nl/-/media/imported/onze-diensten/methoden/gevalideerde-methoden/throughput/documents/2008/10/2008-05-x37-pub.pdf?la=nl-nl

Devaud, D., Tillé, Y. (2019) Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem, *TEST* 28, 1033–1065. https://doi.org/10.1007/s11749-019-00681-3

Eurostat (2003) Handbook on social accounting matrices and labour accounts, Leadership group SAS, *Population and social conditions* 3/2003/E/N°23. https://www.cbs.nl/-/media/imported/onze-diensten/methoden/dataverzameling/aanvullende-onderzoeksbeschrijvingen/documents/2011/38/2011-social-accounting-matrices-and-labour-accounts.pdf

Le Guennec, J., Sautory, O. (2002) Calmar2 : une nouvelle version du macro Calmar de redressement d'échantillon par calage, *Actes des Journées de Méthodologie*, Insee, Paris, pp 33–38. http://www.jms-insee.fr/2002/S01_3_ACTE_LE-GUENNEC-SAUTORY(2)_JMS2002.PDF

Särndal, C.-E. (2007) The calibration approach in survey theory and practice, *Survey Methodology* 33, pp. 99–119. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10488-eng.pdf?st=vAZa5qXq

Sautory, O. (1993) La macro Calmar. Redressement d'un échantillon par calage sur marges, *Document de travail F9310 de la DSDS*, Insee.

Vanderhoeft C (2001) Generalised calibration at Statistics Belgium. SPSS module g-CALIB-S and current practices, Technical report, *Statistics Belgium Working Paper* no. 3.

Zardetto, D. (2015) ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys. *Journal of Official Statistics*, Vol. 31, No. 2, pp. 177–203. http://dx.doi.org/10.1515/JOS-2015-0013

C MODELES DE CALAGE NC : ASPECTS MATHEMATIQUES

L'approche originale d'Eurostat (2015b) pour obtenir la cohérence numérique entre les échantillons redressés BQ et EQ est celle de la cellule. Par “cellule”, nous entendons ici chaque combinaison de variables contextuelles telles que le sexe, la région de résidence, le niveau d'instruction, la classe d'âge, etc. ; par extension, une “cellule” est également toute combinaison de variables contextuelles et du statut BIT sur le marché du travail. Eurostat se limite au sexe et au groupe d'âge comme variables contextuelles.

Dans cette annexe, nous montrons que la méthode d'Eurostat pour adapter l'échantillon redressé BQ à l'échantillon redressé EQ peut être formulée comme un modèle de calage. Cela permet d'adapter l'approche initiale cellule par cellule d'Eurostat si (par exemple) une cellule est vide pour le BQ, mais pas pour le EQ.

C.1 NOTATION

Dans ce qui suit, nous nous limitons aux deux variables contextuelles SEX (sexe) et REG (région de résidence ; le cas échéant, nous distinguons REG1 et REG2 pour BQ et EQ respectivement). SEX peut prendre les valeurs $s = 1$ (homme) et $s = 2$ (femme) ; REG peut prendre les valeurs $r = 1$ (BRU), $r = 2$ (VLA) et $r = 3$ (WAL). Le statut BIT STAT a trois valeurs possibles $b = 1$ (chômeur), $b = 2$ (occupé) et $b = 3$ (inactif) ; selon le contexte, STAT représente STAT1 ou STAT2, c.-à-d. le statut BIT au BQ ou au EQ respectivement. Une “cellule” est une combinaison sr de SEX et de REG, et par extension une combinaison srb de SEX, REG et STAT. L'indice i est utilisé pour identifier les répondants ; avec des notations comme $i \in sr$ (lire : “i élément de la cellule sr”, ou “i dans la cellule sr”, ...) on indique que le répondant i appartient au sous-échantillon de personnes pour lesquelles SEX = s et REG = r. De même, $i \in srb$ indique que le répondant i appartient à la cellule sr et a le statut BIT b, ou pour faire court que i appartient à la cellule srb. Le contexte doit indiquer si un répondant i appartient au BQ, au EQ ou à l'échantillon longitudinal.

Les variables SEX et/ou REG peuvent facilement être remplacées par d'autres variables contextuelles, ou étendues à trois variables contextuelles ou plus.

Les grandeurs suivantes sont essentielles dans le développement mathématique présenté dans cette annexe :

- w_i^{BQ} = poids de calage pour le répondant i de l'échantillon BQ
- w_i^{EQ} = poids de calage pour le répondant i de l'échantillon EQ
- $1_{i \in sr}^{BQ} = 1$ si le répondant i appartient à l'échantillon BQ et se trouve dans la cellule sr, sinon 0
- $1_{i \in sr}^{EQ} = 1$ si le répondant i appartient à l'échantillon EQ et se trouve dans la cellule sr, sinon 0

- $T_{sr}^{BQ} = \sum_{i \in BQ} w_i^{BQ} 1_{i \in sr}^{BQ}$ est le nombre estimé de personnes dans la population de la cellule sr dans le BQ (la somme porte sur l'ensemble de l'échantillon BQ)
- $T_{sr}^{EQ} = \sum_{i \in EQ} w_i^{EQ} 1_{i \in sr}^{EQ}$ est le nombre estimé de personnes dans la population de la cellule sr au EQ (la somme porte sur l'ensemble de l'échantillon EQ)

L'extension de la notation est évidente, et conduit à de nouvelles grandeurs ; par ex :

- Passage de sr à s, ce qui conduit, entre autres, à T_s^{BQ}, le nombre estimé de personnes de la population dans la cellule s au BQ.
- Passage de sr à r, ce qui conduit, entre autres, à T_r^{BQ}, le nombre estimé de personnes de la population dans la cellule r au BQ.
- Transition de sr à srb, ce qui conduit notamment à
  - T_{srb}^{BQ}, le nombre estimé de personnes de la population dans la cellule srb (ou : dans la cellule sr et avec le statut BIT STAT1 = b) au BQ ;
  - $1_{i \in srb}^{BQ} = 1$ si le répondant i appartient à l'échantillon BQ, se trouve dans la cellule sr et a le statut BIT STAT1 = b (soit en bref : se trouve dans la cellule srb), sinon 0 ;

- $1_{i\in srb}^{EQ} = 1$ si le répondant i appartient à l'échantillon EQ, se trouve dans la cellule sr et a le statut BIT STAT2 = b (soit en bref : se trouve dans la cellule srb), sinon 0.

C.2 MÉTHODE CLASSIQUE (NC-C)

Le modèle NC-C SEX * REG1 – un modèle de post-stratification – implique un système d'équations (de calage), plus précisément : une équation de calage pour chaque cellule sr :

$$\sum g_i w_i^{BQ} 1_{i\in sr}^{BQ} = T_{sr}^{EQ}$$

Notez que la sommation dans le membre de gauche de cette équation porte sur l'ensemble de l'échantillon BQ ; c'est le cas pour chaque équation de calage dans cette annexe, puisqu'il s'agit du calage de l'échantillon BQ (sur l'échantillon redressé EQ). Les variables ou inconnues de ce système sont les facteurs de correction g_i (pour les individus i appartenant à l'échantillon BQ) ; elles sont résolues par des méthodes numériques.

Puisque $g_i = g_{sr}$, pour tous $i \in sr$ (théorie de calage !) dans l'échantillon BQ, il s'ensuit que

$$g_i = g_{sr} = {T_{sr}^{EQ}}\Big/{\sum w_i^{BQ} 1_{i\in sr}^{BQ}} = {T_{sr}^{EQ}}\Big/{T_{sr}^{BQ}}$$

où l'on suppose que $T_{sr}^{BQ} > 0$. Si $T_{sr}^{BQ} = 0$ pour une ou plusieurs cellules sr, alors que $T_{sr}^{EQ} > 0$, alors le modèle NC-C SEX * REG1 ne peut pas être appliqué ; en d'autres termes, le système d'équations de calage n'a pas de solution. Les cellules sr pour lesquelles $T_{sr}^{BQ} > 0$ et $T_{sr}^{EQ} = 0$ ne posent aucun problème technique, et pour ces cellules $g_i = g_{sr} = 0$.

Le modèle alternatif NC-C SEX + REG1 implique le système d'équations de calage :

$$\sum g_i w_i^{BQ} 1_{i\in s}^{BQ} = T_s^{EQ} \quad \text{pour tout } s$$
$$\sum g_i w_i^{BQ} 1_{i\in r}^{BQ} = T_r^{EQ} \quad \text{pour tout } r$$

Nous supposons que ce système a une solution. La théorie de calage nous apprend que $g_i = g_{sr}$ pour tous les $i \in sr$, en d'autres termes, que les facteurs de correction sont constants au sein de chaque combinaison sr, mais il est impossible de trouver une expression sous forme fermée comme dans le cas du modèle NC-C SEX * REG1. Les g_{sr} sont ensuite obtenus par résolution itérative du système d'équations de calage, après qu'une fonction cible ait été choisie, qui mesure une (sorte de) quasi-distance entre les poids initiaux w_i^{BQ} et les poids de calage $g_i w_i^{BQ}$. La fonction cible (ou distance) doit être minimalisée sous le système d'équations de calage. (Pour un choix donné de cette fonction cible, la méthode itérative peut être réduite à IPF).

C.3 MÉTHODE EUROSTAT (NC-E)

Le modèle NC-E SEX * REG1 * STAT1 – également un modèle de post-stratification – implique trois équations de calage pour chaque combinaison sr :

$$\sum g_i w_i^{BQ} 1_{i\in srb}^{BQ} = T_{srb}^{BQ} \qquad \text{pour} \quad b = 1 \text{ et } 2$$
$$\sum g_i w_i^{BQ} 1_{i\in srb}^{BQ} = T_{srb}^{BQ} + \left(T_{sr}^{EQ} - T_{sr}^{BQ}\right) = \tilde{T}_{srb}^{BQ} \quad \text{pour} \quad b = 3$$

Là encore, nous supposons que le système d'équations de calage (une équation pour chaque combinaisonsrb) a une solution.

Les équations pour $b = 1$ (chômeurs) et 2 (personnes occupées) conduisent à

$$g_i = g_{srb} = {T_{srb}^{BQ}}\Big/{T_{srb}^{BQ}} = 1$$

si i appartient à la cellule sr et a un statut BIT STAT1 = b au BQ. Ceci bien entendu à la condition que $T_{srb}^{BQ} \neq 0$.

Notez que $T_{srb}^{BQ} = 0$ se produit généralement si l'échantillon BQ ne contient pas de répondants $i \in srb$: dans ce cas, aucun g_i ne doit être déterminé pour $i \in srb$; pour une telle cellule vide srb, il n'est bien sûr pas nécessaire d'inclure une équation de calage dans le système. Le cas $T_{srb}^{BQ} = 0$ peut (exceptionnellement) aussi se produire pour une cellule non vide srb, c.-à-d. si $w_i^{BQ} = 0$ pour tous $i \in srb$; dans ce cas, les g_i peuvent prendre une valeur aléatoire (constante). $T_{srb}^{BQ} < 0$ est exclu, car dans le calage de l'échantillon BQ, les poids initiaux sont positifs, et la méthode de calage est choisie de manière à ce que les facteurs de correction soient non négatifs, de sorte que tous les $w_i^{BQ} \geq 0$.

L'équation pour $b = 3$ (inactifs) entraîne que

$$g_i = g_{srb} = {\tilde{T}_{srb}^{BQ}}\Big/{T_{srb}^{BQ}} = \left(T_{srb}^{BQ} + \left(T_{sr}^{EQ} - T_{sr}^{BQ}\right)\right)\Big/{T_{srb}^{BQ}} = 1 + \frac{T_{sr}^{EQ} - T_{sr}^{BQ}}{T_{srb}^{BQ}}$$

si i appartient à la cellule sr et a un statut BIT STAT1 = $b = 3$ au BQ. De nouveau, nous supposons que $T_{srb}^{BQ} \neq 0$, et les mêmes remarques que pour les équations pour $b = 1$ et $b = 2$ peuvent être faites.

Ainsi, les chômeurs et les personnes occupées de l'échantillon BQ ne reçoivent pas un nouveau poids de calage, les inactifs de l'échantillon BQ par contre peuvent en recevoir un si i appartiennent à une cellule sr pour laquelle $T_{sr}^{EQ} \neq T_{sr}^{BQ}$.

Il est important de noter que le total de calage $\tilde{T}_{sr3}^{BQ}$ peut être négatif – de sorte que le facteur de correction g_i peut également être négatif – c.-à-d. pour une cellule sr pour laquelle $\tilde{T}_{sr3}^{BQ} = T_{sr3}^{BQ} + \left(T_{sr}^{EQ} - T_{sr}^{BQ}\right) < 0$ ou $T_{sr3}^{BQ} < T_{sr}^{BQ} - T_{sr}^{EQ}$, c.-à-d. lorsque le nombre total estimé de personnes dans la cellule sr diminue entre le BQ et le EQ (c.-à-d. $T_{sr}^{EQ} < T_{sr}^{BQ}$), et que cette diminution (en valeur absolue) est supérieure au nombre de personnes inactives T_{sr3}^{BQ} dans la cellule sr au BQ. Par conséquent, nous constatons que la méthode d'Eurotat pour obtenir la cohérence numérique entre les échantillons BQ et EQ, peut conduire à des modèles de calage dont les totaux de calage sont négatifs. Voir par. 2.5.4 au chapitre 2 pour une illustration.

Le modèle alternatif NC-E (SEX + REG1) * STAT1 implique trois équations de calage pour chaque s :

$$\sum g_i w_i^{BQ} 1_{i \in sb}^{BQ} = T_{sb}^{BQ} \qquad \text{pour} \quad b = 1 \text{ et } 2$$
$$\sum g_i w_i^{BQ} 1_{i \in sb}^{BQ} = T_{sb}^{BQ} + \left(T_s^{EQ} - T_s^{BQ}\right) = \tilde{T}_{sb}^{BQ} \qquad \text{pour} \quad b = 3$$

et également trois équations de calage pour chaque r :

$$\sum g_i w_i^{BQ} 1_{i \in rb}^{BQ} = T_{rb}^{BQ} \qquad \text{pour} \quad b = 1 \text{ et } 2$$
$$\sum g_i w_i^{BQ} 1_{i \in rb}^{BQ} = T_{rb}^{BQ} + \left(T_r^{EQ} - T_r^{BQ}\right) = \tilde{T}_{rb}^{BQ} \qquad \text{pour} \quad b = 3$$

En supposant que ce système ait une solution $g_i = g_{srb}$ ($i \in srb$), celle-ci ne peut (en général) pas être exprimée sous forme algébrique, mais peut être trouvée en appliquant un algorithme numérique qui donne simultanément une solution au système et minimise une fonction cible choisie – qui mesure ici aussi une quasi-distance entre les poids initiaux w_i^{BQ} et les poids de calage $g_i w_i^{BQ}$. Statbel utilise actuellement à cet effet la macro SAS® CALMAR2.

Notez que, par exemple, l'équation $\sum g_i w_i^{BQ} 1_{i \in sb}^{BQ} = T_{sb}^{BQ}$ du modèle NC-E (SEX + REG1) * STAT1 est la somme sur r des équations $\sum g_i w_i^{BQ} 1_{i \in srb}^{BQ} = T_{srb}^{BQ}$du modèle NC-E SEX * REG1 * STAT1, pour chaque s et pour $b = 1$ ou $b = 2$. De même, l'équation $\sum g_i w_i^{BQ} 1_{i \in rb}^{BQ} = \tilde{T}_{rb}^{BQ}$ du modèle NC-E (SEX + REG1) * STAT1 est la somme sur s des équations $\sum g_i w_i^{BQ} 1_{i \in srb}^{BQ} = \tilde{T}_{srb}^{BQ}$ du modèle NC-E SEX * REG1 * STAT1, pour chaque r et pour $b = 3$. Cela signifie concrètement que si l'on détermine les totaux de calage pour le modèle NC-E "maximal" (SEX * REG1 * ...) * STAT1, on peut obtenir, par de simples sommations de ces totaux de calage, les totaux de calage pour chaque modèle NC-E "plus limité", comme par exemple (SEX + REG1 + ...) *

STAT1. Cela signifie également que les éventuels totaux de calage négatifs $\tilde{T}_{sr3}^{BQ}$ dans le modèle maximal peuvent se transformer en totaux non négatifs $\tilde{T}_{s3}^{BQ}$ et/ou $\tilde{T}_{r3}^{BQ}$ dans le modèle plus limité.

Les totaux de calage négatifs ne sont pas une pratique courante lorsqu'un échantillon doit être redressé sur les distributions (estimées) de la population, puisque ces dernières sont toujours exprimées en nombres non négatifs. Cependant, l'exposé ci-dessus montre que des totaux de calage négatifs peuvent être le résultat logique si l'on choisit la méthode d'Eurostat pour obtenir une cohérence numérique entre les échantillons BQ et EQ. L'apparition de totaux de calage négatifs est bien sûr liée au modèle NC-E choisi, et, comme expliqué au par. 2.6 du chapitre 2, ce choix est lié aux objectifs du calage de LS, c.-à-d. à la cohérence requise entre les marges des matrices de transition et les distributions des statuts BIT au BQ et au EQ. CALMAR2 permet de travailler avec des totaux de calage négatifs, à condition de choisir une méthode de calage appropriée. La méthode linéaire (tronquée) et la méthode logit autorisent des facteurs de correction négatifs, et donc des poids de calage négatif ; l'exponentielle ou le raking ratio, et la méthode sinus hyperbolique ne le permettent pas. Par conséquent, nous devons choisir soit la méthode linéaire (tronquée) soit la méthode logit pour appliquer les modèles NC-E, et pour ne pas devoir changer de méthode selon que l'on doit travailler ou non avec des totaux de calage négatifs.

Puisque nous ne voyons (actuellement) aucune raison de limiter les facteurs de correction à un certain intervalle, nous choisirons finalement toujours la méthode linéaire lors de l'application des modèles NC-E.

D PERTURBATION DE LS DANS DE PETITES SOUS-POPULATIONS, COMPTE TENU DES EXIGENCES DE COHERENCE

Le fait que le LS, en tant que chevauchement des échantillons BQ et EQ, ne contient qu'environ 50 % des répondants de chacun de ces derniers échantillons, peut poser un problème pour les petites sous-populations pour atteindre la cohérence souhaitée entre les marges de la matrice de transition pour la sous-population en question et les distributions trimestrielles du statut BIT. Une intervention technique – une perturbation du LS – a été élaborée à cet effet. Nous illustrons cela en utilisant la sous-population des 65-74 ans, pour la paire de trimestres 2018T3-2018T4.

Le Tableau B 3 montre ceci :

- Colonne (1) : l'échantillon BQ contient 4.663 répondants dans la tranche d'âge 65-74 ans, dont 2 sont au chômage (au 2018T3) ; colonne (3) : 1 seul de ces 2 chômeurs appartient au LS ;
- Colonne (2) : la distribution estimée du statut BIT au BQ, en termes de nombre de chômeurs et de nombre de personnes occupée, doit être reproduite dans la matrice de transition pour les 65-74 ans (si le modèle LS contient le terme AGE1*STAT1) ;
- Colonne (4) : l'échantillon EQ contient 4.781 répondants dans la tranche d'âge 65-74 ans, dont 1 est au chômage (au 2018T4) ; colonne (6) : ce chômeur n'est pas retenu au LS ;
- Colonne (5) : la distribution estimée du statut BIT au EQ doit être reproduite entièrement dans la matrice de transition pour les 65-74 ans (si le modèle LS contient le terme AGE2*STAT2).

Tableau B 3 Distribution selon le statut BIT des répondants de 65-74 ans dans l'échantillon BQ et EQ – avec les distributions estimées du statut BIT –, et des 65-74 ans au BQ et au EQ dans le LS pour 2018T3-2018T4

	BQ ~ 2018T3			**EQ ~ 2018T4**		
Statut BIT	*Nombre de rép. dans l'échantillon du BQ* (1)	*Distribution estimée statut BIT* (2)	*Nombre de rép. dans le LS* (3)	*Nombre de rép. dans l'échantillon du EQ* (4)	*Distribution estimée statut BIT* (5)	*Nombre de rép. dans le LS* (6)
Chômeur	2	529,69	1	1	50,08	0
Occupé	195	47.611,51	79	209	46.172,90	86
Inactif	4.466	1.079.204,22	2.134	4.571	1.085.211,02	2.189
Total	4.663	1.127.345,42	2.214	4.781	1.131.434,00	2.275

Tableau B 4, le volet "*Avant perturbation*", montre la distribution du LS au EQ des répondants de 65-74 ans selon le statut BIT au BQ et au EQ (c.-à-d. la matrice de transition non pondérée pour les 65-74 ans au EQ). En ce qui concerne la reproduction de la distribution du statut BIT au EQ, l'absence de répondants au chômage au EQ dans ce LS pose un problème : l'estimation trimestrielle de 50,08 pour le nombre de chômeurs au EQ, et par conséquent la distribution du statut BIT au EQ (colonne (5) dans le Tableau B 3) ne peut pas être reproduite.

Tableau B 4 Distribution selon le statut BIT au BQ et au EQ des répondants de 65-74 ans dans le LS pour 2018T3-2018T4, avant et après perturbation du LS

	Statut BIT EQ (2018T4)							
	Avant perturbation				***Après perturbation***			
Statut BIT BQ (2018T3)	*Chômeur*	*Occupé*	*Inactif*	*Total*	*Chômeur*	*Occupé*	*Inactif*	*Total*
Chômeur	0	0	1	1	0	0	1	1
Occupé	0	69	23	92	0	69	23	92
Inactif	0	17	2.165	2.182	1	17	2.164	2.182
Total	0	86	2.189	2.275	1	86	2.188	2.275

Une façon évidente de résoudre ce problème serait de modifier le modèle de calage du LS, par exemple en utilisant d'autres classes d'âge (p.ex. en utilisant la variable AGE2 au lieu de AGE2, de sorte que la classe 65-74 fusionne dans la classe plus large 55-74 ; voir par. 2.1). Mais cette méthode ne permet pas de reproduire les distributions prédéfinies des statuts BIT (par exemple, pour les 55-64 ans et les 65-74 ans séparément). Par conséquent, dans le cadre du calage pour l'estimation des matrices de transition, nous choisissons une méthode alternative : le modèle de calage n'est pas modifié, mais nous effectuons une perturbation aléatoire (minimale) de l'échantillon LS.

La perturbation mise en œuvre par Statbel est la suivante : l'un des inactifs du EQ dans le LS (il y en a 2.189) est sélectionné de manière aléatoire et son statut BIT dans le EQ passe d'inactif à chômeur. Un résultat possible de ceci est visible dans le volet "*Après la perturbation*" du Tableau B 4 ; dans ce cas, un répondant qui est également inactif au BQ a été sélectionné de manière aléatoire (mais avec une grande probabilité !), mais ce n'est pas nécessairement toujours le cas. L'échantillon ainsi modifié (pour les 65-74 ans) nous permet de travailler avec le terme AGE2*STAT2 dans le modèle de calage LS – à condition, bien sûr, que d'autres sous-populations ne posent pas non plus de problèmes.

Au Tableau B 5, nous montrons la matrice de transition estimée pour la sous-population des 65-74 ans, après avoir appliqué le modèle de calage final LS-4 (après application de NC-E-3a) à l'ensemble du LS pour 2018T3-2018T4. Notez que la structure de cette matrice est la même que la structure du LS dans le volet "*Après*" du Tableau B 4. Les marges de la matrice de transition reproduisent, comme souhaité, les distributions du statut BIT au BQ et au EQ qui ont été montrées dans le Tableau B 3 (colonnes (2) et (5)).

Tableau B 5 Matrice de transition estimée pour les personnes âgées de 65 à 74 ans au QE, après application du modèle de calage final au LS pour 2018T3-2018T4, et après perturbation du LS

	Statut BIT EQ (2018T4)			
Statut BIT EQ (2018T3)	*Chômeur*	*Occupé*	*Inactif*	*Total*
Chômeur	-	-	529,69	529,69
Occupé	-	33.523,40	14.088,11	47.611,51
Inactif	50,08	12.649,50	1.070.593,24	1.083.292,82
Total	50,08	46.172,90	1.085.211,04	1.131.434,01

Il convient de noter que pour 2018T3-2018T4, l'application du modèle LS-3 (après NC-E-3a) nécessite également une perturbation, car les modèles LS-3 et LS-4 impliquent les mêmes exigences de cohérence.

Au moment de la finalisation de cette analyse, 27 calages distincts de LS ont déjà été effectués : pour l'estimation des transitions trimestrielles de 15 paires (de 2017T1-2017T2 à 2020T3-2020T4) et pour l'estimation des transitions annuelles d'un trimestre spécifique de 12 paires (de 2017T1-2018T1 à 2019T4-2020T4). Pour 11 de ces 27 paires, nous avons utilisé la technique de perturbation pour pouvoir appliquer le modèle de calage final (par. 2.7). Dans 5 de ces 11 cas, pour un répondant aléatoire du LS, le statut BIT au EQ a été modifié d'inactif à chômeur ; dans 5 autres cas, pour un répondant aléatoire du LS, le statut BIT au BQ a été modifié d'inactif à chômeur ; dans 1 cas, pour un répondant aléatoire du LS, le statut BIT au EQ a été modifié d'inactif à chômeur et également pour un autre répondant aléatoire du LS, le statut BIT au BQ a été modifié d'inactif à chômeur.

Il est clair qu'une telle perturbation peut également être appliquée à d'autres sous-populations si nécessaire et/ou si un autre statut BIT n'est pas représenté, éventuellement si un autre modèle de calage – par exemple pour répondre à d'autres exigences de cohérence – doit être appliqué.

À PROPOS DE STATBEL

Statbel, l'office belge de statistique, collecte, produit et publie des chiffres fiables et pertinents sur l'économie, la société et le territoire belges.

Sur la base de sources de données administratives et d'enquêtes, Statbel produit des statistiques reposant sur des fondements scientifiques. Les résultats statistiques sont publiés de manière conviviale et sont disponibles pour tous en même temps.

Statbel utilise les données collectées uniquement à des fins statistiques. En tant qu'office de statistique, nous garantissons à tout moment la protection de la vie privée et des données confidentielles.

Visitez notre site internet
www.statbel.fgov.be

ou contactez-nous
e-mail: statbel@economie.fgov.be

Statbel (Direction générale Statistique - Statistics Belgium)
North Gate - Boulevard du Roi Albert II, 16, 1000 Bruxelles
E-mail: statbel@economie.fgov.be

Numéro d'entreprise
0314.595.348

Editeur responsable
Philippe Mauroy

North Gate
Boulevard du Roi Albert II, 16
1000 Bruxelles