Distr.
GÉNÉRALE

CES/AC.68/2002/10
18 février 2002

FRANÇAIS
Original: ANGLAIS

COMMISSION DE STATISTIQUE et COMMISSION ÉCONOMIQUE POUR L'EUROPE

CONFÉRENCE DES STATISTICIENS EUROPÉENS

ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (OCDE)

COMMISSION DES COMMUNAUTÉS EUROPÉENNES (EUROSTAT)

Réunion commune CEE/EUROSTAT/OCDE sur la comptabilité nationale
(Genève, 24-26 avril 2002)

DÉTECTION DE VALEURS ABERRANTES DANS LES DONNÉES ADMINISTRATIVES À VARIABLES MULTIPLES OBTENUES PAR EXTRAPOLATION, DANS LE CADRE DE LA COMPTABILITÉ NATIONALE BELGE

Communication sollicitée émanant de la Banque nationale de Belgique*

Introduction

1. Le présent document a pour objet d'exposer les principaux résultats obtenus en appliquant un système de détection des valeurs aberrantes, dans le cadre de la comptabilité nationale belge, aux principales variables du compte de production du secteur S11. L'accent est placé sur les résultats et non pas sur des calculs théoriques.

2. L'estimation des variables production (P1), consommation intermédiaire (P2) et valeur ajoutée (B1g) du secteur S11 s'appuie en grande partie (20 % de la valeur ajoutée totale) sur l'extrapolation de données administratives partielles. Parce que les données obtenues par extrapolation comportent toujours un risque quelconque d'erreur, on a besoin dans ce contexte d'un système solide et si possible automatisé de détection des valeurs aberrantes. Comme les variables à l'étude sont de toute évidence liées, la préférence est donnée à une analyse à plusieurs variables.

* Document établi par M. Michel Soudan, Département de la statistique de la Banque nationale de Belgique.

3. La structure du document est simple. Premièrement, des renseignements d'ordre général sont communiqués au sujet du système belge d'estimation, puis les données considérées sont brièvement présentées à l'aide de quelques statistiques de base et d'une analyse des principales composantes. Deuxièmement, des explications sont fournies en ce qui concerne l'emploi de la distance de Mahalanobis aux fins de la détection des valeurs aberrantes. La dernière partie du document porte sur la manière dont ce concept a été mis en œuvre dans le cadre de la comptabilité nationale belge.

Renseignements d'ordre général

4. Des estimations des variables P1 et P2 de S11 sont établies en se fondant à la fois sur des sources administratives presque exhaustives et des enquêtes. Les sources les plus importantes de ces estimations sont les suivantes: comptes annuels des sociétés, déclarations de TVA, déclarations de cotisations sociales et enquêtes sur la structure des entreprises.

5. Toutes les unités économiques sont classées par secteur institutionnel, branche d'activité, région et catégorie. Les estimations initiales de P1, P2 et B1g pour le secteur S11 sont donc élaborées à un niveau très détaillé: par branche d'activité (249 branches d'activité), district (NUTSIII, 44 districts) et «catégorie» (10 catégories).

6. Le concept de «région» est pris en compte en tant que variable de classification pour assurer un maximum de cohérence entre les comptes nationaux et les comptes régionaux. Toutefois, l'inclusion de cette variable de classification depuis l'année 2000 a entraîné une multiplication par 19 du nombre d'agrégats: on est passé d'environ 1 300 agrégats à plus de 25 000. Une inspection manuelle n'est donc plus envisageable.

7. La catégorie reflète la taille de la société (grandes entreprises par opposition aux petites et moyennes entreprises) et la quantité de données que l'on peut extraire de sources administratives. Pour certaines catégories, seules des informations sur la valeur ajoutée ou sur les salaires sont disponibles. Pour d'autres, on peut obtenir des informations se rapportant aux variables P1, P2 et B1g.

8. Lorsque des données ne sont disponibles que pour une seule des trois principales variables du compte de production, on procède à une extrapolation fondée sur les observations relatives à un groupe analogue d'entreprises afin d'obtenir des estimations pour l'ensemble des trois variables. En l'occurrence, le terme «analogue» signifie que les entreprises en question sont classées dans le même secteur, la même branche d'activité et la même région. Elles ne sont différentes que sur le plan de la catégorie.

Étude des données

9. L'analyse des valeurs aberrantes devrait prendre en considération deux types distincts de valeurs atypiques. Premièrement, P1 et P2 devraient être à peu près linéairement liées (au niveau des agrégats) et cette relation devrait présenter par une certaine stabilité dans le temps. Toute modification devrait être progressive. Deuxièmement, l'évolution des agrégats ne devrait pas dépasser des limites raisonnables. Les agrégats qui accusent des taux de croissance extrêmement élevés ou faibles sont sujets à caution.

10. Pour pouvoir effectuer ce genre d'analyse, il faudrait prendre en considération aussi bien la variable P1 que la variable P2. Afin d'inclure la dimension chronologique, l'analyse devrait tenir compte d'observations de variables portant sur plus d'une année. Pour effectuer une analyse sur plusieurs variables, celles-ci doivent être groupées en vecteurs.

11. La population statistique considérée correspond donc à un ensemble de vecteurs complets, strictement positifs englobant les variables ci-après:

$$X = (P1_t \;\; P2_t \;\; P1_{t-1} \;\; P2_{t-1})'$$

12. Pour la période 1998-1999, X devient $(P1_{1999}\ P2_{1999}\ P1_{1998}\ P2_{1998})'$. B1g n'est pas pris en compte comme variable car sa valeur peut être directement calculée sur la base de P1 et P2. La population comprend 23 432 vecteurs sur un total possible de 28 390 (soit 83 %). La différence entre les deux chiffres est imputable au fait que les caractéristiques individuelles ne sont pas stables dans le temps. Par exemple, certaines combinaisons (secteur, branche d'activité, région, catégorie) peuvent exister durant une période considérée mais pas au cours de l'autre période en raison de reclassements, de fusions et d'acquisitions ou de changements intervenus dans la disponibilité des données. En outre, en 1999, deux catégories nouvelles ont été créées pour lesquelles on ne dispose d'aucune information se rapportant à 1998. Il en résulte qu'un nombre considérable d'agrégats doivent encore être contrôlés manuellement, essentiellement sur la base de ratios.

13. En raison de l'étendue de la gamme de variables, celles-ci font l'objet d'une transformation logarithmique. Le plus petit agrégat de la population statistique est 34 euros, le plus grand 16,5 milliards d'euros. Toutes les variables sont exprimées en log(BEF). Ainsi qu'il ressort des graphiques 1a à 1d, la distribution des variables transformées (et normalisées) semble être quasiment normale, une légère asymétrie vers la droite étant présente. Toutefois, le test de Kolmogorov-Smirnov rejette l'hypothèse de la normalité pour l'ensemble des quatre variables avec une marge d'erreur de 1 %.

14. Les tableaux 1 et 2 révèlent que les variables sont très semblables et fortement corrélées. Une analyse des principales composantes basée sur la matrice de corrélation montre que le nuage de points dans l'espace quadridimensionnel peut aisément être réduit à un espace bidimensionnel, voire unidimensionnel (voir le tableau 3).

15. Les vecteurs propres associés du tableau 4 montrent que la structure du nuage de points est simple. La première composante principale est juste une variable de taille. La deuxième correspond au contraste entre les deux années considérées. Les autres composantes sont aussi des contrastes. Toutefois, elles n'expliquent qu'une portion négligeable de la variation.

16. Il est clair que les données ayant subi une transformation logarithmique constituent un nuage de points dans l'espace quadridimensionnel, qui prend la forme d'un ellipsoïde très allongé avec seulement quelques écarts très faibles par rapport à son axe principal.

Méthode de détection des valeurs aberrantes

17. Pour repérer la présence de valeurs aberrantes, la distance de Mahalanobis est calculée pour chacune des observations. Sur le plan théorique, on peut diviser cette opération en

deux étapes. Premièrement, les variables considérées sont soumises à une transformation de Mahalanobis qui normalise (moyenne = 0, variance = 1) et orthogonalise les données originelles (c'est-à-dire que leurs intercorrélations deviennent égales à 0).

$$y = C^{-1/2} \cdot (x - x_{avg})$$

où

x = le vecteur des données (logtransformées originelles)

C = la matrice de covariance estimée

y = le vecteur des données transformées

x_{avg} = le vecteur estimé des moyennes

18. Deuxièmement, la distance de Mahalanobis entre les données transformées et l'origine 0 est calculée en appliquant la formule suivante:

$$D^2 = y' \cdot y = y_i^2 = (x-x_{avg})' \cdot C^{-1} \cdot (x-x_{avg})$$

où D (= $+\sqrt{D^2}$) est égal à la distance de Mahalanobis.

19. Géométriquement parlant, les observations présentant la même distance de Mahalanobis se trouvent sur un ellipsoïde. La distance de Mahalanobis est faible pour les observations qui se trouvent sur l'axe principal de l'ellipsoïde ou s'en approchent. Les observations qui sont plus éloignées de l'axe principal ont une distance de Mahalanobis beaucoup plus élevée. Le graphique 2 illustre cette constatation pour les deux variables. L'ellipsoïde est formé par toutes les observations pour lesquelles la distance de Mahalanobis est égale à 1. La distance de Mahalanobis de la valeur aberrante (fictive) à (7,5;8,5) − ce qui signifie que P2 est 10 fois supérieure à P1 – est égale à 7,12.

Application pratique

20. L'objectif final d'un système de détection des valeurs aberrantes consiste à repérer les observations douteuses afin de pouvoir les soumettre à une inspection manuelle. Pour ce faire, la population totale des observations est divisée en deux en fonction de la taille moyenne d'une observation. Les observations pour lesquelles la valeur moyenne des quatre variables est supérieure à 1 milliard de francs belges (environ 25 millions d'euros) constituent ce que l'on appelle «les grandes observations» et les autres observations «les petites observations».

21. On compte 2 558 grandes observations et 20 874 petites. Les observations qu'il convient de contrôler manuellement sont celles dont la distance de Mahalanobis se situe dans le quantile de 5 % supérieur de chaque groupe. Dans le cas des grandes observations, il s'agit de celles dont la distance de Mahalanobis est supérieure à 3. S'agissant des petites observations, la limite est de 3,8. Le tableau 5 contient les tables de fréquence pour les variables D, $D_{grandes}$ et $D_{petites}$.

22. On trouvera dans le tableau 6 un extrait d'un tableau de sortie typique. Outre les données originelles et la distance de Mahalanobis elle-même, les données transformées y sont également incluses. Comme la distance de Mahalanobis au carré est simplement égale à la somme des variables transformées au carré, les données transformées permettent de repérer, parmi les quatre

variables incluses dans le vecteur, celle(s) qui est (sont) responsable(s) de l'écart important. Il convient de noter que Y1 est la variable transformée P1_98, Y2 la variable transformée P1_99, etc. Les variables transformées sont exprimées en écarts types.

23. Par exemple, la distance de Mahalanobis élevée de la première observation est principalement attribuable aux variables P1_98 et P2_98 (Y1 et Y3 respectivement). De toute évidence, P1_98 est beaucoup trop faible par rapport à P2_98.

24. La matrice reproduite dans le tableau 6 a été créée en utilisant un composant com intégré dans une macro Excel. Ainsi, tous ceux qui ont accès à Excel, au sein du Département de la statistique, sont en mesure de créer ce produit. Les autres options (matrice de covariance et de corrélation, produit graphique,…) sont progressivement mis en œuvre. Auparavant, un programme en SAS-IML était utilisé, ce qui limitait la facilité d'accès.

Conclusions

25. Bien que toutes les observations soumises à une inspection manuelle au cours de l'exercice décrit ci-dessus se caractérisent par un comportement atypique, dans la majorité des cas, aucune correction n'a été effectuée après l'inspection. Les fluctuations marquées des agrégats, d'une année à l'autre, sont principalement dues au fait que les sociétés changent de catégorie plutôt qu'à des erreurs d'extrapolation. À partir de l'expérience acquise en s'appuyant sur la distance de Mahalanobis, la division de la comptabilité nationale étudie différents moyens de donner une plus grande stabilité dans le temps à la variable de classification «catégorie».

26. L'expérimentation de la technique exposée ci-dessus a montré que l'emploi de la distance de Mahalanobis est très efficace pour détecter les valeurs aberrantes. C'est une technique très polyvalente de sorte que son utilisation ne se limite pas au cas considéré ci-dessus. Parmi d'autres utilisations potentielles, on peut citer la comparaison de deux ou plusieurs variables censées mesurer le même phénomène (par exemple le chômage) ou la comparaison de deux versions de certaines variables pour identifier les principales révisions.

27. Il faut néanmoins tenir compte de quelques restrictions, la principale étant que les variables retenues pour l'analyse devraient présenter une distribution symétrique et unimodale. Les corrélations entre les variables devraient être très élevées (de préférence $|r| > 90\ \%$). Les relations entre les variables devraient être linéaires, encore qu'une transformation linéarisante permette de résoudre ce problème. Enfin, il est recommandé que le nombre des variables incluses dans l'analyse demeure limité (<10) afin de ne pas compliquer l'interprétation des variables *y* transformées.

TABLEAUX

Tableau 1. Statistiques de base

	P1_99	P2_99	P1_98	P2_98
Moyenne	7,83	7,64	7,80	7,63
Écart type	1,05	1,08	1,04	1,07
Asymétrie	-0,22	-0,19	-0,19	-0,19
Aplatissement	0,15	0,06	0,08	0,06

Tableau 2. Matrice de corrélation

	P1_99	P2_99	P1_98	P2_98
P1_99	100 %			
P2_99	98,8 %	100 %		
P1_98	92,4 %	91,6 %	100 %	
P2_98	91,7 %	92,6 %	98,8 %	100 %

Tableau 3. Analyse des principales composantes

Composantes	Valeurs propres	Proportion	Proportion cumulative
PRIN1	3,829	95,7 %	95,7 %
PRIN2	0,147	3,7 %	99,4 %
PRIN3	0,021	0,5 %	99,9 %
PRIN4	0,004	0,1 %	100,0 %

Tableau 4. Vecteurs propres

	PRIN1	PRIN2	PRIN3	PRIN4
P1_99	0,500	-0,500	0,497	-0,503
P2_99	0,500	-0,500	-0,492	0,508
P1_98	0,500	0,505	0,504	0,492
P2_98	0,500	0,495	-0,508	-0,497

Tableau 5. Tableau de fréquences

Bin	D		$D_{grandes}$		$D_{petites}$	
	Fréq.	% cum	Fréq.	% cum	Fréq.	% cum
0	0	0	0	0	0	0
1	8 676	37	569	22	8 107	39
2	10 109	80	1 613	85	8 496	80
3	2 757	92	240	95	2 517	92
4	961	96	100	99	861	96
5	409	98	8	99	401	98
6	199	99	14	99	185	99
7	121	99	4	100	117	99
8	64	99	10	100	54	99
9	40	100	0	100	40	100
10 et plus	96	100	0	100	96	100
	23 432		2 558		20 874	

Tableau 6. Extrait d'un tableau de sortie (10 premières lignes)

P1_98	P1_99	P2_98	P2_99	lgP1_98	lgP1_99	lgP2_98	lgP2_99	Y1	Y2	Y3	Y4	Dist
5 000	115 492 000	10 001 000	121 220 000	3,70	8,06	7,00	8,08	-22,76	9,63	14,79	-3,93	29,07
17 312	1 527	5 330	818 299	4,24	3,18	3,73	5,91	5,76	-18,38	-10,36	16,03	27,12
20 998	12 389 988	87 354	49 274	4,32	7,09	4,94	4,69	-12,38	14,43	6,80	-13,46	24,27
295 365 000	185 706 000	30 417 000	241 000	8,47	8,27	7,48	5,38	-0,31	11,62	2,90	-14,50	18,81
43 112 339	162 874 745	20 907 989	589 571	7,63	8,21	7,32	5,77	-3,97	11,41	4,70	-12,86	18,26
27 863 551	5 350 146	117 642	2 799 390	7,45	6,73	5,07	6,45	10,76	-4,75	-12,58	4,24	17,73
300 828	809 454	1 392	430 287	5,48	5,91	3,14	5,63	8,32	-4,02	-14,28	4,96	17,72
905 405	13 127 648	188 538 380	39 897 241	5,96	7,12	8,28	7,60	-12,81	1,89	11,58	-1,86	17,47
54 000	7 000	2 299 000	2 463 000	4,73	3,85	6,36	6,39	-4,57	-13,22	3,26	9,50	17,22
17 569 618	1 249 355	15 723 584	15 563	7,24	6,10	7,20	4,19	-2,77	6,32	6,39	-12,75	15,85

Figure I.a
Histogram log(P1_98)

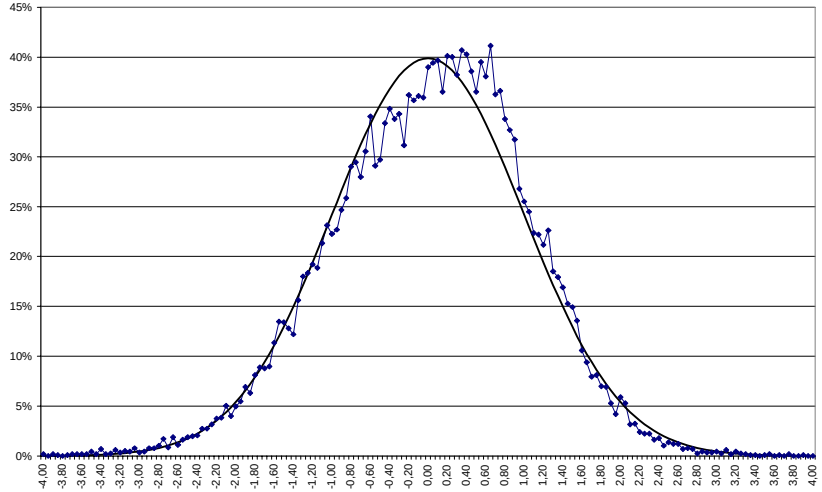

Figure I.b
Histogram log(P1_99)

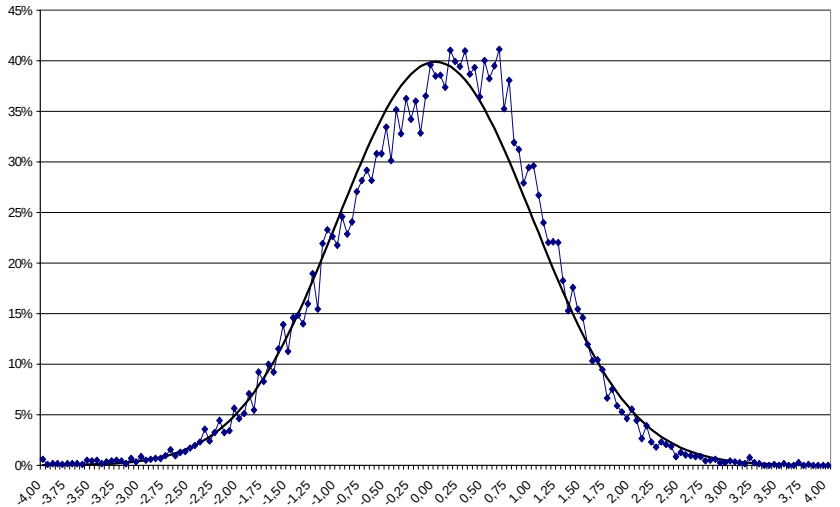

Figure I.c
Histogram log(P2_98)

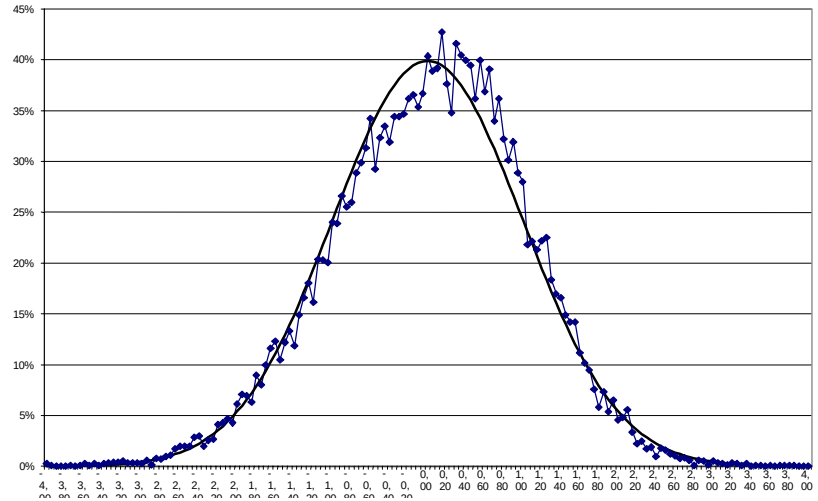

Figure I.c
Histogram log(P2_98)

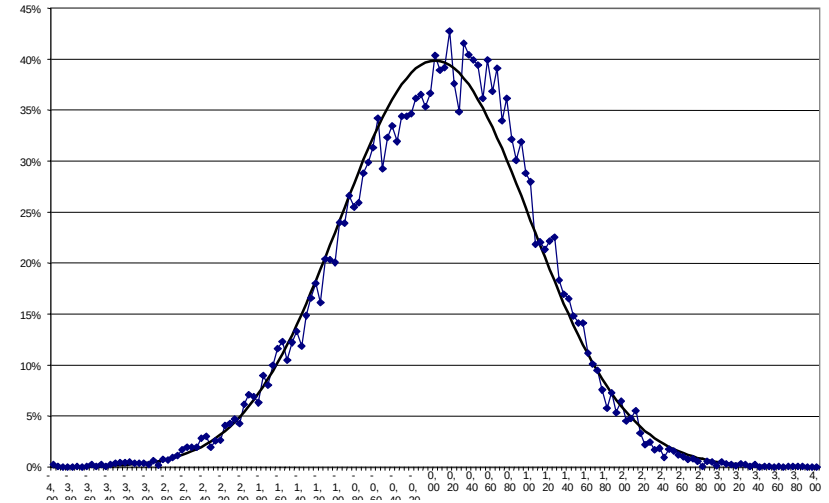

Figure 2: logP1_99 vs. logP2_99