L'Actualité économique

Sur l’estimation des équations de CANDIDE-R
The estimation of CANDIDE-R

Alban D’Amours

Volume 51, Number 4, octobre–décembre 1975

Le modèle CANDIDE (parties 1 et 2)

URI: https://id.erudit.org/iderudit/800650ar
DOI: https://doi.org/10.7202/800650ar

See table of contents

Publisher(s)

HEC Montréal

ISSN

0001-771X (print)
1710-3991 (digital)

Explore this journal

Cite this article

D’Amours, A. (1975). Sur l’estimation des équations de CANDIDE-R. *L'Actualité économique*, *51*(4), 626–633. https://doi.org/10.7202/800650ar

Article abstract

CANDIDE-R is a huge simultaneous macro-economic model which raises estimations difficulties. We avoid the problem of identification assuming that the great number of variables in our model makes it impossible that the necessary condition be not satisfied. We assume that our system converges to a solution solving this way the problem of identification. The core of the paper gives justifications of the procedure we adopted to estimate CANDIDE-R. Because of the presence of regional equations and the limited amount of regional data, we are bound to pool cross sections and time series data. We then justified the use of Zellner's approach instead of the error components models within the class of regional models built on national premises.

SUR L'ESTIMATION DES ÉQUATIONS DE CANDIDE-R

CANDIDE-R est un modèle macro-économique à équations simultanées dont l'estimation soulève un certain nombre de questions d'usage sur l'identification et les méthodes d'estimation à utiliser. L'étudiant sera sans doute choqué de constater l'apparente facilité avec laquelle nous semblons répondre à ces questions mais le caractère particulier de CANDIDE-R nous oblige à formuler un certain nombre d'hypothèses pour assurer la poursuite des travaux. Le professeur John Chipman avait l'habitude de dire à ses étudiants, dont j'étais, que l'économétrie resterait à la disposition des économistes tant et aussi longtemps qu'on leur laisserait la liberté de faire des hypothèses. Bien que la remarque soit judicieuse, elle ne constitue pas pour autant une licence à toutes les prouesses dont trop souvent les vérifications économétriques font usage dans la littérature économique. C'est pourquoi nous croyons utile de rappeler les préoccupations qui ont présidé à l'estimation de CANDIDE-R.

L'identification

L'identification d'un modèle s'intéresse à déterminer si la distribution de probabilités des variables endogènes peut être le fruit d'un ensemble unique de paramètres. Pour répondre à cette question, nous devons habituellement vérifier si les équations du système satisfont les conditions nécessaires et suffisantes pour l'identification [1]. La grande difficulté posée par l'exercice d'identification des équations de CANDIDE-R nous vient du très grand nombre d'équations qui rend impossible la vérification des conditions d'ordre et de rang. Vu le très grand nombre de variables dans le modèle, nous avons supposé presque nulle la probabilité que la condition d'ordre [2] (condition nécessaire) ne soit pas satisfaite. Ainsi nous accordons un ajournement à l'identification. Si le système converge lors de la recherche d'une solution, nous supposerons acquise l'identification de l'ensemble.

1. Voir Johnston, J., *Econometric Methods*, 2e édition, McGraw Hill, pp. 352-375.
2. Voir Johnston, *op. cit.*, pp. 377-423.

Les méthodes d'estimation à équations multiples

Les méthodes d'estimation appropriées aux modèles à équations multiples sont nombreuses et présentent un ensemble de qualités asymptotiques reconnues [3]. La première préoccupation de l'économètre concerne le choix de l'une de ces méthodes lorsqu'il a pour tâche l'estimation d'un modèle comme CANDIDE-R. Cependant, les grandes dimensions de notre modèle limitent sensiblement ce choix. Les moindres carrés à deux étapes sont la seule méthode qui puisse se trouver dans les limites du possible. Mais le grand nombre de variables nous confronte inévitablement au problème de l'insuffisance des degrés de liberté (i.e. le nombre de variables prédéterminées excède le nombre d'observations). Il est bien connu, dans un tel cas, qu'on ne peut réaliser le calcul de la première étape de l'estimation [4]. Kloek et Meuner [5] ont suggéré de remplacer l'ensemble des variables prédéterminées par un nombre restreint et approprié de composantes principales [6].

Nous avons écarté cette avenue pour des raisons à la fois pratiques et théoriques. D'abord, le calcul des composantes principales requises pour chaque équation dépasse nettement les ressources disponibles. A la suggestion de Klein, nous pourrions calculer un seul ensemble de composantes principales dont un nombre très restreint [7] serait utilisé pour fin d'estimation. Vu le grand nombre d'équations et les différents degrés de désagrégation du modèle, cette approche serait évidemment stérile [8].

La distance que nous avons donc prise par rapport aux méthodes d'estimation à équations simultanées ne constituent pas pour autant leur procès. Le choix des techniques d'estimation dans CANDIDE-R est beaucoup plus le fruit d'une recherche que le résultat d'une capitulation devant les difficultés [9].

La désagrégation régionale des données

Une des caractéristiques du modèle touche les divers degrés de désagrégation au niveau spatial et sectoriel. Le principe de la désagrégation

---

3. Voir Johnston, *op. cit.*, pp. 377-423.
4. Voir Johnston, *op. cit.*, pp. 381 et 123.
5. T. Kloek et L.B.M. Meuner, « Simultaneous Equation Estimation Based on Principal Components of Predetermined Variables », *Econometrica*, vol. 28, 1960, pp. 45-61.
6. Voir Johnston, *op. cit.*, pp. 323-334 (sur les composantes principales) et pp. 393-394. Voir aussi L.R. Klein, *A Textbook of Econometrics*, 2e éd., Prentice-Hall, 1974, pp. 183-188. Klein discute, dans ce chapitre, de la possibilité d'étendre l'usage des composantes principales aux estimateurs à information limitée et complète (LIML et FIML).
7. Quatre ou huit composantes principales correspondant aux quatre ou huit plus grandes racines caractéristiques de la matrice des moments des variables prédéterminées. Voir Klein, pp. 228 et suivantes.
8. R. Bridgen M., « Estimation of Large Econometric Models by Principal Components and Instrumental Variable Methods », *R.E.S.*, mai 1971, pp. 140-146.
9. Cette recherche se situe au niveau exclusivement des équations régionales, car nous avons retenu les estimations de CANDIDE pour les équations nationales.

des données pose le problème des interactions potentiellement présentes au niveau de chaque dimension. Comme les données régionales dont nous disposons sont généralement multi-dimensionnelles, c'est-à-dire qu'elles décrivent une industrie, dans une région et ce pour chaque année, nous nous attendons à diverses formes d'interactions tant industrielles, régionales que temporelles. L'usage d'un tableau canadien d'échanges interindustriels au cœur de CANDIDE-R permet de prendre en charge d'une façon adéquate les interactions de dimension industrielle. Concrètement, cette dernière hypothèse nous laisse la liberté de stratifier la masse des données régionales par industrie et, le cas échéant, de conduire les analyses par industrie.

Oubliant pour l'instant le phénomène des interactions régionales et temporelles, nous pouvons illustrer [10] à l'aide des cinq équations suivantes, la spécification des équations régionales de CANDIDE-R.

$$Y_{jt} = B_1 X_{jt,1} + B_2 X_{jt,2} + ---- + B_k X_{jt,k} + \mu_{jt} \qquad (1)$$

où :

$j = 1, ----, R$; régions, $R = 5$
$t = 1, ----, T$; 1961 — 1971
$R = 1, ----, K$; nombre de variables prédéterminées
Y : variable endogène
X : variable prédéterminée
μ : terme d'erreur

Ayant eu soin de poser les hypothèses requises sur les termes d'erreur, i.e.

$E\ (\mu_{jt}) = 0$ pour tout (j,t)

$E\ (\mu_{jt}\ \mu_{rt}) = 0$ pour tout $(j,r,t\ ;\ j \neq r)$

$E\ (\mu^2_{jt}) = \sigma^2_j$ pour tout (j,t)

l'estimation indépendante de ces cinq équations ignorerait les interactions de dimensions régionales dont nous avons fait mention et qui ne sont pas prises en charge par la présence de variables prédéterminées appropriées [11]. Comme aucune équation régionale ne contient de ces variables [12], le seul usage possible des estimations obtenues ne viserait qu'à tester l'homogénéité des comportements relatifs à chaque région

10. Compte tenu de la qualité parfois douteuse des données régionales et de l'ignorance des frontières régionales optimales, nous avons dû accentuer le caractère d'homogénéité des équations régionales pour solidifier nos raisonnements et pour disposer d'un critère additionnel au moment de tester les hypothèses. L'approche globale éloigne, dans ce cas, le danger de partir à la chasse d'équations jugées bonnes exclusivement par la qualité de leur performance statistique (*fit*).

11. La présence de ces variables poserait de toute évidence le problème encore plus délicat de la simultanéité.

12. Variables d'autres régions ou variables de dispersion.

(Test de Chow) [13]. L'utilisation prévue de CANDIDE-R ne peut se satisfaire de ces exercices d'estimation dont l'efficacité gagnerait d'une augmentation du nombre d'observations et de l'utilisation de techniques pouvant tenir compte des phénomènes d'interactions régionales.

L'usage de séries temporelles de coupes instantanées

Les phénomènes dimensionnels d'interactions régionales et temporelles dont nous avons fait état plus haut sont essentiellement de caractère aléatoire et comme nous l'avons expliqué ils ne font l'objet d'aucun raffinement au niveau de la spécification des équations [14] en termes de variables explicatives. Les raffinements habituels dont fait usage l'économétrie pour traiter ces phénomènes aléatoires consistent à purger les données à l'aide de transformations appropriées. Il est bien connu, par exemple, que lorsque les termes d'erreurs d'un modèle n'utilisant que des séries temporelles souffrent du phénomène d'autocorrélation du premier ordre, la transformation des données [15] se présente comme l'unique remède. En général, ces transformations visent à reconstituer les hypothèses d'homoscédasticité et d'indépendance des erreurs et à permettre l'usage de méthodes d'estimation aux qualités reconnues.

Comme nous l'avons déjà souligné, les spécifications dans CANDIDE-R se caractérisent par leur homogénéité. D'une part, nous voulions donner plus de poids à nos raisonnements et, d'autre part, nous voulions faire sortir les similitudes régionales de comportement tout en permettant la possibilité de mesurer, à l'aide de coefficients proprement régionaux, des différences d'intensité. Encadrés par les exigences de ces hypothèses, nous ne pouvions traiter les phénomènes d'interactions qu'en intégrant les dimensions régionales et temporelles par le croisement des séries en coupes transversales (régionales) et des séries chronologiques.

Le premier modèle à retenir notre attention pour intégrer les dimensions régionales et temporelles de nos données est celui connu sous le nom de *modèle à erreurs composées* [16]. Dans ces modèles, on ajoute au terme d'erreur habituel deux composantes servant à représenter les écarts aléatoires par rapport à la moyenne imputables à chacune des dimensions.

13. G.C. Chow, « Tests of Equality Sets of Coefficients in Two Linear Regressions », *Econometrica*, vol. 28, 60, pp. 591-605.

14. La seule exception touche l'utilisation de variables sous la forme de ratio de valeurs régionales sur des valeurs canadiennes.

15. Voir Johnston, *op. cit.*, pp. 259-261. Cette transformation est liée aux restrictions imposées à la matrice de variance-covariance des termes d'erreurs.

16. Il existe une littérature abondante sur ces modèles. Voir, par exemple, G.S. Maddala, « The Uses of Variance Components Models in Pooling Cross Section and Time Series Data » *Econometrica*, vol. 39, nº 2, mars 1971, pp. 341-358 ; J. Kmenta, *Elements of Econometrics*, 1971, pp. 514-517 ; T.D. Wallace et A. Hussain, « The Use of Error Components Models in Combining Cross Section with Time Series Data », *Econometrica*, vol. 37.

Soit l'équation (1) écrite sous forme matricielle :

$$Y_j = X_j B + \mu_j \tag{2}$$

que nous récrivons en supposant l'usage de séries temporelles de coupes instantanées.

$$Y = XB + \mu \tag{3}$$

$$\underset{(5 \times T) \times 1}{Y} = \underset{(5 \times T) \times (K \times 1)}{XB} + \underset{(5 \times T) \times 1}{\mu}$$

où :

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_5 \end{bmatrix} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_5 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_5 \end{bmatrix}$$

Dans le cadre hypothétique des modèles à erreurs composées, nous supposons que $\mu_{jt} = \varepsilon_j + \eta_t + \nu_{jt}$ où

ε_j sont les effets de dimension régionale
η_t sont les effets de dimension temporelle
ν_{jt} sont les effets combinés

Bien que ce genre de modèle soit satisfaisant au niveau des hypothèses intégrant les dimensions régionales et temporelles, il ne l'est pas par rapport aux objectifs poursuivis par la régionalisation de CANDIDE. En effet, l'estimation de ce modèle ne nous livrerait qu'un ensemble de coefficients unique pour toutes les régions. Nonobstant le fait que ces estimations prennent en charge les divers niveaux d'interaction, elles n'offrent pas la possibilité de distinguer suffisamment les comportements régionaux pour satisfaire aux exercices de simulation envisagés avec CANDIDE-R.

La contrainte que nous nous sommes imposée nous force à rechercher des estimations pour chaque équation régionale tout en tenant compte des problèmes d'intégration des données. Utilisant une notation matricielle, nous devons estimer les vecteurs B_j de chacune des équations :

$$\begin{aligned} Y_1 &= X_1 B_1 + \mu_1 \\ Y_2 &= X_2 B_2 + \mu_2 \\ \cdot \quad & \quad \cdot \quad \quad \cdot \\ \cdot \quad & \quad \cdot \quad \quad \cdot \\ \cdot \quad & \quad \cdot \quad \quad \cdot \\ Y_5 &= X_5 B_5 + \mu_5 \end{aligned}$$

ou encore

$$Y_j = X_j B_j + \mu_j \tag{4}$$

Tout exercice du genre implique la présence d'un certain nombre d'hypothèses au niveau de la matrice Ω de variance-covariance des termes d'erreur. Il est clair que si nous n'imposons aucune contrainte à Ω, nous n'arriverons pas à l'estimer. Nous sommes forcés de considérer une situation où Ω dépend d'un nombre restreint de paramètres [17]. Dans ce cas seulement, pouvons-nous penser à construire un estimateur de type Aitken. Le modèle de Zellner [18], tout en offrant le cadre hypothétique pour réaliser l'intégration des données, nous rend possible l'obtention des estimations désirées.

Zellner suggère de combiner les équations en laissant croire qu'elles sont apparemment indépendantes et de passer dans une deuxième étape à des estimations de type Aitken [19].

Combinant les équations (4), nous obtenons :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} X_1 & 0 & 0 & 0 & 0 \\ 0 & X_2 & 0 & 0 & 0 \\ 0 & 0 & X_3 & 0 & 0 \\ 0 & 0 & 0 & X_4 & 0 \\ 0 & 0 & 0 & 0 & X_5 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix}$$

ou encore :

$$Y = Z\rho + \mu \tag{5}$$

Aux hypothèses de normalité des erreurs et de moyennes nulles nous ajoutons les précisions suivantes sur les covariances :

i) $E(\mu_{jt}\,\mu_{js})$ [20] $= 0,$ pour tout j,r

$$\left.\begin{array}{ll} \text{ii)} & E(\mu_{jt}\,\mu_{rs}) = 0 \\ \text{iii)} & E(\mu_{jt}\,\mu_{rt}) = \sigma_{jr} \end{array}\right\} \quad \text{pour tout } j,$$

c'est-à-dire :

$$E(\mu_j\,\mu_r') = \sigma_{jr} \quad I \tag{6}$$

17. Tel est le cas, par exemple, où les termes d'erreur suivent un schéma autorégressif markovien du premier ordre.

18. A. Zellner, « An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias », *J.A.S.A.*, 57, 1962, pp. 348-368.

19. En plus de nous offrir un cadre hypothétique hautement désirable, l'approche de Zellner nous permet d'accroître le nombre de degrés de liberté et de contourner le problème du nombre restreint d'observations annuelles par région.

20. Il est possible d'imaginer à ce niveau l'introduction d'un schéma autorégressif markovien du premier ordre sans nuire à la généralité des hypothèses. Voir J. Kmenta, *op. cit.*, p. 528.

d'où,

$$\Omega = E(\mu\mu') = \begin{bmatrix} \sigma_{11} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \sigma_{15} \\ \cdot & \cdot & \cdot & \sigma_{jr} & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & & & & & \cdot \\ \cdot & & & & & & & \cdot \\ \sigma_{51} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \sigma_{55} \end{bmatrix} \otimes \mathrm{I} \tag{7}$$

où I est une matrice $T \otimes T$ [21]

Si nous appliquons les moindres carrés généralisés sur l'équation (5), nous obtenons l'estimateur d'Aitken :

$$\hat{\gamma} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} X \tag{8}$$

comme meilleur estimateur linéaire sans biais avec, comme matrice de variance-covariance :

$$\mathrm{Var}\ (\hat{\gamma}) = (X' \Omega^{-1} X)^{-1} \tag{9}$$

A la lecture des hypothèses, il est clair que les seuls liens existant entre les équations se réalisent à l'équation (6) sous la forme de covariance entre les termes d'erreur contemporains. Le caractère homoscédastique de ces covariances dans le temps, tout en consacrant l'intégration désirée des dimensions régionales et temporelles, représente la frontière nous séparant des modèles à erreurs composées.

Pour calculer $\hat{\gamma}$ nous devons estimer Ω. Zellner nous suggère l'approche à deux étapes suivante : d'abord, calculer les résidus à partir de l'application des moindres carrés ordinaires sur l'équation (5), i.e., $e_j = Y_j - X_j \hat{B}_j$, pour ensuite généraliser les moindres carrés dans une deuxième étape en utilisant S comme estimateur convergent de Ω où :

$$S = \frac{1}{T} \begin{bmatrix} e_1' \\ \cdot \\ e_2' \\ \cdot \\ \cdot \\ e_5' \end{bmatrix} [e_1\, e_2 - - - - e_5] \tag{10}$$

L'estimateur $\hat{\gamma}$ ainsi obtenu est asymptotiquement équivalent à l'estimateur des moindres carrés généralisés d'Aitken. Il est donc convergent, asymptotiquement efficace et de distribution asymptotique normale [22].

21. Le symbole ⊗ signifie qu'il s'agit d'un produit matriciel de type Knonecker.
22. Voir J. Kmenta, *op. cit.*, p. 525.

A certaines occasions nous avons poussé le raffinement de la technique de Zellner jusqu'à supposer la présence de schémas autorégressifs du premier ordre au niveau des régions. Ainsi nous pouvons tenir compte d'un phénomène additionnel d'interactions de dimension temporelle.

Il suffit de remplacer l'hypothèse i) $E(\mu_{jt}\,\mu_{js}) = 0$ par

$$E[\mu_j\,\mu_j'] = \sigma_{jj} \begin{bmatrix} 1 & \rho_j & \dots\dots & \rho_j^{T-1} \\ \rho_j & 1 & & \rho_j^{T-2} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \rho_j^{T-1} & \rho_j^{T-2} & \dots\dots & 1 \end{bmatrix}$$

et $E(\mu_j\,\mu_r') = \sigma_{jr}$ I par

$$E(\mu_j\,\mu_r') = \sigma_{jr} \begin{bmatrix} 1 & \rho_r & \dots\dots & \rho_r^{T-1} \\ \rho_j & 1 & \dots\dots & \rho_r^{T-2} \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \rho_j^{T-1} & \rho_j^{T-2} & \dots\dots & 1 \end{bmatrix}.$$

où ρ_j est le coefficient d'autocorrélation de l'équation j. Nous pouvons estimer ρ_j à l'aide d'un estimateur convergent [23] comme

$$\hat{\rho}_j \simeq \frac{\sum_t e_{jt}\,e_{j,t-1}}{\sum_t e_{j,t-1}^2}, \quad t = 2, \dots T$$

et l'utiliser comme base de transformation pour purger les données régionales et rétablir les hypothèses (6) et (7).

Conclusion

C'est à la suite de la démarche exposée plus haut que nous avons choisi d'estimer nos équations régionales en réalisant le croisement des séries de coupes transversales et des séries chronologiques à l'aide de l'approche de Zellner. Nons avons ainsi refusé de traiter le problème d'estimation comme secondaire. Même plus, nous l'avons inscrit au cœur même de la régionalisation de CANDIDE en réservant aux techniques d'estimation la tâche de prendre en charge les phénomènes d'interactions régionales.

Alban D'AMOURS,
Université de Sherbrooke.

23. Voir H. Theil, *Principle of Econometrics*, Wiley, 1971, p. 254.