PARIS-SACLAY
APPLIED ECONOMICS

PSAE BRIEF

PSAE Brief n°8

Construire des intervalles de confiance en présence de dépendance multi-grappe ? Quelques clés méthodologiques.

Yannick Guyonvarch

Edito

Les économistes utilisent de manière croissante des ensembles de bases de données très riches et hétérogènes (données issues des réseaux sociaux, données géolocalisées, données administratives individuelles, etc). Ces données posent un véritable challenge méthodologique car elles peuvent présenter des structures complexes de dépendance, en particulier la dépendance multi-grappe (ou multiway clustering) qui est l'objet de cette note. Cette dépendance n'est pas encore bien comprise d'un point de vue statistique. Elle doit cependant être prise en compte pour mesurer l'incertitude qui accompagne l'estimation de l'effet d'une politique économique à partir des nouveaux types de données auxquels les chercheurs ont accès. Nous présentons quelques avancées récentes qui donnent des pistes pour guider les économistes dans le monde complexe de la dépendance multi-grappe.

Sommaire

Introduction

L'évaluation de politiques publiques à l'aide de bases de données est au cœur du travail quotidien du chercheur en économie. Ce travail d'évaluation repose sur deux piliers fondamentaux et complémentaires : estimer l'effet de la politique en tant tel et quantifier l'incertitude qui entoure l'effet mesuré à l'aide d'intervalles de confiance.

Pour être pertinent, un intervalle de confiance se doit de refléter le degré d'incertitude contenu dans les données. L'incertitude dépend à la fois du nombre d'observations présentes dans la base mais aussi également du degré de dépendance entre les différentes observations. Bien appréhender la structure de dépendance dans les données n'est cependant pas aisé : les économistes ont en effet recours à des bases de plus en plus riches (données de réseau, panels d'individus à haute fréquence...) dont la structure de dépendance peut s'avérer complexe.

Dans cette note, nous allons nous intéresser plus particulièrement à la dépendance multi-grappe ou *multiway clustering*. C'est une généralisation de la dépendance par grappe simple. Cette dernière survient lorsque les observations appartenant à des groupes qui vérifient les propriétés suivantes : les observations peuvent être arbitrairement dépendantes si elles appartiennent au même groupe et sont indépendantes sinon. Par exemple, si nous observons les volumes produits par des exploitations agricoles et nous supposons qu'il existe de la dépendance entre les exploitations d'un même département, nous faisons face à un cas de dépendance par grappe simple (avec le département comme grappe). Par extension, la dépendance multi-grappe survient quand plusieurs structures de groupe distinctes et non-imbriquées se superposent. Pour reprendre l'exemple précédent, si nous supposons que les productions des exploitations agricoles peuvent être corrélées si ces-dernières appartiennent au même département ou à la même filière de production (faisons ici l'hypothèse que les exploitations produisent un unique type de bien), alors nous sommes en présence de dépendance bi-grappe. Les départements et le type d'exploitation forment bien deux grappes distinctes et non imbriquées car il existe *plusieurs* types d'exploitation par département et chaque type d'exploitation est présent dans plusieurs départements. La Figure 1 fournit une illustration graphique de la dépendance bi-grappe pour l'exemple de la production agricole.

La dépendance multi-grappe apparaît également avec certaines données de réseau. L'article « International climate aid and trade » co-écrit par Clément Nedoncelle (PSAE), Basak Bayramoglu (PSAE), Jean-François Jacques (UGE) et Lucille Neumann-Noel (U. Paris-Nanterre) en est une bonne illustration. Dans ce travail, les auteurs s'intéressent aux flux d'aide climatique entre un groupe de pays donateurs et de pays récipiendaires. Dans ce cadre, deux flux d'aide peuvent a priori être dépendants s'ils partagent le même pays donateur ou récipiendaire. Les pays donateurs et récipiendaires induisent donc une structure de dépendance bi-grappe.

En présence de dépendance par grappe simple, la construction d'intervalles de confiance se fait à l'aide d'écarts-types *cluster-robust* (robustes à la dépendance par grappe simple en français) bien connus des économistes appliqués. Comment cette approche se généralise-t-elle avec de la dépendance multi-grappe ?

Par souci de simplicité, nous nous intéressons uniquement à la dépendance bi-grappe dans ce qui suit et nous nous concentrons sur l'exemple des exploitations agricoles et des dépendances entre celles-ci via le département ou la filière d'exploitation d'appartenance. L'objectif poursuivi par l'économiste ici est de construire un intervalle de confiance sur le volume de production moyen d'une exploitation agricole en prenant en compte la dépendance bi-grappe liée au département et à la filière de production.

	départ 1	départ 2	départ 3	départ 4	départ 5	départ 6
filière 1						
filière 2						
filière 3						
filière 4						
filière 5						

Figure 1 : exemple d'une structure de dépendance bi-grappe. La production agricole dans la filière 2 et le département 5 (cellule verte) est corrélée à toute la production agricole dans la filière 2 (cellules bleues) et le département 5 (cellules jaunes).

L'approche de Cameron, Gelbach et Miller (CGM)

En 2011, les chercheurs Colin Cameron, Jonah Gelbach et Douglas Miller publient un article intitulé « Robust inference with multiway clustering ». Dans cet article, cité près de 4000 fois depuis sa parution, les auteurs proposent des écarts-types *two-way cluster robust* (robustes à la dépendance bi-grappe).[1] Appliqué à notre exemple des exploitations agricoles, l'estimateur prend la forme suivante

$$se^2_{CGM} = se^2_{département} + se^2_{filière} - se^2_{département \times filière}$$

Le terme $se_{département}$ est un écart-type robuste à la dépendance par grappe simple le long de la dimension ``département''. Le terme $se_{filière}$ un écart-type robuste à la dépendance par grappe simple le long de la dimension ``filière''. Enfin, le terme $se_{département\text{-}filière}$ est un écart-type robuste à la dépendance par grappe simple le long de la dimension ``département x filière''.

Si l'estimateur se_{CGM} est simple à calculer, l'intuition derrière sa formule mérite d'être explicitée. Les deux premières composantes $se_{département}$ et $se_{filière}$ capturent la variabilité inter-départements et inter-filières respectivement. La troisième composante $se_{département\text{-}filière}$ capture la variabilité dans l'échantillon entre des couples département-filière différents. Pourquoi ce terme est-il soustrait ? En étudiant plus finement termes $se_{département}$ et $se_{filière}$, il est possible de voir que ces deux termes capturent à la fois la variabilité inter-départements/inter-filières mais également la variabilité au niveau département x filière. La variabilité au niveau département x filière est donc comptée deux fois au travers du terme $se^2_{département} + se^2_{filière}$ et est donc soustraite pour ne compter qu'une seule fois au final.

L'approche CGM est aujourd'hui très largement répandue en pratique. Elle est par exemple proposée comme option de la commande regress sous Stata et peut être utilisée sous R via le package multiwayvcov. Cette approche admet néanmoins un défaut majeur : du fait de la soustraction du terme $se_{département\text{-}filière}$, l'estimateur se^2_{CGM} peut être négatif, ce qui est une très mauvaise propriété pour un estimateur de variance... En particulier, un estimateur de variance négatif ne permet pas de construire des intervalles de confiance.

Ce problème survient-il en pratique ? Il est malheureusement impossible de répondre à cette question pour l'heure. En effet, aucune méta-analyse n'a été réalisée sur le sujet à ce jour. Un éclairage théorique peut toutefois déjà être apporté. Dans un document de travail « Analytic inference with two-way clustering » à paraître en 2024, nous présentons un exemple simple pour lequel se^2_{CGM} est négatif près de 40% du temps ! Nous discutons également cet exemple plus en détail à la fin de cette note avec la figure 2. Dans le document de travail, nous montrons plus précisément les limites de se^2_{CGM} dans le cas où les données au niveau département x filière se décomposent en un produit d'un choc économique au niveau du département et de la filière. Plus généralement, nous expliquons qu'en présence de dépendance bi-grappe, les données de production agricole admettent une représentation en termes de chocs à l'échelle du département, de chocs au niveau de la filière, de chocs département x filière et de produits des chocs au niveau département et filière. Lorsque les termes produits sont absents, l'approche CGM fonctionne, lorsqu'ils sont présents, cette approche est mise en défaut en général.

Le défi est donc de construire des intervalles de confiance informatifs sans connaître à l'avance la structure précise de dépendance bi-grappe dans les données. Dans la section qui suit nous proposons une alternative simple à se_{CGM} qui permet d'obtenir des intervalles de confiance informatifs y compris dans les situations où l'approche CGM n'est plus valide.

Une méthode alternative

L'alternative à CGM que nous présentons est issue du document de travail « Analytic inference with two-way clustering » mentionné dans la section précédente. L'écart-type modifié se_{modif} vérifie

$$se_{modif} = \max(se_{département}, se_{filière}, se_{CGM})$$

avec se_{CGM} qui n'est pris en compte que s'il est non-négatif. Etant donné que $se_{département}$ et $se_{filière}$ sont des écarts-types standards, ils sont toujours positifs. Ainsi, se_{modif} est assuré d'être positif également.

Comment justifier la construction de se_{modif} ? Plus $se_{département}$ est grand, plus l'incertitude associée aux chocs économiques au niveau département est élevée. Plus $se_{filière}$ est grand, plus l'incertitude associée aux chocs économiques au niveau filière est élevée. Dans le document de travail, nous montrons que prendre le maximum de ces deux écarts-types lorsque se_{CGM} est négatif permet de construire des intervalles de confiance valides mais qui peuvent être conservateurs.[2] Lorsque se^2_{CGM} est positif (avec grande probabilité), il capture l'incertitude dans les deux dimensions (département et filière) et domine donc les termes $se_{département}$ et $se_{filière}$. Il permet par ailleurs de gagner en précision via la soustraction de $se_{département\text{-}filière}$.

En résumé, dans les cas défavorables où se_{CGM} peut être négatif, se_{modif} permet de construire des intervalles de confiance conservateurs qui capturent la source principale d'incertitude dans les données (dimension département ou filière) et dans les cas favorables, se_{modif} permet de construire des intervalles à la fois valides et plus précis.

En Figure 2, nous illustrons graphiquement les comportements de se_{CGM} et se_{modif} sur des données simulées. Nous générons des données de production au niveau département x filière qui se décomposent en un produit d'un choc économique au niveau du département et de la filière. Nous constatons bien le comportement pathologique de se_{CGM} décrit dans la section précédente, à savoir que l'estimateur de variance est négatif dans près de 40% des cas. A l'inverse, se_{modif} demeure positif. Avec les mêmes données, nous concluons qu'un intervalle de confiance à 95% construit à partir de se_{CGM} ne contient le paramètre en réalité que 59% du temps, l'inférence basée sur cette approche est donc trompeuse. A l'inverse, un intervalle à 95% basé sur se_{modif} contient le paramètre 99.7% du temps, ce qui met en avant l'aspect valide mais néanmoins conservateur de cette méthode.

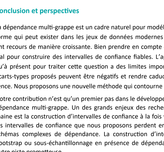

Figure 2 : Distribution en histogramme des estimateurs de variance se^2_{CGM} et se^2_{modif} à partir de 10000 réplications d'un processus générateur de données où les données au niveau département x filière se décomposent en un produit d'un choc économique au niveau du département et de la filière. En abscisse sont indiquées les valeurs prisent par les estimateurs et en ordonnée nous avons le nombre d'estimateurs dans chaque plage de valeurs.

Conclusion et perspectives

La dépendance multi-grappe est un cadre naturel pour modéliser la dépendance protéiforme qui peut exister dans les jeux de données modernes auxquels les économistes ont recours de manière croissante. Bien prendre en compte cette dépendance est crucial pour construire des intervalles de confiance fiables. L'approche CGM utilisée jusqu'à présent pour traiter cette question a des limites importantes : en particulier, les écarts-types proposés peuvent être négatifs et rendre caduque toute tentative d'inférence. Nous proposons une nouvelle méthode qui contourne cette difficulté.

Notre contribution n'est qu'un premier pas dans le développement d'outils adaptés à la dépendance multi-grappe. Un des grands enjeux des recherches futures dans ce domaine est la construction d'intervalles de confiance à la fois valides et précis. En effet, les intervalles de confiance que nous proposons perdent en précision face à certains schémas complexes de dépendance. La construction d'intervalles de confiance par bootstrap ou sous-échantillonnage en présence de dépendance multi-grappe est une autre piste prometteuse.

Pour en savoir plus :

Davezies, L., D'Haultfoeuille, X. et Guyonvarch, Y. « Analytic inference with two-way clustering ». Document de travail à paraître.

Bibliographie

Bayramoglu, A., Jacques, J-F., Nedoncelle, C. and Neumann-Noel, L. (2023) « International climate aid and trade », Journal of Environmental Economics and Management, Vol. 117.

Cameron, C., Gelbach, J. and Miller, D. (2011) « Robust inference with multiway clustering », Journal of Business and Economic Statistics, Vol. 29, No.2.

Miglioretti, D. and Heagerty, P. (2007) « Marginal modeling of nonnested multilevel data using standard software », American Journal of Epidemiology, Vol. 165, No.4.

Thompson, S. (2011) « Simple formulas for standard errors that cluster by both firm and time », Journal of Financial Economics, Vol. 99, No.1.

Notes de fin :

1. Bien que l'article de Cameron, Gelbach et Miller soit le plus connu, deux autres articles ont contribué à la même période - à savoir à la formule des écarts-types two-way cluster-robust. Il s'agit de « Marginal modeling of nonnested multilevel data using standard software » écrit par Diana Miglioretti et Patrick Heagerty en 2007, et « Simple formulas for standard errors that cluster by both firm and time » écrit par Samuel Thompson en 2011.

2. Un intervalle à 95% est dit conservateur s'il contient le paramètre d'intérêt plus que 95% du temps. Un intervalle conservateur est donc moins précis (plus large) qu'un intervalle exact.