

La rareté : un défi pour évaluer les effets des accords commerciaux

Camilo Umana-Dajud*

Les modèles de gravité sont l'un des outils privilégiés pour évaluer les effets *ex post* des accords commerciaux. Très efficaces pour prédire les flux commerciaux, ils rencontrent un écueil de taille pour apprécier les effets des accords commerciaux : la rareté des données des pays signataires. Dès lors, les effets qui ont été estimés jusqu'à présent peuvent être très éloignés de la réalité, et ce dans les deux sens : soit qu'ils n'aient pas été détectés, soit qu'ils aient été surestimés. Dans ces circonstances, le recours à des méthodes statistiques adaptées aux petits échantillons est indispensable. En étant plus exigeantes sur la significativité des résultats obtenus, elles permettront de ne pas surestimer les effets des accords commerciaux. En revanche, face à la rareté des données, elles ne pourront garantir qu'un accord commercial, qui semble ne pas avoir d'effet, n'en ait pas en réalité.

Lorsque les pays s'engagent dans des accords commerciaux pour réduire les entraves aux échanges entre eux, c'est dans l'espoir d'accroître leur commerce, conformément à ce que prédit la théorie économique. Pourtant, les études empiriques peinent parfois à observer un tel effet. Dans certaines, près de 40 % des accords commerciaux n'ont pas d'effet statistiquement significatif, et certains auraient même des effets négatifs, en contradiction avec la théorie. Ce paradoxe provient d'un problème statistique fondamental qui affecte non seulement l'étude du commerce international, mais aussi celle d'autres questions comme l'analyse des flux migratoires et des mouvements de capitaux.

La source du problème ? Chaque accord commercial, comme beaucoup d'autres politiques internationales, ne concerne qu'un nombre très limité de paires de pays par rapport au nombre de paires de pays possibles dans le monde. Cette rareté crée un piège statistique à double tranchant. D'un côté, elle rend très difficile la détection des véritables effets, même quand ils existent. De l'autre, elle produit parfois de fausses certitudes : des résultats qui paraissent statistiquement solides mais qui sont, en fait, très éloignés de la réalité.

■ Les modèles de gravité, un outil indispensable mais fragile

Les modèles de gravité sont devenus un outil omniprésent dans l'analyse économique internationale. Utilisés pour évaluer l'impact des accords commerciaux sur les échanges entre signataires, analyser les flux migratoires entre pays ou encore comprendre les mouvements de capitaux internationaux, ils constituent l'outil sur lequel s'appuient de nombreuses recommandations de politique économique.

La méthode est simple. Elle consiste à comparer les échanges commerciaux entre deux pays avant et après la signature de leur accord, tout en tenant compte de ce qui se passe pour les autres paires de pays qui n'ont pas signé d'accord. Concrètement, après avoir rassemblé les données sur les flux commerciaux entre toutes les paires de pays du monde, sur plusieurs années, on crée une variable qui prend la valeur 1 quand deux pays ont un accord commercial en vigueur, et 0 sinon. En théorie, cela permet d'évaluer l'effet propre de l'accord sur les échanges, tout en contrôlant d'autres facteurs qui les déterminent aussi (distance, taille des économies, langue commune, etc.).

* Camilo Umana-Dajud est économiste au CEPII.

Encadré – Les modèles de gravité

- Pourquoi parle-t-on de « gravité » ?

L'idée est simple et intuitive, comme la loi de la gravitation de Newton : plus deux pays sont grands économiquement, plus ils commercent entre eux ; plus ils sont éloignés, moins ils échangent. C'est exactement comme deux planètes qui s'attirent d'autant plus qu'elles sont massives et proches.

Cette analogie n'est pas qu'une image : les modèles de gravité prédisent remarquablement bien les flux observés. Ils sont aujourd'hui utilisés non seulement pour le commerce de biens, mais aussi pour analyser les flux migratoires (où vont les travailleurs ?), les investissements directs étrangers (où les entreprises s'implantent-elles ?) ou encore les flux de capitaux entre pays.

- Comment mesure-t-on l'effet d'un accord commercial ?

À partir de bases de données massives qui couvrent les échanges entre tous les pays du monde sur plusieurs décennies, on note pour chaque paire de pays et chaque année la valeur de leurs échanges commerciaux (en dollars), leurs caractéristiques économiques (taille du PIB, population, niveau de développement), leurs caractéristiques géographiques et culturelles (distance, langue commune, frontière commune, passé colonial) et les politiques commerciales : présence d'un accord de libre-échange, union douanière, etc.

L'analyse statistique permet alors de répondre à la question suivante : « À caractéristiques égales, deux pays qui signent un accord commercial commercent-ils davantage que deux pays similaires sans accord ? »

Par exemple, si la France et l'Italie signent un accord, on compare l'évolution de leurs échanges avec celle d'autres paires de pays similaires (même taille économique, même distance) qui n'ont pas signé d'accord. On isole ainsi l'effet propre de l'accord en contrôlant pour tous les autres facteurs.

- Le rôle crucial des « effets fixes »

Pour affiner encore l'analyse, des « effets fixes » sont utilisés : des variables statistiques qui neutralisent toutes les particularités propres à chaque pays et à chaque paire de pays.

Concrètement, imaginons que l'on veuille évaluer l'effet d'un accord entre la France et l'Italie. Ces deux pays ont une longue histoire commune, des liens culturels profonds, une frontière partagée. Comment s'assurer que l'effet mesuré provient vraiment de l'accord, et non de tous ces facteurs difficilement observables ? Les effets fixes résolvent en partie ce problème : au lieu de comparer la France et l'Italie à d'autres paires de pays, on compare la France-Italie à elle-même, avant et après l'accord. On élimine ainsi tout ce qui est constant dans leur relation (histoire, géographie, culture) pour ne garder que ce qui change : l'accord lui-même.

Cette technique est extrêmement puissante. Elle permet de contrôler simultanément des milliers de facteurs impossibles à mesurer individuellement. Mais elle a un prix statistique élevé : en se concentrant uniquement sur les variations au sein de chaque paire de pays, elle réduit drastiquement l'information disponible. Avec des accords peu fréquents, cette exigence amplifie encore le problème de rareté des données.

Une méthode simple, en apparence seulement, car de nombreux pièges ont été identifiés et corrigés au fil des années, comme ceux liés aux flux commerciaux nuls (quand deux pays n'échangent pas du tout) ou au biais de sélection (les pays qui signent des accords ne sont pas choisis au hasard). Ces défis ont été surmontés grâce à des innovations méthodologiques (encadré).

Pourtant, cette méthode fait face à un défi majeur qui demeure largement sous-estimé : le problème de la rareté ! Imaginons une base de données avec 200 pays observés sur 10 ans, ce qui représente 398 000 observations de paires de pays ($200 \times 199 \times 10$ années). Prenons l'exemple de l'accord commercial entre les États-Unis et le Japon signé en 2020. Si nous observons cet accord 5 ans avant et 5 ans après sa signature, la variable « accord » ne prend la valeur 1 que pour 10 observations (1 paire \times 2 directions du commerce \times 5 années) sur les 398 000 observations totales. C'est comme tenter d'évaluer l'efficacité d'une nouvelle méthode pédagogique en ne l'appliquant qu'à deux élèves sur un campus de 40 000 étudiants, que l'on suivrait 5 ans avant et 5 ans après l'introduction de la méthode ($40\,000 \text{ étudiants} \times 10 \text{ années} = 400\,000$ observations, avec seulement 10 observations pour la nouvelle méthode pédagogique).

Ce problème de rareté se manifeste dès que la variable d'accord commercial prend la valeur 1 moins de 100 à 500 fois selon le contexte spécifique de l'analyse. Cela signifie qu'il peut affecter non seulement les accords récents, mais aussi des accords signés de longue date comme l'ALENA (signé en 1994 entre les États-Unis, le Canada et le Mexique) ou l'accord de libre-échange entre l'Union européenne et la Corée du Sud (entré en vigueur formellement fin 2015), qui ne concernent chacun qu'un nombre limité de paires de pays sur l'ensemble des observations disponibles.

■ Un piège statistique à double détente

Avec si peu d'observations où l'accord existe, l'incertitude statistique explose au point de rendre toute conclusion hasardeuse.

Il est important de souligner que le problème n'est pas la proportion d'observations avec un accord, mais bien le nombre absolu. Dans l'exemple de l'accord entre les États-Unis et le Japon, avec seulement 10 observations sur 5 ans, réduire la taille de l'échantillon à 50 pays au lieu de 200 ne résoudrait rien : on aurait toujours seulement 10 observations avec l'accord, et donc le même problème de rareté. Ce qui compte, c'est d'avoir suffisamment de cas où la politique est en vigueur pour pouvoir en mesurer l'effet avec précision.

Sinon, c'est comme chercher une aiguille dans une botte de foin : même quand l'aiguille existe bel et bien (l'accord a un effet positif), elle est si petite qu'il est extrêmement difficile de la trouver et de la mesurer avec précision. L'information pertinente (les quelques observations avec accord) se perd dans la masse des données (les dizaines de milliers d'observations sans accord).

Mais le piège peut également jouer dans l'autre sens, en créant de fausses certitudes. Imaginons que l'on observe les échanges entre les États-Unis et le Japon avant et après leur accord de 2020. Supposons que le commerce augmente fortement après la signature. Les tests statistiques pourraient indiquer que cet effet est « statistiquement significatif », c'est-à-dire qu'il est peu probable qu'il soit dû au hasard. Pourtant, avec seulement 10 observations où l'accord est en vigueur, cette hausse pourrait tout simplement refléter d'autres facteurs conjoncturels qui ne sont pas correctement contrôlés dans l'estimation.

Pour comprendre le piège statistique, imaginons que l'on lance une pièce de monnaie 5 fois et que l'on obtienne 4 fois « pile ».

Un test statistique mal conçu pourrait conclure que ce résultat est « significatif » et que la pièce est probablement truquée, car la probabilité d'obtenir 4 piles sur 5 lancers avec une pièce équilibrée n'est que de 15 %. Pourtant, avec seulement 5 lancers, cette conclusion serait hâtive : il est tout à fait possible qu'il s'agisse simplement du hasard. En revanche, si l'on avait lancé la pièce 100 fois et obtenu 80 fois « pile », on aurait alors une vraie preuve que la pièce est déséquilibrée. Le problème n'est pas le pourcentage (80 %), mais le nombre absolu d'observations qui permet de distinguer un vrai effet d'une fluctuation aléatoire.

De la même manière, avec seulement 10 observations pour un accord commercial, on peut trouver un « effet » qui semble statistiquement significatif, mais qui en réalité ne reflète qu'une fluctuation aléatoire ou des circonstances favorables temporaires. Les tests statistiques standards ne sont pas conçus pour gérer ces situations de rareté extrême : ils donnent une fausse impression de fiabilité et peuvent en conséquence conduire à des conclusions erronées. On risque alors d'attribuer à l'accord des effets qui n'existent pas, ou d'en surestimer considérablement l'ampleur.

Pour savoir ce qu'il en est vraiment, on recourt à des simulations. Le principe est le suivant : on crée des données artificielles où l'on connaît par construction l'effet d'un accord commercial, par exemple une augmentation de 10,5 %. On applique ensuite les méthodes statistiques utilisées dans les études et on compare l'effet estimé avec l'effet fixé au départ. En répétant ce processus des milliers de fois, on peut mesurer la fiabilité des estimations lorsque les données sont rares.

Les résultats sont sans appel et révèlent deux problèmes majeurs. Premier problème : l'instabilité extrême des estimations (graphique 1). Lorsque seulement 5 paires de pays ont signé un accord, l'estimation de l'effet peut varier considérablement d'une simulation à l'autre. Dans certaines simulations, parmi les 1 000 réalisées, on estime que l'accord baisse le commerce de 35 %, dans d'autres qu'il l'augmente de 80 %, alors que l'effet est, par hypothèse, de 10,5 %. D'ailleurs seules 18 % des simulations se situent entre 5,13 % et 16,18 % (zone noire sur le graphique), intervalle autorisant une marge d'erreur de 10 % autour de l'effet fixé (10,5 %). Cette volatilité rend toute conclusion hasardeuse : on ne sait plus si ce que l'on mesure reflète la réalité ou simplement le bruit statistique.

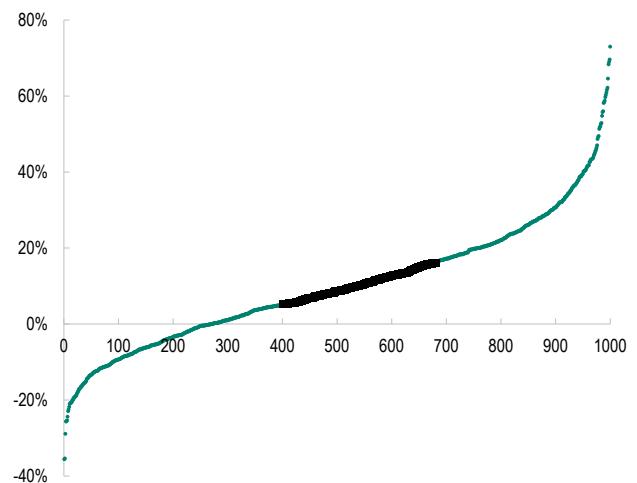
Second problème : la défaillance des tests de significativité. Les tests statistiques sont censés indiquer si un résultat est fiable ou non. Mais avec des données rares, les tests standards sont inadaptés : ils peuvent conclure qu'un effet est « statistiquement significatif », alors qu'en réalité l'estimation est très éloignée de la réalité. À l'inverse, ils peuvent rejeter un effet parce que l'incertitude est trop grande.

Cela parce que les tests statistiques classiques s'appuient sur des formules mathématiques dérivées de théorèmes qui ne fonctionnent que lorsque les échantillons sont grands : avec beaucoup de données, les moyennes et les proportions que l'on calcule suivent des distributions prévisibles et bien connues (comme la fameuse

« courbe en cloche »). Ces tests ont l'avantage de calculer rapidement la probabilité qu'un résultat soit dû au hasard. Mais quand les données sont rares (comme nos accords commerciaux), cette hypothèse ne tient plus. Avec si peu d'observations, les distributions réelles peuvent être très différentes des distributions théoriques supposées par les tests standards.

Graphique 1 – Une instabilité extrême des estimations

Estimations (1 000) de l'effet d'un accord commercial, fixé par hypothèse à 10,5 %, lorsque peu de pays ont un accord (0,34 % soit environ 5 paires)



Notes : Les effets estimés sont classés par ordre croissant et représentés en vert. La zone noire délimite les estimations conformes à l'effet fixé par hypothèse à 10,5 %, avec une marge d'erreur de 10 %.

Source : Umana-Dajud, C (2025).

La méthode du bootstrap adopte une approche différente : au lieu de se fier à des distributions qui supposent un monde idéal avec beaucoup de données, il simule directement des milliers de tirages possibles à partir des données observées.

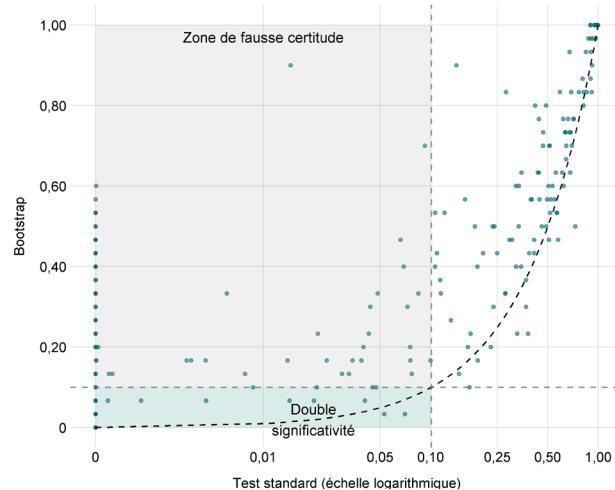
Concrètement, dans le cas où l'on dispose de données pour les échanges de 200 pays, dont seulement 5 paires avec un accord, le bootstrap va créer des milliers de « versions alternatives » de ces données en rééchantillonnant : on tire au hasard des observations parmi les données originales (en autorisant les remises, ce qui crée à chaque fois un nouvel échantillon de même taille). Pour chacun de ces échantillons alternatifs, on calcule l'effet estimé de l'accord. Après des milliers de répétitions, on obtient une distribution complète des résultats possibles à partir de laquelle on construit le test permettant de savoir si l'estimation est statistiquement significative.

Le graphique 2 illustre les résultats obtenus en recourant à ces deux façons de calculer la significativité des estimations : la méthode standard (utilisée dans la plupart des études) et le bootstrap.

Chaque point du graphique représente une simulation avec un nombre de paires de pays ayant un accord compris entre un et cinq. Si les deux méthodes donnaient les mêmes résultats, tous les points seraient alignés sur la courbe. Mais ce n'est pas ce que l'on observe.

Graphique 2 – Les tests statistiques standards donnent une fausse impression de certitude

Significativité statistique (valeur-p) des estimations avec les tests standards et le bootstrap



Notes : Chaque point représente une simulation dans laquelle le nombre de paires de pays ayant un accord est compris entre un et cinq. La courbe en pointillés représente la diagonale où les deux méthodes donnent le même résultat. Le cadran vert (en bas à gauche) correspond aux cas où les deux tests concourent à un effet significatif (valeur-p < 10 %) ; la zone griseée, les cas de « fausse certitude », où seuls les tests standards concourent à un résultat significatif.

Source : Umana-Dajud, C (2025).

Tout d'abord, on constate que de nombreux points se situent au-dessus de la courbe, ce qui signifie que les tests standards sont moins « exigeants » que le bootstrap : la valeur-p qui mesure la probabilité que l'effet estimé soit dû au hasard plutôt qu'à un véritable effet, est plus faible avec les tests standards. Et comme plus cette valeur est faible (proche de 0), plus on peut être confiant que l'effet observé n'est pas le fruit du hasard, les tests standards surestiment la fiabilité de l'estimation.

Ensuite, les tests concourent tous deux à un résultat significatif (valeur-p < 10 %, cadran vert sur le graphique 2) pour seulement 20 estimations sur 215, soit 9 %.

Enfin, on observe une zone particulièrement préoccupante, celle de la « fausse certitude » : de nombreux points (38 %) se situent dans la partie grisée du graphique, celle où la méthode standard indique un résultat significatif (valeur-p < 10 %), alors que le bootstrap conclut à

un résultat non significatif (valeur-p > 10 %). En d'autres termes, en utilisant les tests standards, on conclurait avec confiance que l'accord a un effet, alors qu'en réalité cette conclusion est fragile.

■ Que faire de ces résultats ?

Pour la recherche existante, les implications sont considérables. De nombreuses études sur les accords commerciaux, les traités d'investissement ou les politiques migratoires bilatérales pourraient être affectées. Des résultats supposés solides et significatifs pourraient ne pas l'être et les résultats non significatifs ne sont peut-être pas dus à l'absence d'effet, mais au problème de rareté. Si le bootstrap permet de disposer de distributions plus adaptées au problème de faible échantillon que les tests standards et d'être plus exigeant quant à la significativité des résultats, il ne fait pas non plus de miracle. En présence de très peu de pays ayant signé un accord depuis peu de temps, un effet non significatif ne permet pas de conclure que l'accord n'a pas d'effet. Avec davantage de données – par exemple un accord sur une plus longue période d'observation –, l'estimation de l'effet de l'accord sur le commerce pourrait être significative.

Pour ceux qui prennent les décisions, le message est clair : si une étude ne trouve pas d'effet d'un accord commercial, cela ne signifie pas forcément que l'accord est inutile. Les études empiriques ont leurs limites, surtout quand les données sont rares. Il est donc important de croiser différentes sources d'information : chiffres, études de cas, témoignages d'entreprises.

Pour ceux qui font les analyses, plusieurs bonnes pratiques s'imposent : indiquer clairement le nombre de paires de pays concernées par la politique étudiée et utiliser les techniques adaptées aux données peu nombreuses. La transparence sur ce qu'on peut et ne peut pas conclure est essentielle, tout comme la prudence à ne pas exagérer la portée des résultats. À la limite, il vaut parfois mieux renoncer à évaluer l'effet d'un accord spécifique, pour lequel la rareté des données est un obstacle à l'obtention d'un résultat fiable, et regrouper des accords pour augmenter le nombre d'observations et gagner en fiabilité, même si cela ne permet plus que d'avoir un effet moyen pour les différents accords regroupés.

La Lettre du
CEPII

© CEPII, PARIS, 2025

Centre d'études prospectives
et d'informations internationales
20, avenue de Ségur
TSA 10726
75334 Paris Cedex 07

contact@cepii.fr
www.cepii.fr – @CEPII_Paris
Contact presse : presse@cepii.fr

Le CEPII (Centre d'Études Prospectives et d'Informations Internationales) est le principal centre français d'étude et de recherche en économie internationale. Les analyses et études du Centre contribuent au débat public et à la formulation des politiques économiques en matière de politique commerciale, compétitivité, macroéconomie, finance internationale et croissance.

RÉDACTEURS EN CHEF :
ISABELLE BENSIDOUN
ANTOINE VATAN

ISSN 2493-3813

Décembre 2025

DIRECTEUR DE LA PUBLICATION :
ANTOINE BOUËT

Pour s'inscrire à
La Newsletter du CEPII :
www.cepii.fr/Resterinformé

RESPONSABLE DES PUBLICATIONS :
ISABELLE BENSIDOUN

RÉALISATION :
LAURE BOIVIN

Cette Lettre est publiée sous la responsabilité de la direction du CEPII.
Les opinions qui y sont exprimées sont celles des auteurs.

RECHERCHE ET EXPERTISE
SUR L'ÉCONOMIE MONDIALE

