Inférence causale et évaluation d'impact

Causal Inference and Impact Evaluation

Denis Fougère* et Nicolas Jacquemet**

Résumé – Cet article décrit de manière non technique les principales méthodes d'évaluation d'impact, expérimentales et quasi-expérimentales, et le modèle statistique qui les sous-tend. Sont ensuite recensés les articles utilisant ces méthodes que la revue *Economie et statistique / Economics and Statistics* a publiés durant ces quinze dernières années. Dans une seconde partie sont présentées certaines des avancées méthodologiques les plus importantes récemment proposées dans ce champ de recherche. Pour finir, l'accent est mis sur la nécessité d'être particulièrement attentif à la précision des effets estimés, mais aussi sur l'obligation de répliquer les évaluations, réalisées par expérimentation ou quasi-expérimentation, en vue de distinguer les faux-positifs des effets avérés.

Abstract – *This paper describes, in a non-technical way, the main impact evaluation methods, both experimental and quasi-experimental, and the statistical model underlying them. In the first part, we provide a brief survey of the papers making use of those methods that have been published by the journal* Economie et Statistique / Economics and Statistics *over the past fifteen years. In the second part, some of the most important methodological advances to have recently been put forward in this field of research are presented. To finish, we focus not only on the need to pay particular attention to the accuracy of the estimated effects, but also on the requirement to replicate évaluations, carried out by experimentation or quasi-experimentation, in order to distinguish false positives from proven effects.*

Codes JEL / JEL Classification: C1, C2, C3, C54
Mots-clés : inférence causale, méthodes d'évaluation, effets causaux
Keywords: causal inference, evaluation methods, causal effects

** CNRS, Sciences Po Paris-OSC/LIEPP, CEPR et IZA (denis.fougere@sciencespo.fr)*
*** Université Paris 1- Centre d'Economie de la Sorbonne et École d'Economie de Paris (nicolas.jacquemet@univ-paris1.fr)*

Nous remercions un relecteur anonyme pour ses commentaires qui ont permis d'enrichir notablement une première version de l'article. Ce projet bénéficie du soutien apporté par l'Agence nationale de la recherche (ANR) et l'État au titre du programme d'investissements d'avenir dans le cadre du Labex LIEPP (ANR-11-LABX-0091, ANR-11-IDEX-0005-02).

Citation : Fougère, D. & Jacquemet, N. (2019). Causal Inference and Impact Evaluation. *Economie et Statistique / Economics and Statistics*, 510-511-512, 181–200.
https://doi.org/10.24187/ecostat.2019.510t.1996

Au cours des vingt dernières années, le nombre de travaux d'évaluation d'impact, expérimentaux ou quasi expérimentaux, a crû de façon exponentielle. Ces méthodes permettent d'identifier, à partir de données individuelles d'enquête, des relations entre variables pouvant être rigoureusement interprétées comme des liens de cause à effet. Elles reposent sur des schémas d'observation et de recherche qui garantissent que les différences de résultat estimées (par exemple en termes de salaire, d'employabilité, de productivité ou de résultats scolaires) sont essentiellement dues à l'intervention ou à la politique mise en place, et que les biais de sélection et d'auto-sélection qui entachent nombre d'études empiriques sont notablement réduits, voire éliminés. Ces méthodes visent en particulier à identifier statistiquement les résultats dits contrefactuels, c'est-à-dire ceux qui seraient survenus en l'absence de la mise en place de l'intervention considérée. L'identification de l'effet causal de l'intervention sur la variable de résultat (son « impact ») est alors déduite de la comparaison avec les résultats observés pour les unités d'observation (chômeurs, salariés, entreprises, élèves, etc.) qui bénéficient de cette politique.

Une brève recension des techniques usuelles

Pour parvenir à cela, la méthode expérimentale la plus simple, qui consiste à tirer au sort les bénéficiaires de la politique évaluée et à comparer leur situation avec celle des individus ou entreprises que le tirage au sort a exclus, garantit la mise en évidence d'une relation de causalité entre la politique et l'effet observé, et ce sans que l'analyste ait besoin de faire des hypothèses par trop contraignantes. Les autres méthodes, dites quasi-expérimentales, cherchent à identifier des situations où, conditionnellement à un certain nombre de facteurs, le fait de bénéficier de l'intervention est indépendant des caractéristiques, observables ou non, des agents visés par cette intervention. Ces méthodes peuvent être regroupées en quatre catégories, qui sont présentées ci-dessous de manière non technique[1].

Les méthodes de variables instrumentales

Supposons que l'on observe les salaires de deux groupes de personnes, le premier groupe ayant récemment bénéficié d'une politique d'emploi telle qu'une période de formation continue, l'autre groupe n'en ayant pas bénéficié. Il est possible d'estimer par la méthode de la régression linéaire les effets de plusieurs variables caractéristiques des individus, tels que l'âge, le genre, la situation familiale, le niveau d'éducation, le lieu de résidence, etc., mais aussi l'effet du passage par une formation continue sur le salaire perçu au moment de l'enquête. Mais cette méthode simple risque de produire des estimations biaisées. Le problème est que l'accès à la période de formation n'est pas « exogène » : il peut non seulement être corrélé aux caractéristiques observées que nous venons de citer, mais aussi à des variables non observées par l'analyste, telles que le souhait de changer de profession, le goût pour l'apprentissage de nouvelles connaissances, la productivité du salarié évaluée par son employeur, etc. En conséquence, le fait d'être passé par une période de formation est vraisemblablement corrélé avec le terme d'erreur de la régression, terme d'erreur dont la valeur dépend généralement de ces caractéristiques non observées. Cette corrélation est la cause d'un biais dit d'endogénéité. Pour faire face à ce problème, les économètres ont longtemps utilisé la méthode des variables instrumentales. Par définition, une variable instrumentale doit jouer très significativement sur l'accès au programme évalué, ici la période de formation, mais ne pas directement affecter le niveau de salaire perçu après participation à ce programme. La méthode utilisée en ce cas est celle dite des doubles moindres carrés. La première étape consiste à régresser l'accès au programme sur l'ensemble des variables exogènes (âge, genre, etc.) mais aussi sur la valeur de la variable instrumentale (qui peut être, par exemple, la date d'une réforme significative des conditions d'accès à la formation continue). En un second temps, il faut régresser le salaire sur les mêmes variables exogènes et sur l'accès au programme de formation, non pas tel qu'il est effectivement observé, mais tel qu'il est prédit en tant que résultat de la première régression. Le coefficient associé à cette valeur « instrumentée » peut être interprété, sous certaines conditions très restrictives, comme « l'effet causal » du programme de formation sur le salaire des bénéficiaires.

Les méthodes d'appariement d'échantillons (*matching*)

Il s'agit ici avant tout de comparer bénéficiaires et non-bénéficiaires en neutralisant les différences liées aux caractéristiques observables. Ces méthodes reposent sur deux hypothèses. La

1. Ces méthodes sont décrites en détail, par exemple, dans l'ouvrage de Crépon & Jacquemet (2018), chapitre 9.

première stipule que l'affectation au groupe des bénéficiaires dépend exclusivement de caractéristiques exogènes observables et non des résultats anticipés de l'intervention : cette hypothèse est appelée hypothèse d'indépendance conditionnelle. La seconde hypothèse signifie que tout individu ou entreprise a une probabilité non nulle (comprise entre 0 et 1) d'être *a priori* bénéficiaire de l'intervention, quelles que soient ses caractéristiques, qu'il ou elle soit effectivement bénéficiaire ou non *a posteriori* : cette hypothèse est appelée hypothèse de superposition ou de support commun (*overlap assumption*). Ces deux hypothèses étant supposées valides, la méthode consiste à comparer le résultat de chaque bénéficiaire à la moyenne des résultats des non-bénéficiaires « proches » du point de vue des caractéristiques observables (âge, genre, niveau d'éducation, etc.), puis à faire la moyenne de tous ces écarts dans le groupe des bénéficiaires. La proximité au bénéficiaire considéré, i.e. le choix de ses « plus proches voisins », peut être réalisée à l'aide d'une distance (telle que la distance euclidienne ou celle de Mahalanobis), ou plus simplement encore à l'aide d'un score de propension, défini comme la probabilité d'être bénéficiaire de l'intervention compte-tenu des variables observables caractérisant l'individu ; cette probabilité peut être estimée en un premier temps, à l'aide par exemple d'un modèle logit ou probit, et ce indépendamment de la valeur des variables de résultat (*outcomes*) observées.

La méthode des doubles différences (*difference-in-differences*)

L'hypothèse sur laquelle repose cette méthode est simple. Supposons que l'on observe les variations entre deux dates d'une variable de résultat telle que le salaire au sein de deux groupes distincts. Le premier de ces groupes, appelé groupe cible ou groupe traité, bénéficie d'une intervention ou d'une politique d'emploi donnée ; le second, appelé groupe témoin ou groupe de contrôle[2], n'en bénéficie pas. La politique d'emploi est mise en place entre les deux dates considérées. La méthode repose sur une hypothèse stipulant qu'en l'absence de cette politique, l'évolution moyenne des salaires des individus du groupe traité aurait été identique à celle observée au sein du groupe de contrôle (hypothèse de « tendances parallèles », *parrallel trends)*. La validité de cette hypothèse, non vérifiable, peut être confortée par le fait qu'avant la mise en place de la politique, les salaires ont évolué de la même façon dans les deux groupes (hypothèse de *common pre-trends*). À l'inverse de la précédente, cette seconde hypothèse peut être testée à partir des données observées préalablement à la mise en place de l'intervention, à condition de disposer d'observations répétées au cours de cette période. Cette méthode exploite ainsi la dimension longitudinale (ou pseudo-longitudinale[3]) des données.

La méthode de la régression sur discontinuité

Cette méthode peut être appliquée lorsque l'accès à une intervention ou à une politique publique est conditionné par un seuil exogène fixé par les autorités en charge de cette politique. Ce seuil peut être une condition d'âge (pour un départ en retraite par exemple), un seuil de niveau d'emploi (par exemple, une politique de réduction des charges destinée aux entreprises de moins de vingt salariés) ou un niveau de ressources donnant l'accès à une bourse d'études ou à un crédit d'impôt. Dans sa forme la plus simple, la régression sur discontinuité permet de comparer la valeur moyenne de la variable de résultat dans le groupe des personnes bénéficiaires, par exemple celles dont le revenu ou l'âge est juste inférieur au seuil d'éligibilité fixé, et la valeur moyenne de cette variable dans le groupe de contrôle comparable, formé des personnes dont le revenu ou l'âge est juste supérieur à ce seuil. L'hypothèse sous-jacente est que, pour des personnes ayant par ailleurs les mêmes caractéristiques du point de vue de la qualification, du niveau d'éducation ou du genre, celles situées juste en-dessous et au-dessus du seuil sont identiques. Seul un pur aléa, telle qu'une date de naissance, les distingue. Dans ces conditions, une simple différence entre les moyennes de la variable de résultat (par exemple, le niveau de salaire ou d'éducation après mise en œuvre de la politique) permet d'estimer l'effet causal de l'intervention considérée. Cette différence n'est toutefois qu'une mesure locale, au voisinage du seuil, et son extrapolation à des niveaux de revenu ou des âges éloignés de ce seuil n'a pas de validité scientifique. Pour cette raison, on dit que la régression sur discontinuité permet d'estimer un effet local moyen (*local average treatment effect,* discuté en détail plus bas).

Chaque type de méthode correspond donc à des hypothèses bien spécifiques. En pratique, notamment lorsqu'il n'est pas possible de conduire une expérience randomisée, il importe de reconnaître l'information dont dispose l'analyste et de savoir

2. Ces dénominations sont les mêmes dans chacune des méthodes d'inférence causale utilisées.

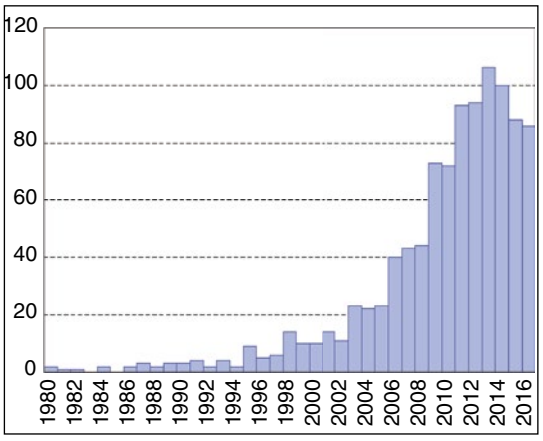
3. Les observations peuvent ne pas être celles des mêmes individus mais de répétitions d'échantillons aléatoires d'une même population et former un « pseudo panel ».

laquelle de ces hypothèses est la plus vraisemblable afin de choisir la méthode la mieux adaptée aux données disponibles. Depuis l'article pionnier de LaLonde (1986), plusieurs travaux ont été consacrés à la comparaison des évaluations réalisées à l'aide de méthodes expérimentales et quasi-expérimentales, et notamment aux biais d'estimation pouvant résulter de l'emploi des méthodes quasi-expérimentales. Par manque de place, il ne nous est pas possible de résumer ici les résultats de ces comparaisons. Sur ce sujet, le lecteur intéressé pourra consulter, par exemple, les articles de Glazerman *et al.* (2003), Hill (2008), Chabé-Ferret (2015), Wong *et al.* (2017), et Chaplin *et al.* (2018).

Une littérature scientifique internationale en plein essor

Ces méthodes ont été appliquées dans de nombreux champs de recherche. Par exemple, dans le domaine des politiques éducatives, le nombre d'expérimentations aléatoires contrôlées (« *randomized controlled trials* », ou RCT) ayant donné lieu à des publications internationales est passé de quelques unités en 1980 à plus de 80 par an depuis 2010 (figure I). Les évaluations quasi expérimentales ont suivi une tendance similaire et l'ensemble constitue aujourd'hui ce que certains ont appelé une « révolution empirique »[4]. Ces travaux et les évaluations chiffrées qu'ils contiennent sont des ressources de première importance pour le choix, la conception et la mise en œuvre des politiques publiques.

Figure I
Nombre d'expérimentations aléatoires contrôlées réalisées entre 1980 et 2016 dans le domaine des politiques éducatives, ayant donné lieu à une publication scientifique internationale, d'après Connolly *et al.* (2018)



Un autre indice de la maturité et de la richesse des méthodes économétriques d'évaluation est la publication récente de plusieurs ouvrages de référence. Parmi ceux-là, citons les livres d'Imbens & Rubin (2015), Lee (2016) et Frölich & Sperlich (2019), qui font suite aux articles de synthèse d'Angrist & Krueger (1999), Heckman *et al.* (1999), Heckman & Vytlacil (2007a, 2007b), Abbring & Heckman (2007), et Imbens & Wooldrige (2009). Le *Handbook of Fields Experiments* édité par Duflo & Banerjee en 2017 est l'ouvrage de référence sur les expérimentations randomisées de terrain. Pour les expérimentations en laboratoire, l'ouvrage de Jacquemet & L'Haridon (2018) est la référence la plus récente. Enfin, la liste des articles consacrés aux méthodes d'inférence causale et publiés ces trente dernières années dans les meilleures revues internationales d'économie ou de statistique est trop longue pour être reprise ici. Le lecteur intéressé la trouvera dans les bibliographies des ouvrages cités ci-dessus. Des synthèses en langue française, plus ou moins formalisées, sont également disponibles. Parmi celles-ci, citons les articles de Brodaty *et al.* (2007), Givord (2014) et Chabé-Ferret *et al.* (2017).

De nombreux travaux d'évaluation ont été publiés dans *Économie et Statistique*

La revue *Économie et Statistique* (alors pas encore « / *Economics and Statistics* ») a accompagné cette progression et ces avancées tout au long de ces vingt dernières années, en publiant avec une fréquence soutenue des articles appliquant les méthodes économétriques d'évaluation à des données françaises, principalement produites par les services de la statistique publique. Certains de ces articles ont trouvé un réel écho dans le débat citoyen. Il est certes risqué d'en établir la liste exhaustive, certaines de ces publications ayant pu échapper à notre attention. Il est toutefois possible d'en citer quelques-uns en les regroupant en fonction des méthodes utilisées.

La technique des variables instrumentales a été utilisée par Crépon *et al.* (2004) pour mesurer les effets de la réduction du temps de travail sur la productivité et l'emploi dans les entreprises. Leclair & Roux (2007) l'ont ensuite mobilisée pour mesurer la productivité relative et l'utilisation des emplois de courte durée dans les entreprises. Des variables instrumentales ont été également utilisées par Beffy *et al.* (2009) pour estimer les effets du travail salarié des étudiants

4. *Angrist & Pischke (2010).*

sur la réussite et la poursuite des études universitaires, et par Fougère & Poulhès (2014) pour étudier l'influence de la propriété sur le portefeuille financier des ménages.

Le lecteur trouvera des applications de la méthode des doubles différences dans plusieurs articles de la revue. Les premières publications ayant mis en œuvre cette méthode sont les articles de Bénabou *et al.* (2004), consacré à l'évaluation des zones d'éducation prioritaire (ZEP), et de Behaghel *et al.* (2004), dont le but était d'estimer les effets de la contribution Delalande sur les transitions des salariés entre emploi et chômage. Fack & Landais (2009) l'ont utilisée pour évaluer l'efficacité des incitations fiscales aux dons. Carbonnier (2009) a évalué les conséquences incitatives et redistributives des incitations fiscales portant sur l'emploi d'un salarié à domicile. La méthode a permis à Bozio (2011) de mesurer l'impact de l'augmentation de la durée d'assurance ayant fait suite à la réforme des retraites de 1993. Geniaux & Napoleone (2011) ont utilisé une méthode de doubles différences couplée avec une méthode d'appariement pour évaluer les effets des zonages environnementaux sur la croissance urbaine et l'activité agricole. Toujours grâce à la méthode des doubles différences, Simonnet & Danzin (2014) ont évalué l'effet du RSA sur le retour à l'emploi des allocataires, puis Bérard & Trannoy (2018) ont mesuré l'impact de la hausse des droits de mutation immobiliers de 2014 sur le marché du logement français.

Parmi les articles ayant mis en application les méthodes d'appariement (*matching*), notons particulièrement ceux de Crépon & Desplatz (2001) qui ont utilisé une méthode de ce type pour estimer les effets des allègements de charges sociales sur les bas salaires, d'Even & Klein (2007) qui ont estimé les effets à moyen terme des contrats aidés sur l'emploi des bénéficiaires, de Rathelot & Sillard (2008) qui ont évalué les effets de la politique des zones franches urbaines (ZFU) sur l'emploi salarié et les créations d'établissements, et de Bunel *et al.* (2009) qui ont consacré leur étude aux effets des allégements de cotisations sociales sur l'emploi et les salaires.

La méthode de la régression sur discontinuité a été utilisée pour la première fois dans *Économie et Statistique* par Lorenceau (2009) afin d'évaluer les effets des baisses de charges salariales accordées dans le cadre des zones de revitalisation rurale sur la création d'établissements et sur le volume d'emploi. Elle a également été utilisée par Baraton *et al.* (2011) pour évaluer les effets de la réforme de 2003 sur les départs en retraite des enseignants du second degré public.

À notre connaissance, *Economie et Statistique / Economics and Statistics* n'a pas publié d'articles relatifs à des expérimentations randomisées *stricto sensu*. Ceci ne signifie pas que les économistes français n'ont pas produit de recherches de grande qualité en ce domaine. Bien au contraire, sous l'influence et parfois avec la collaboration d'Esther Duflo, professeure d'économie au M.I.T., les économistes français ont publié dans les meilleures revues internationales des articles consacrés à des expérimentations randomisées, notamment dans le domaine des politiques d'emploi ou d'éducation. Le lecteur en trouvera des exemples notables dans les travaux de Crépon *et al.* (2013 ; 2015), d'Avvisati *et al.* (2014), de Goux *et al.* (2017), ou encore de Barone *et al.* (2019). *Économie et Statistique* a toutefois publié trois articles consacrés à des expériences de *testing*, qui sont certes des expériences aléatoires mais qui ne peuvent être assimilées à des expérimentations randomisées de terrain. Le *testing* est une forme d'expérimentation sociale en situation réelle destinée à déceler une situation de discrimination. Dans le cas le plus simple, le statisticien compare le comportement d'un tiers, en général un employeur ou un bailleur, envers deux personnes ayant exactement le même profil pour toutes les caractéristiques pertinentes, à l'exception de celle que l'on soupçonne de donner lieu à discrimination, par exemple une origine ethnique, un handicap, une religion, un âge, un genre, une orientation sexuelle, etc. L'article de Petit *et al.* (2011) consacré aux effets du lieu de résidence sur l'accès à l'emploi, ainsi que ceux de Petit *et al.* (2013) et Edo & Jacquemet (2013) portant sur les effets du genre et de l'origine sur la discrimination à l'embauche, sont particulièrement représentatifs de ce type d'approche, dont les limites, tant méthodologiques que conceptuelles, ont été rappelées par Aeberhardt *et al.* (2011) dans un commentaire publié dans la revue à la suite de l'article de Petit *et al.* (2011).

La liste des publications, notamment internationales, utilisant des méthodes statistiques d'inférence causale s'allonge chaque jour. À côté des études les mettant directement en application avec des données expérimentales ou quasi-expérimentales, beaucoup de travaux ont été consacrés ces dix dernières années à raffiner ces méthodes, ou à proposer des solutions permettant de dépasser certaines de leurs limites. Le reste de cet article est consacré à la présentation des développements qui nous semblent en ce domaine particulièrement prometteurs. Faute de place, nous n'avons pu

aborder ici tous les thèmes émergents, notamment celui des interactions sociales et des interférences dans les expérimentations randomisées. Ce sujet jusqu'alors malheureusement sous-estimé est, par exemple, abordé dans les articles d'Hudgens & Halloran (2008), Aronow (2012), Manski (2013), Liu & Hudgens (2014), et Baird *et al.* (2018). Une recension extensive des avancées récentes et des voies de recherche futures peut être trouvée dans les articles d'Athey & Imbens (2017a, 2017b) et d'Abadie & Cattaneo (2018).

Le modèle canonique de l'évaluation d'impact

Dès sa formulation originelle par Rubin (1974), le modèle canonique d'évaluation d'impact met l'accent sur l'hétérogénéité de la réponse des agents économiques à la suite d'une intervention les concernant[5]. Dans ce modèle, chaque unité d'observation est caractérisée par deux « résultats potentiels » qui lui sont propres : y_{i0} est le résultat qui serait observé en l'absence d'intervention pour l'unité i et y_{i1} celui qui serait observé pour cette même unité par suite de l'intervention. Pour chaque unité, seul un de ces deux effets est observé. Plutôt qu'à un « effet causal », l'intervention est donc associée à une distribution de changements de situation $\Delta_i = y_{i1} - y_{i0},\ i = 1,\ldots,N,\ N$ étant ici la taille de l'échantillon. La démarche d'évaluation nécessite donc de choisir le paramètre de cette distribution que l'analyste souhaite identifier. Parmi les paramètres résumant la distribution de l'effet de l'intervention (ou du traitement), les plus communs sont l'effet moyen du traitement et l'effet moyen du traitement sur les traités.

L'effet moyen du traitement (*Average Treatment Effect*, ou ATE) correspond à l'espérance mathématique de cette distribution : il mesure donc le changement moyen de situation pour un individu tiré au hasard dans la population. L'effet moyen du traitement sur les traités (*Average Treatment effect on the Treated*, ou ATT) est quant à lui spécifique à la sous-population des individus qui bénéficient effectivement du programme (et correspond formellement à l'espérance conditionnelle au fait d'être effectivement traité). Les deux paramètres ne sont égaux que sous des hypothèses très restrictives. Ils concordent par exemple de manière triviale si l'intervention concerne l'ensemble de la population (c'est, par exemple, le cas d'une augmentation de l'âge minimum de sortie du système scolaire, mesure qui concerne tous les élèves), ou si le traitement est supposé agir de la même façon sur tous les individus ($\Delta_i = \Delta,\ i = 1,\ldots,N$). Dans toutes autres circonstances, ces deux paramètres sont distincts. Ils informent de manière différente sur la distribution de l'effet causal : l'effet moyen du traitement sur les traités mesure l'efficacité du programme au travers du changement de situation des bénéficiaires, tandis que l'effet moyen du traitement indique quelle serait son efficacité si le programme était généralisé à l'ensemble de la population. La méthode d'évaluation choisie conditionne fortement le paramètre pouvant être mesuré. Les expériences randomisées permettent d'estimer l'ATE à condition que l'affectation aléatoire aux groupes soit réalisée dans l'ensemble de la population et que tous les individus sélectionnés pour participer à l'expérience y participent effectivement. Toutefois, elles permettent d'estimer uniquement l'ATT lorsque certains des individus sélectionnés refusent de participer à l'expérience, ou plus généralement lorsque seul est observé un sous-échantillon non aléatoire de l'échantillon prélevé (pour une illustration, voir Chabé-Ferret *et al.*, 2017). L'estimateur en doubles différences ou les estimateurs par appariement mesurent quant à eux le changement de situation spécifique aux bénéficiaires, à savoir l'ATT.

Au-delà de l'importance du choix du paramètre à estimer (qui doit primer sur le choix de la méthode d'identification), l'hétérogénéité de l'effet du traitement constitue une limite importante à la capacité à généraliser les effets estimés d'une intervention dans le cadre d'une étude empirique particulière (voir plus loin).

L'effet local moyen du traitement (LATE, ou *local average treatement effect)*

Depuis les travaux d'Imbens & Angrist (1994) qui ont introduit l'estimateur LATE (*local average treatment effect*), l'interprétation de l'estimateur par variable instrumentale comme « effet moyen du traitement sur les traités » est remise en question. Elle n'est valide que si l'effet du programme est le même pour tous les individus, quelles que soient leurs caractéristiques d'âge, de genre, d'expérience, etc., ce qui est évidemment une hypothèse fort peu réaliste. Imbens & Angrist (1994), et nombre d'économètres à leur suite, montrent que dans le cas où l'effet d'une intervention ou d'une politique publique est susceptible de varier d'un groupe d'individus à l'autre, et plus généralement d'être hétérogène au sein d'une population, seul peut être produit un estimateur

5. Ce modèle est différent du modèle introduit par Judea Pearl qui utilise le formalisme des graphes orientés acycliques, ou directed acyclic graphs, *souvent utilisés en épidémiologie ou en psychométrie (cf. Peters* et al., *2017, ou Pearl & Mackenzie, 2018).*

local pour ceux des individus qui décideraient d'être bénéficiaires du programme lorsque celui-ci deviendrait accessible à la suite d'une variation de l'instrument. Ces individus sont appelés les *compliers*, terme qui n'a pas de traduction directe en français, sauf à dire, de manière plus ou moins satisfaisante, qu'il s'agit des personnes qui se conforment ou adhèrent au programme lorsque la valeur de l'instrument évolue. Le groupe des *compliers* est vraisemblablement mieux défini lorsque lui sont opposés les personnes qui refusent systématiquement le programme (*never-takers*) et celles qui sont toujours prêtes à y participer (*always-takers*), quelle que soit la valeur de l'instrument. La mise en œuvre de l'estimateur LATE suppose qu'il n'existe pas d'individus prêts à participer au programme lorsque celui-ci n'est pas proposé, mais qui refuseraient de le faire une fois le programme introduit. Ce groupe de personnes, appelé les *defiers*, est supposé ne pas exister : cette hypothèse correspond à ce qu'Imbens & Angrist (1994) nomment l'hypothèse de *monotonicité*. L'estimateur LATE mesure donc l'effet de l'intervention pour le seul groupe des *compliers*, qui n'est malheureusement pas toujours identifiable. Lorsqu'il l'est, notamment dans le cas où une loterie ou une procédure aléatoire modifie l'affectation au traitement (i.e. à l'intervention ou au programme proposé), l'estimateur LATE peut être obtenu à l'aide des doubles moindres carrés. Angrist & Imbens (1995) proposent une méthode plus générale permettant de tenir compte de l'effet d'autres variables exogènes (telle que l'âge) dans le cadre de la mise en œuvre du LATE. Angrist *et al.* (2000) appliquent cette approche à l'estimation des modèles à équations simultanées.

La validité externe des méthodes d'évaluation d'impact

Plusieurs des méthodes qui ont été citées sont caractérisées par une forte validité interne : elles permettent d'obtenir des estimateurs crédibles des effets moyens des interventions pour les échantillons considérés. La possibilité d'extrapoler leurs résultats à une population plus large, i.e. leur validité externe, est toutefois souvent questionnée.

Dans le cas des expérimentations randomisées, cette critique tient au fait que les échantillons sont généralement d'assez faible taille et concernent des groupes particuliers, par exemple des personnes vivant dans des environnements ou présentant des caractéristiques spécifiques ; ils ne sont pas représentatifs de la population dans son ensemble, tout au moins de la totalité des personnes potentiellement éligibles. La question de la validité externe est fondamentalement liée à celle de l'hétérogénéité des effets des interventions (voir ci-dessous). Supposons que l'on conduise une expérimentation dans un cadre A, qui peut concerner une localité, une période, ou une sous-population d'individus donnée. En quoi les estimations des effets de cette intervention particulière conduite dans ce cadre particulier nous informent-elles de ce que seraient les effets de la même intervention dans une autre localité, à un autre moment, pour un groupe d'individus différent, c'est-à-dire dans un cadre B différent de A ? Les différences peuvent provenir de caractéristiques observées et non observées de ces autres localités, périodes, individus, et éventuellement d'une modification, même légère, des modalités de l'intervention. Pour répondre à ces questions, il est utile d'avoir accès aux résultats de multiples expérimentations, menées dans des cadres différents, et si possible, avec des échantillons d'assez grande taille et représentatifs de la population éligible (au moins du point de vue des principales caractéristiques observables). Un exemple particulièrement intéressant est celui de la microfinance. Meager (2019) a analysé les résultats de sept expérimentations conduites sur ce thème, et a constaté que les effets estimés étaient remarquablement cohérents.

Une autre approche consiste à tenir explicitement compte des différences entre les distributions des caractéristiques spécifiques aux groupes ou aux périodes considérées. Hotz *et al*. (2005) et Imbens (2010) proposent un cadre théorique dans lequel les différences d'effets constatées au sein d'un groupe de plusieurs localités proviennent du fait que les unités établies dans ces localités ont des caractéristiques différentes. Au moyen d'une procédure d'ajustement qui consiste en une repondération des unités individuelles (personnes, ménages, entreprises, etc.), ils peuvent comparer les effets de l'intervention considérée dans ces localités différentes. Cette technique est proche des méthodes de pondération inverses de probabilité (*inverse probability weighting*)[6] préconisées par Stuart et ses co-auteurs (Imai *et al.*, 2008 ; Stuart *et al.*, 2011 ; Stuart *et al.*, 2015).

Rappelons que l'estimateur par variables instrumentales est souvent interprété comme un estimateur local de l'effet moyen du traitement, c'est-à-dire comme un estimateur LATE qui mesure l'effet moyen du traitement pour ceux

6. La pondération inverse de probabilité est une technique statistique permettant de calculer des statistiques standardisées pour une pseudo-population différente de celle dans laquelle les données ont été collectées.

des individus, les *compliers*, dont l'affectation au traitement est modifiée par une variation de la valeur de l'instrument. Sous quelles conditions cet estimateur peut-il être interprété comme l'effet moyen du traitement dans la population totale ? En d'autres termes, quelles sont les conditions qui assurent sa validité externe ? Il existe deux groupes qui ne sont jamais affectés par la variable instrumentale, les *always-takers* qui reçoivent toujours le traitement, et les *never-takers* qui ne le reçoivent jamais. Pour répondre à la question, Angrist (2004) suggère de tester si la différence entre les résultats moyens des *always-takers* et des *never-takers* est égale à l'effet moyen du traitement sur le résultat des *compliers.* Angrist & Fernandez-Val (2013) cherchent à exploiter une condition d'ignorabilité (*conditional effect ignorability*) stipulant que, conditionnellement à certaines variables exogènes, l'effet moyen pour les *compliers* est identique à l'effet moyen pour les *always-takers* et les *never-takers*. Bertanha & Imbens (2019) suggèrent de tester la combinaison de deux égalités, à savoir celle du résultat moyen des *compliers* non traités et du résultat moyen des *never-takers*, et l'égalité du résultat moyen des *compliers* traités au résultat moyen des *always-takers*.

Dans le cas de la régression sur discontinuité, l'absence de validité externe provient principalement du fait que cette méthode produit des estimateurs locaux, qui ne sont valides qu'au voisinage du seuil d'éligibilité considéré. Si ce seuil est par exemple une condition d'âge, la régression sur discontinuité ne permet pas d'inférer ce que serait l'effet moyen de l'intervention pour des personnes dont l'âge diffère fortement de l'âge définissant le seuil d'éligibilité. Sous quelles conditions peut-on généraliser les estimations d'effets obtenus avec la régression sur discontinuité ? Dong & Lewbel (2015) font remarquer que dans beaucoup de cas, la variable qui définit le seuil d'éligibilité (appelée « variable de forçage » ou *forcing variable*) est une variable continue telle que l'âge ou le niveau de revenu. Ces auteurs font remarquer qu'en ce cas, au-delà de l'ampleur de la discontinuité de la variable de résultat au voisinage du seuil, il est également possible d'estimer la variation de la dérivée première de la fonction de régression, et même de fonctions dérivées d'ordre supérieur. Ceci permet d'extrapoler les effets causaux du traitement pour des valeurs de la variable de forçage plus éloignées du seuil d'éligibilité. Angrist & Rokkanen (2015) proposent de tester si, conditionnellement à des variables exogènes additionnelles, la corrélation entre la variable de forçage et la variable de résultat disparaît. Un tel résultat signifierait que l'affectation au traitement pourrait être considérée comme indépendante des résultats potentiels (*unconfoundedness property*)[7] conditionnellement à ces variables exogènes additionnelles, ce qui permettrait une fois encore d'extrapoler le résultat pour des valeurs de la variable de forçage plus éloignées du seuil. Bertanha & Imbens (2019) proposent une approche fondée sur la régression floue sur discontinuité[8]. Ils suggèrent de tester la continuité de l'espérance conditionnelle de la variable de résultat, pour une valeur donnée du traitement et de la variable de forçage au niveau du seuil, ajustée par les variations des caractéristiques exogènes.

Doubles différences et contrôle synthétique

Comme rappelé précédemment, la mise en œuvre des doubles différences suppose que l'on dispose d'un groupe de contrôle dont l'évolution au cours du temps reflète celle qu'aurait connue le groupe de traitement en l'absence d'intervention. Cette hypothèse ne peut être testée sur la période qui suit l'intervention, au cours de laquelle les différences de résultat entre groupes reflètent également l'effet de la politique. Une composante testable de cette hypothèse est que l'évolution passée de la variable de résultat (avant mise en œuvre de la politique évaluée) est en moyenne similaire à celle de cette même variable dans le groupe de traitement. Lorsqu'elle est rejetée, il est possible de créer par un système adéquat de pondérations une unité de contrôle artificielle, dite *contrôle synthétique*, à partir des observations du groupe de contrôle. Ce contrôle synthétique est construit de telle sorte que l'évolution passée de la variable de résultat en son sein soit identique à celle de cette variable dans le groupe de traitement.

La méthode a été introduite par Abadie & Gardeazabal (2003) dans une étude visant à évaluer l'effet de l'activité terroriste de l'ETA sur l'évolution du PIB du pays basque entre 1975 et 2000, période caractérisée par l'intensité et la fréquence des actes violents commis par cette organisation. Le problème est qu'entre 1960 et 1969, décennie qui a précédé le début de la période d'activité terroriste, le PIB de la région basque a évolué de manière très différente de la moyenne des PIB des seize autres régions

7. *«* The unconfoundedness assumption states that assignment is free from dependence on the potential outcomes *» (Imbens & Rubin, 2015, p. 257).*

8. *La régression sur discontinuité stricte (*sharp regression discontinuity design*) correspond au cas où nul ne peut déroger à la contrainte du seuil d'éligibilité. À ce cas, s'oppose celui de la régression sur discontinuité floue (*fuzzy regression discontinuity design*) dans lequel on observe des individus traités, ou des individus non traités, des deux côtés du seuil.*

espagnoles, conduisant au rejet de l'hypothèse de tendance commune pré-traitement. Abadie & Gardeazabal (2003) proposent alors de construire une région de contrôle synthétique dont l'évolution du PIB entre 1960 et 1969 serait similaire à celle du PIB du pays basque. Cela peut être réalisé en minimisant la distance entre les observations annuelles du PIB basque entre 1960 et 1969 et celles de cette région synthétique. De manière plus formelle, les valeurs annuelles du PIB dans le pays basque entre 1960 et 1969 sont notées $y_{1,t}$ ($t = 1960,\ldots,1969$) et regroupées dans un vecteur $\boldsymbol{Y}_{1,0} = [Y_{1,1960} \ldots Y_{1,1969}]$. De façon analogue, les observations annuelles du PIB dans chacune des seize autres régions espagnoles sont notées $Y_{j,t}$ $\left(j = 2,\ldots,17; t = 1960,\ldots,1969\right)$ et rangées dans une matrice notée $\boldsymbol{Y}_{0,0}$ de dimension $\left(10 \times 16\right)$. La région de contrôle synthétique est construite à partir d'un vecteur de pondérations $\boldsymbol{w} = \left[w_1,\ldots,w_{16}\right]'$ de dimension $\left(16 \times 1\right)$ qui minimise la norme euclidienne pondérée suivante pour une matrice $\boldsymbol{V}$ donnée :

$$\left\|\boldsymbol{Y}_{1,0} - \boldsymbol{Y}_{0,0}\boldsymbol{w}\right\| = \sqrt{\left(\boldsymbol{Y}_{1,0} - \boldsymbol{Y}_{0,0}\boldsymbol{w}\right)' \boldsymbol{V} \left(\boldsymbol{Y}_{1,0} - \boldsymbol{Y}_{0,0}\boldsymbol{w}\right)}$$

Dans une première application simple, Abadie & Gardeazabal (2003) choisissent pour matrice $\boldsymbol{V}$ la matrice identité. Cela leur permet de trouver aisément le système de pondérations $\boldsymbol{w}^*$ qui minimise cette norme[9]. Ils vérifient que les dix PIB annuels de cette région synthétique, calculés comme $\boldsymbol{Y}^*_{0,0} = \boldsymbol{Y}_{0,0} \times \boldsymbol{w}^*$ au cours de la période 1960-1969, sont similaires aux PIB de la région basque observés durant la même période. Cela leur permet de calculer ensuite les PIB contrefactuels à ceux de la région basque durant la période d'activité terroriste, 1975-2000. Ces PIB contrefactuels sont notés $\boldsymbol{Y}^*_{0,1}$ et calculés dans le vecteur de dimension $\left(26 \times 1\right)$ $\boldsymbol{Y}^*_{0,1} = \boldsymbol{Y}_{0,1} \times \boldsymbol{w}^*$, où $\boldsymbol{Y}_{0,1}$ est la matrice de dimension $\left(26 \times 16\right)$ qui regroupe les observations des 26 PIB annuels[10] de chacune des seize régions espagnoles autres que le pays basque. L'effet causal du terrorisme sur le PIB basque est alors mesuré comme $\boldsymbol{Y}_{1,1} - \boldsymbol{Y}^*_{0,1}$, où $\boldsymbol{Y}_{1,1}$ est la matrice de dimension $\left(26 \times 1\right)$ qui regroupe les 26 observations annuelles du PIB basque de 1975 à 2000.

En général, $\boldsymbol{V}$ est une matrice diagonale dont les éléments diagonaux sont non négatifs. Dans une version élargie de cette méthode, Abadie & Gardeazabal (2003) et Abadie *et al.* (2010 ; 2015) proposent de choisir des matrices $\boldsymbol{V}$ dont les éléments sont fondés sur les données (*data-driven*). Le nombre d'unités traitées peut être supérieur à l'unité : en ce cas, il faut calculer un contrôle synthétique pour chaque unité traitée. Toutefois, dans le cas où le nombre d'unités traitées est très grand, il est possible que le contrôle synthétique d'une unité traitée ne soit pas unique. Abadie & L'Hour (2019) proposent une variante tenant compte de cette difficulté. Leur estimateur s'écrit :

$$\left\|\boldsymbol{Y}_{1,0} - \boldsymbol{Y}_{0,0}\boldsymbol{w}\right\|^2 + \lambda \sum_{j=2}^{J+1} w_j \left\|\boldsymbol{Y}_{j,0} - \boldsymbol{Y}_{1,0}\right\|^2, \text{ avec } \lambda > 0.$$

Dans cette expression, $\boldsymbol{Y}_{j,0}$ est le vecteur dont les éléments sont les valeurs observées de la variable de résultat pour l'unité de contrôle j $\left(j = 2,\ldots,J+1\right)$ au cours de chacune des périodes qui précèdent la mise en œuvre de l'intervention. L'estimateur proposé par Abadie & L'Hour (2019) inclut une pénalisation λ pour les écarts entre les valeurs de la variable de résultat d'une unité traitée et celles de chaque unité de contrôle au cours de la période précédant la mise en œuvre de l'intervention. Abadie & L'Hour (2019) montrent que, sous ces conditions, et sauf dans quelques cas spécifiques, leur estimateur fournit un contrôle synthétique unique.

Des versions élargies de l'estimateur du contrôle synthétique ont été également proposées par Amjad *et al.* (2018) et Athey *et al.* (2018) qui suggèrent d'utiliser des techniques de complétion de matrices, mais aussi par Hahn & Shi (2017) qui fondent leur approche sur des méthodes d'inférence fondées sur les échantillons (*sampling-based inferential methods*).

Le rôle et le choix des variables explicatives

Quel que soit le type d'intervention ou de méthode d'évaluation choisie par l'analyste, les individus, ménages, entreprises, etc., qui sont échantillonnés, bénéficiaires ou non de l'intervention, membres du groupe cible (i.e. de traitement) ou du groupe témoin (i.e. de contrôle), peuvent toujours différer du point de vue de certaines caractéristiques exogènes (telles que l'âge, le genre, le nombre d'années d'expérience sur le marché du travail, etc., pour des individus, ou le nombre de salariés, la date de création, le niveau d'endettement à court terme, etc., pour une entreprise). Dans le cas d'une expérimentation contrôlée randomisée non stratifiée ou d'une régression sur discontinuité stricte, une régression simple de la variable de résultat observée sur une constante et une variable indicatrice de l'appartenance au

9. Les seules régions qui ont des poids bien supérieurs à zéro sont Madrid et la Catalogne.
10. 2000 – 1974 = 26 années.

groupe de traitement suffit à obtenir un estimateur convergent de l'effet moyen de ce traitement dans l'échantillon. L'ajout à cette régression des variables exogènes a en théorie pour principal effet d'améliorer la précision de l'estimateur de l'effet moyen du traitement.

Mais dans des cas autres que la randomisation non stratifiée ou que la régression sur discontinuité stricte, il est nécessaire pour obtenir des estimateurs convergents d'ajouter des hypothèses relatives au rôle des variables exogènes. L'hypothèse la plus communément utilisée est celle de l'indépendance conditionnelle. Celle-ci stipule que l'affectation au groupe de traitement, représentée par une variable aléatoire T, et les résultats potentiels de l'intervention, notés y_{1i} pour un individu traité et y_{0i} pour un individu non traité, sont indépendants conditionnellement à l'ensemble $\boldsymbol{x}$ des variables exogènes pertinentes, c'est-à-dire toutes celles affectant la probabilité de bénéficier de l'intervention. Cette hypothèse est cruciale pour la mise en œuvre d'une technique telle que celle de l'appariement d'échantillons (*matching*). Une fois cette hypothèse admise, si l'échantillon est suffisamment grand et/ou si le nombre de variables exogènes n'est pas trop élevé, il est possible de mettre en œuvre une méthode d'appariement exact : celle-ci repose sur la comparaison du résultat de chaque individu traité avec celui d'un individu non traité dont les caractéristiques observables sont exactement identiques. Lorsque cette méthode ne peut être mise en œuvre, notamment lorsque le nombre de variables exogènes est trop élevé, cet appariement exact est souvent remplacé par un critère de distance permettant d'associer à chaque individu traité son plus « proche voisin » au sens de la distance choisie, ou bien de mettre en œuvre la technique du score de propension, tel que défini plus haut : au résultat de chaque individu traité est comparé celui de l'individu non traité qui a un score de propension dont la valeur est très proche de celle du score de propension de cet individu traité[11]. L'ensemble des variables exogènes pouvant être utilisées pour la construction d'un score de propension valide, assurant notamment l'indépendance conditionnelle des variables exogènes et de l'affectation au groupe de traitement pour une valeur donnée de ce score[12], est potentiellement très vaste. Outre ces variables, il est ainsi possible d'inclure dans cet ensemble certaines de leurs interactions, des indicatrices dichotomiques pour celles qui ont plusieurs modalités (par exemple, les niveaux d'éducation ou les catégories socio-professionnelles), certaines transformations de ces variables telles que leurs puissances ou leur logarithme, etc.

Face à la multiplicité de variables exogènes pouvant être mobilisées, plusieurs travaux récents ont recommandé de mettre en œuvre des méthodes de sélection de modèles et de variables telles que les méthodes d'apprentissage automatique, ou *machine learning* (McCaffrey *et al.*, 2004 ; Wyss *et al.,* 2014 ; Athey & Imbens, 2017a ; Chernozhukov *et al.*, 2018), et les méthodes LASSO[13] (Belloni *et al.*, 2014, 2017 ; Farrell, 2015). Par exemple, McCaffrey *et al.* (2004), comme Wyss *et al.* (2014), combinent la méthode des forêts d'arbres décisionnels[14] (*random forests*) avec la technique LASSO pour estimer le score de propension. Il est à noter que ces méthodes s'appliquent à d'autres procédures d'évaluation que le *matching.* C'est le cas notamment de la méthode proposée par Belloni *et al.* (2017) qui consiste en une double procédure de sélection de variables. La régression LASSO est utilisée dans un premier temps pour sélectionner les variables qui sont corrélées avec la variable de résultat, puis une fois encore pour sélectionner celles qui sont corrélées avec la variable indicatrice de traitement. Après cela, les moindres carrés ordinaires peuvent être appliqués en réunissant ces deux ensembles de variables, ce qui permet d'améliorer les propriétés des estimateurs usuels de l'effet moyen du traitement, notamment par rapport à des techniques plus simples de régression régularisée telles que la régression *ridge*.

L'hétérogénéité des effets d'une intervention

Les travaux récents ont souvent mis l'accent sur l'hétérogénéité des effets d'une intervention entre groupes d'individus éligibles. La figure II illustre cette situation à partir d'un exemple fictif inspiré de Leamer (1983). Pour faciliter la représentation graphique, l'hétérogénéité de l'effet du traitement est supposée être liée à une variable x dont les valeurs différencient les individus les uns des autres. La partie gauche de la figure II décrit l'identification de l'effet causal réalisée à

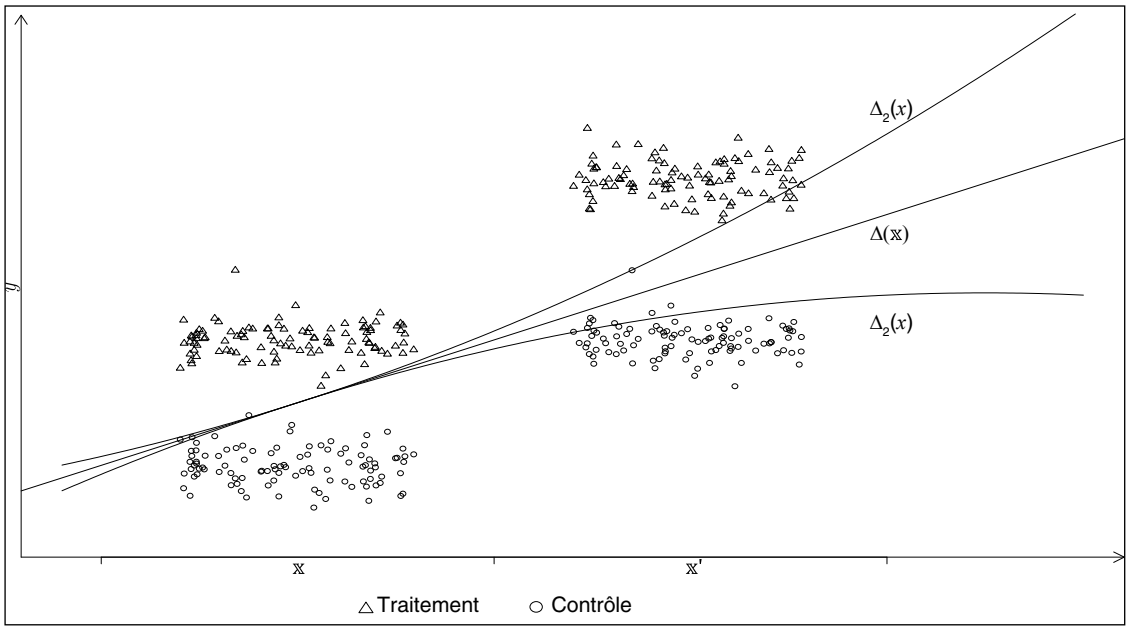
11. Il est parfois préférable de la comparer à une moyenne pondérée des résultats des individus non traités dont les scores de propension ont des valeurs voisines. C'est le principe qui est mis en œuvre dans le cas d'un appariement avec fonction noyau (kernel matching).

12. Cette propriété est dite propriété d'équilibrage (balancing score property).

13. LASSO est l'acronyme de « Least Absolute Shrinkage and Selection Operator *». Cette méthode, introduite par Tibshirani (1996), est une méthode de contraction des coefficients de la régression qui consiste pour l'essentiel à estimer le vecteur de coefficients en minimisant la somme des carrés des résidus sous une contrainte supplémentaire de régularisation.*

14. Pour mettre en œuvre cette technique, le lecteur peut notamment utiliser le package R randomForest *(https://cran.r-project.org/web/packages/randomForest/index.html).*

Figure II
Identification empirique de l'effet d'un traitement à l'aide d'une variable exogène x faiblement dispersée ($x \in \mathbb{x}$) et largement dispersée ($x \in \mathbb{x} \cup \mathbb{x}'$)



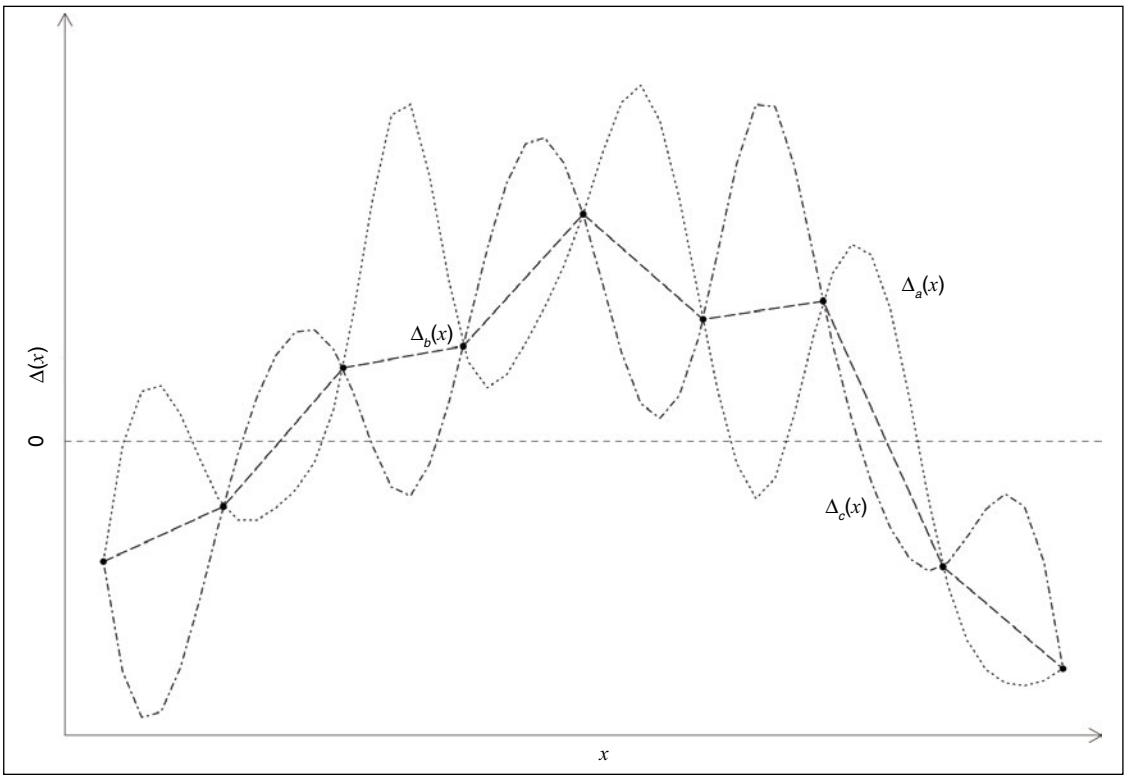
l'aide d'un échantillon d'individus pour lequel les valeurs de la variable exogène, reportées en abscisse, sont faiblement dispersées. La variation de la variable de résultat entre les individus du groupe de contrôle et ceux du groupe de traitement (i.e., l'hétérogénéité de l'effet du traitement) est mesurée par la pente de la droite de régression $\Delta(\mathbb{x})$, mais ne permet pas de trancher entre les multiples projections possibles de cet effet pour des valeurs alternatives de l'hétérogénéité (dont deux exemples sont présentés sur la figure II). Si l'on considère maintenant aussi la partie droite de la figure II, on voit que l'accès à des données supplémentaires, correspondant à une plus grande hétérogénéité des individus ($x \in \mathbb{x} \cup \mathbb{x}'$), permet d'affiner l'analyse et de mesurer la déformation de l'effet du traitement dans la population.

Un éventail plus large de situations observées permet donc d'affiner l'estimation de l'effet causal du traitement, et de caractériser son hétérogénéité en fonction des caractéristiques observables des individus. Quelle que soit la richesse des données disponibles, l'identification de la distribution de l'effet du traitement ne saurait cependant être résolue empiriquement. À titre d'illustration, la figure III présente diverses mesures de l'effet d'un traitement, estimées pour un large éventail de valeurs de la variable exogène x. Il n'en reste pas moins que ces valeurs ponctuelles de l'effet du traitement sont compatibles avec une infinité de distributions sous-jacentes, dont la figure III présente trois exemples : $\Delta_a(x), \Delta_b(x)$, et $\Delta_c(x)$. Aussi fines soient les informations fournies par les données, et aussi hétérogène l'échantillon soit-il, la capacité à décrire la totalité de la distribution de l'effet du traitement nécessite une modélisation préalable destinée à sélectionner la forme de la relation entre la variable de résultat et le traitement.

Dans le cas où l'échantillon est de grande taille et contient des informations sur de nombreuses variables, comme c'est le cas avec des *big data*, il est possible d'estimer des effets hétérogènes de traitement en combinant des méthodes d'inférence causale quasi-expérimentales avec les méthodes LASSO et plus généralement avec les techniques d'apprentissage automatique (voir, par exemple, Wager & Athey, 2018 ; Knaus *et al.*, 2017, 2018). Cette approche statistique peut être généralisée au cas avec plusieurs traitements (Lechner, 2018).

Des travaux empiriques récents ont été consacrés à la mesure de l'hétérogénéité des effets, et ce souvent en conjonction avec la question de la validité externe des estimateurs utilisés. Des exemples particulièrement convaincants de cette approche sont contenus dans les travaux de Dehejia *et al.* (2019) et Bisbee *et al.* (2017) qui examinent, à l'aide d'estimateurs de type LATE et de données provenant de plus d'une centaine de recensements internationaux, le lien de causalité entre fécondité et participation des femmes au marché du travail. Leurs résultats sont relativement convergents. Un autre exemple est fourni par l'étude d'Allcott (2015) qui évalue la

Figure III
De l'estimation à l'identification de la distribution de l'effet du traitement



variation de l'effet d'une politique de réduction de la consommation d'énergie qui a été progressivement mise en œuvre dans 111 sites des États-Unis : il trouve que l'effet de cette politique a été plus fort dans les dix sites dans lesquels le dispositif a été initialement appliqué, ce qui laisse penser que ces premiers sites ont été sélectionnés en raison de leurs caractéristiques particulières.

Précision des effets estimés : la qualité de l'identification au-delà de l'absence de biais

L'attention portée à l'estimation d'effets causaux dans la littérature d'évaluation d'impact a eu tendance à cantonner la réflexion sur l'identification aux propriétés d'absence de biais des effets estimés, la question de la précision étant souvent traitée sur la base de la significativité statistique des effets estimés – une intervention étant considérée comme digne d'intérêt à condition que son effet estimé soit significativement différent de 0.

Une première limite de la significativité statistique, bien connue mais restant encore largement sous-estimée dans la littérature empirique (McCloskey & Ziliak, 1996 ; Ziliak & McCloskey, 2004), est qu'elle ne permet pas de se prononcer sur l'importance quantitative des effets mesurés. Pour chacun de ces effets, la significativité statistique dépend uniquement de la précision de leur estimation. Une estimation ponctuelle de très faible ampleur peut ainsi être statistiquement très significative tandis qu'un effet de très grande ampleur peut être non significatif en raison d'une très faible précision de l'estimation. De fait, les tests d'hypothèse ne sont rien d'autre qu'une formulation alternative des intervalles de confiance (à seuil de confiance et niveau du test équivalents). En ce sens, la significativité statistique ne fait que renseigner sur l'appartenance de la valeur 0 à l'intervalle de confiance déduit du coefficient estimé, c'est-à-dire à l'ensemble des effets sous-jacents compatibles avec l'estimation ponctuelle. S'appuyer uniquement sur la significativité statistique, que ce soit pour écarter une intervention ou pour la considérer comme bénéfique, revient à accorder un poids disproportionné à l'une des nombreuses valeurs appartenant à l'intervalle de confiance, un grand nombre d'entre elles conduisant à prendre une décision contraire à celle qu'indique la significativité statistique *stricto sensu* : en d'autres termes, un intervalle de confiance trop large, i.e. une trop grande imprécision de l'estimation d'un effet dont la valeur

ponctuelle est élevée, peut conduire à écarter l'intervention évaluée si cet intervalle inclut la valeur zéro, ou à la considérer comme bénéfique si cet intervalle, bien que composé de valeurs négligeables, est suffisamment étroit pour exclure le zéro (Amrhein *et al.,* 2019).

L'attention portée à la précision statistique doit être tout aussi rigoureuse que la réflexion menée sur l'identification des effets causaux. L'amélioration de cette précision requiert en particulier de minimiser les sources de variation non contrôlées. Le contrôle de l'environnement – c'est-à-dire la neutralisation des sources de variation autres que celles des variables d'intérêt, tels que le niveau d'un « traitement » ou ses modalités d'application – constitue une démarche expérimentale qui a le mérite de garantir l'identification tout en accroissant la précision des estimations (voir sur ce thème l'article de Deaton & Cartwright, 2018). La randomisation, souvent présentée de manière excessive, voire militante, comme la « règle d'or » de l'évaluation, fait essentiellement reposer l'identification de l'effet causal sur la similarité statistique des unités qui appartiennent aux deux groupes, cible et témoin. Elle ne contrôle pas pour autant l'ensemble des facteurs inobservés qui peuvent « bruiter » l'estimation[15].

L'importance accordée à la significativité des effets estimés peut également conduire à un certain nombre de dérives dans l'interprétation des tests. En particulier, la valeur limite de la statistique de test qui conduit à rejeter l'hypothèse nulle d'absence d'effet ne mesure en aucune manière la probabilité que l'hypothèse alternative, stipulant l'existence d'un effet, soit vraie. Cette probabilité est mesurée par la puissance du test, dont la valeur dépend de la distribution qui produit la statistique de test lorsque l'hypothèse alternative est vraie, et donc de la valeur vraie (inconnue) dont résulte l'estimation. À cela s'ajoute un autre problème : la probabilité critique ne correspond pas à la probabilité que l'hypothèse nulle (i.e., l'absence d'effet) soit vraie. Cette probabilité est en effet conditionnelle à l'hypothèse nulle : la distribution de la statistique de test associée à l'estimation est déduite de la valeur de l'effet sous l'hypothèse nulle. Si l'on note $\hat{s}$ la valeur calculée de la statistique de test, et H_0 l'hypothèse nulle, la probabilité critique mesure donc formellement la quantité $Pr(\hat{s} \mid H_0)$. La probabilité que l'hypothèse nulle soit vraie correspond quant à elle au conditionnement inverse, $Pr(H_0 \mid \hat{s})$. La confusion entre ces deux probabilités peut être illustrée par ce que la littérature en sciences du comportement appelle le sophisme du procureur (*prosecutor fallacy*), concept introduit par Thompson & Schumann (1987) : bien que, par exemple, la probabilité de gagner à la roulette sans tricher soit très faible, il est évidemment erroné d'en déduire qu'un gagnant à la roulette est un tricheur. L'évaluation de la probabilité que l'hypothèse nulle soit vraie requiert une mesure de la probabilité inconditionnelle de cet événement, comme l'illustre la section suivante.

Le risque croissant d'apparition de « faux-positifs », et la nécessité de travaux de réplication

Les tests de significativité sont sujets à deux types de risque d'erreur : les « faux positifs » correspondent aux situations dans lesquelles l'estimation conduit, à tort, à penser à l'existence d'un effet non-nul, et les « faux négatifs » à la situation inverse dans laquelle l'absence de relation estimée n'est qu'apparente. Les probabilités respectives de ces cas correspondent au risque de première espèce (aussi appelé « niveau » du test), souvent noté α et dont la valeur la plus couramment choisie est 5 %, et au risque de deuxième espèce, β, qui correspond à l'inverse de la puissance, $\mathcal{P} = 1 - \beta$. La puissance mesure la probabilité de détecter l'effet de l'intervention et dépend de l'intensité de cet effet : elle ne correspond pas à une probabilité, mais à une fonction qui dépend également de manière cruciale de la taille de l'échantillon[16].

Un effet estimé est « statistiquement significatif au seuil de 5 % » si la probabilité d'observer cette valeur estimée de l'effet alors qu'il est en réalité nul est inférieure à 5 %. Cette propriété implique une probabilité de 5 % de se tromper lorsque l'on conclut à la significativité statistique de l'effet estimé d'une intervention. Cette probabilité est souvent interprétée comme mesurant la proportion de résultats statistiquement significatifs qui sont erronés. Cette conclusion n'est vraie que dans des circonstances très particulières, et les conséquences du risque de première espèce sur la crédibilité des travaux empiriques sont en réalité souvent beaucoup plus sérieuses que ne le laisse apparaître sa valeur.

---

15. Dans un article relativement critique à l'égard des applications mécaniques de la procédure d'expérimentation randomisée, Deaton (2010) passe en revue les problèmes d'identification qui demeurent en dépit de l'affectation aléatoire aux groupes de traitement et de contrôle.

16. Le niveau de puissance de référence dans les travaux appliqués est de 80 %, même si Ioannidis et al. *(2017) montrent que dans plus de la moitié des travaux d'économie appliquée, la puissance médiane est de 18 %, ou même moins.*

Pour illustrer cet argument, Wacholder *et al.* (2004) décrivent les composantes de la probabilité d'occurrence d'un faux-positif (*False-Positive Report Probability*, FPRP ci-après) en fonction des propriétés statistiques des tests de significativité. La FPRP correspond à la probabilité que l'effet d'une intervention soit en réalité nul, alors même que l'estimation produit un effet statistiquement significatif. Le calcul de cette probabilité fait intervenir une quantité inconnue (et dont il est inhabituel de débattre bien qu'elle soit fondamentale) qui correspond à la proportion, notée $\bar{y}$, d'interventions qui ont un effet non nul parmi toutes les interventions qui font l'objet d'une évaluation. Le tableau suivant décrit la probabilité d'occurrence des quatre types de situations possibles : la détection légitime d'une absence (vrai négatif) ou de la présence (vrai positif) d'un effet de l'intervention, ainsi que l'apparition de faux positifs, ou de faux négatifs.

Compte tenu de la combinaison des risques de première et de deuxième espèce, la probabilité d'occurrence d'un faux-positif (la proportion d'effets qui ne sont qu'apparents parmi toutes les interventions dont l'effet est significatif) est mesurée par :

$$FPRP = \frac{\alpha(1-\bar{y})}{\alpha(1-\bar{y})+(1-\beta)\bar{y}}$$

La plupart des tests couramment utilisés sont convergents, c'est-à-dire que leur puissance tend vers la valeur 1 à mesure que la taille de l'échantillon s'accroît. Dans cette situation très favorable (où $\beta = 0$), cette probabilité n'est inférieure au niveau α du test qu'à condition que la moitié au moins de toutes les interventions évaluées aient effectivement un effet non nul. Si cette fréquence est supérieure, la probabilité d'occurrence de faux positifs est inférieure au niveau du test. Elle est supérieure à ce niveau sous l'hypothèse inverse (et certainement plus crédible) selon laquelle, parmi toutes les interventions évaluées, moins d'une sur deux a un effet non nul, situation qui a d'autant plus de chances de se réaliser que les travaux d'évaluation se multiplient. Il est bien évidemment impossible de quantifier $\bar{y}$, et très difficile de recueillir une information objective sur ce point. Mais les conséquences des variations de la proportion $\bar{y}$ sur la crédibilité accordée aux résultats des évaluations ne sont pas anodines : sous l'hypothèse extrême qu'une intervention sur 1 000 a un effet non nul ($\bar{y} = 0{,}001$), la probabilité d'occurrence de faux positifs est supérieure à 98 %.

Cette situation peut être encore aggravée par les conditions dans lesquelles les résultats de l'évaluation sont rendus publics[17]. Ioannidis (2005) met en particulier l'accent sur deux types de biais qui font croître la probabilité d'occurrence de faux positifs : les biais de publication et les biais de communication. Les biais de publication font référence à l'attrait particulier qu'exercent les travaux qui mettent en évidence les effets non nuls d'une intervention, et ce à toutes les étapes du processus – depuis les décisions de financement du projet, jusqu'à la communication des résultats au grand public, en passant par la validation académique que confèrent les publications dans des revues scientifiques prestigieuses. Ces biais de publication conduisent à fausser la proportion de résultats positifs. Ils sont renforcés

17. Nous laissons délibérément de côté la question des pratiques douteuses qui consistent à forcer délibérément la significativité des résultats, par exemple en choisissant à dessein la variable de résultat parmi l'ensemble des variables sur lesquelles l'intervention peut agir, pratique qui fait croître mécaniquement la proportion de faux positifs (voir, par exemple, List et al.*, 2001). Christensen & Miguel (2018) présentent un panorama des pratiques qui conduisent à affaiblir la crédibilité des résultats empiriques en économie, et listent un certain nombre de solutions possibles.*

Tableau
Composantes de la probabilité d'occurrence d'un faux positif

Véracité de l'hypothèse alternative	Test de significativité statistique		Total
	Significatif	Non significatif	
Effet non nul de l'intervention	$(1-\beta)\bar{y}$ [Vrai positif]	$\beta\bar{y}$ [Faux négatif]	$\bar{y}$
Effet nul de l'intervention	$\alpha(1-\bar{y})$ [Faux positif]	$(1-\alpha)(1-\bar{y})$ [Vrai négatif]	$(1-\bar{y})$
Total	$(1-\beta)\bar{y} + \alpha(1-\bar{y})$	$\beta\bar{y} + (1-\alpha)(1-\bar{y})$	1

Note : conditionnellement à l'existence ou à l'absence d'effet de l'intervention, chacune des cellules décrit la probabilité que l'effet estimé soit statistiquement significatif (première colonne) ou statistiquement non-significatif (deuxième colonne) compte tenu du niveau α du test, de sa puissance β, et de la proportion $\bar{y}$ d'interventions qui ont un effet non nul parmi toutes celles qui sont évaluées.
Source : Wacholder *et al.* (2004, p. 440).

par les biais de communication, qui consistent à ne rendre compte d'une évaluation qu'à condition qu'elle conduise à des effets estimés positifs, et simultanément à ne pas rendre publics les résultats d'évaluation concluant à l'absence d'effets pour d'autres interventions. Comme le souligne Roth (1994), ce risque est particulièrement élevé lorsqu'une intervention est élaborée à l'issue d'un processus de tâtonnement, qui conduit à modifier les modalités d'une intervention « pilote » après en avoir constaté l'absence d'effets, et ce jusqu'à l'élaboration d'une proposition finale qui produit les objectifs attendus. Ce processus est légitime parce qu'il permet de construire des politiques publiques efficaces ; il n'affecte pas la probabilité d'apparition de faux positifs si l'ensemble de tous les essais est rendu public en même temps que l'évaluation finale. Dans le cas contraire, ce processus conduit à un biais de communication puisque seules les évaluations de l'intervention qui s'avèrent positives sont rendus publiques, tandis que les tentatives infructueuses qui l'ont précédée sont passées sous silence.

Les biais de publication comme les biais de communication conduisent à une augmentation de la proportion de faux-positifs. Pour illustrer cet argument, notons B la proportion de résultats positifs qui résultent de l'un de ces deux biais. Parmi les $\bar{y}$ interventions qui ont réellement un effet, l'analyse permettra de conclure avec justesse à l'existence d'un effet non nul dans une proportion $(1-\beta)$ de cas, alors qu'un nombre $(B \times \beta)$ apparaitra comme ayant un effet en raison de l'un des biais. De même, une proportion α d'interventions parmi les $(1-\bar{y})$ dont l'effet est réellement nul apparaîtra comme n'ayant pas d'effet, tandis qu'un nombre $B \times (1-\alpha)$ apparaîtra comme ayant un effet non nul en raison des biais. Au total, le FPRP devient :

$$FPRP = \frac{(1-\bar{y})\left[\alpha + B(1-\alpha)\right]}{(1-\bar{y})\left[\alpha + B(1-\alpha)\right] + (1-\beta)\bar{y} + B\beta\bar{y}}$$

* *
*

Pour réussir pleinement la « révolution de crédibilité » annoncée par certains auteurs (Angrist & Pischke, 2010), l'évaluation des politiques publiques ne peut pas s'appuyer seulement sur des stratégies d'identification convaincantes. La réplication des résultats d'évaluation, permettant de distinguer les faux-positifs des effets avérés d'une intervention (Clemens, 2017), reste indispensable, tout comme l'est le souci de la précision des effets estimés. □

BIBLIOGRAPHIE

Abadie, A. & Cattaneo, M. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10, 465–503. https://dx.doi.org/10.1146/annurev-economics-080217-053402

Abadie, A., Diamond, A. & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., Diamond, A. & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2), 495–510. https://doi.org/10.1111/ajps.12116

Abadie, A. & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review,* 93(1), 113–32. https://doi.org/10.1257/000282803321455188

Abadie, A. & L'Hour, J. (2019). A penalized synthetic control estimator for disaggregated data. *Mimeo*.

Abbring, J. H. & Heckman, J. J. **(2007).** Econometric evaluation of social programs, Part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In: Heckman, J. J. & Leamer, E. (Eds.), *Handbook of Econometrics*, vol. 6, part B, chapter 72, pp. 5145–5303. Amsterdam: Elsevier

Amrhein, V., Greenland, S. & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Aeberhardt, R., Fougère, D. & Rathelot, R. (2011). Les méthodes de testing permettent-elles d'identifier et de mesurer l'ampleur des discriminations ? *Économie et statistique*, 447, 97–101. https://www.insee.fr/fr/statistiques/1377350?sommaire=1377352

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *Quarterly Journal of Economics*, 130(3), 1117–1165. https://doi.org/10.1093/qje/qjv015

Amjad, M. J., Shah, D. & Shen, D. (2017). Robust Synthetic Control. *Journal of Machine Learning Research*, 19(22), 1–51. http://www.jmlr.org/papers/volume19/17-777/17-777.pdf

Angrist, J. (2004). Treatment Effect Heterogeneity In Theory And Practice. *Economic Journal*, 114(494), 52–83. https://doi.org/10.1111/j.0013-0133.2003.00195.x

Angrist, J. & Fernandez-Val, I. (2013). ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In: Acemoglu, D., Arellano, M. & Dekel, E. (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress, Volume III: Econometrics*. Cambridge: Cambridge University Press.

Angrist, J., Graddy, K. & Imbens, G. (2000). The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish. *Review of Economic Studies*, 67(3), 499–527. https://doi.org/10.1111/1467-937X.00141

Angrist, J. & Imbens, G. (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association*, 90(140), 431–442. https://scholar.harvard.edu/imbens/publications/two-stage-least-squares-estimation-average-causal-effects-models-variable-treatm

Angrist, J. & Krueger, A. B. (1999). Empirical strategies in labor economics. In: Ashenfelter, O. C. & Card, D. (Eds.), *Handbook of Labor Economics*, vol. 3, part A, chapter 23, pp. 1277–1366. Amsterdam: Elsevier.

Angrist, J. & Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics. *Journal of Economic Perspectives*, 24(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Angrist, J. & Rokkanen, M. (2015). Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff. *Journal of the American Statistical Association*, 110(512), 1331–1344. https://doi.org/10.1080/01621459.2015.1012259

Aronow, P. (2012). A General Method for Detecting Interference in Randomized Experiments. *Sociological Methods and Research*, 41(1), 3-16. https://doi.org/10.1177%2F0049124112437535

Athey, S., Bayatiz, M., Doudchenkox, N., Imbens, G. & Khosravik, K. (2018). Matrix Completion Methods for Causal Panel Data Models. NBER Working Paper No. 25132

Athey, S. & Imbens, G. (2017a). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, 31(2), 3–32. https://doi.org/10.1257/jep.31.2.3

Athey, S. & Imbens, G. (2017b). Econometrics of randomized experiments. In: Banerjee, A. V. & Duflo, E. (Eds.), *Handbook of Economic Field Experiments*, vol. 1, chapter 3, pp. 73–140. Amsterdam : North-Holland.

Avvisati, F., Gurgand, M., Guyon, N. & Maurin, E. (2014). Getting parents involved: A field experiment in deprived schools. *Review of Economic Studies*, 81(1), 57–83, 2014. https://doi.org/10.1093/restud/rdt027

Baird, S., Bohren, J. A., McIntosh, C. & Özler, B. (2018). Optimal Design of Experiments in the Presence of Interference. *The Review of Economics and Statistics,* 100(5), 844–860. https://doi.org/10.1162/rest_a_00716

Baraton, M., Beffy, M. & Fougère, D. (2011). Une évaluation de l'effet de la réforme de 2003 sur les départs en retraite. Le cas des enseignants du second degré public. *Économie et statistique*, 441-442, 55–78. https://www.insee.fr/fr/statistiques/fichier/1377513/ES441D.pdf

Barone, C., Fougère, D. & Pin, C. (2019). Social origins, shared book reading and language skills in early childhood: evidence from an information experiment. *European Sociological Review,* à paraître.

Beffy, M., Fougère, D. & Maurel, A. (2009). L'impact du travail salarié des étudiants sur la réussite et la poursuite des études universitaires. *Economie et statistique*, 422, 31–50. https://www.insee.fr/fr/statistiques/1376784?sommaire=1376788

Behaghel, L., Crépon, B. & Sédillot, B. (2004). Contribution Delalande et transitions sur le marché du travail. *Économie et statistique*, 372, 61–88. https://www.insee.fr/fr/statistiques/1376608?sommaire=1376612

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650. https://doi.org/10.1093/restud/rdt044

Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2017). Program Evaluation and Causal Inference with High-Dimensional Data. *Econometrica*, 85(1), 233–298. https://doi.org/10.3982/ECTA12723

Bénabou, R., Kramarz, F. & Prost, C. (2004). Zones d'éducation prioritaire : quels moyens pour quels résultats ? *Économie et Statistique*, 380, 3–29. https://www.insee.fr/fr/statistiques/1376492?sommaire=1376498

Bérard, G. & Trannoy, A. (2018). The impact of the 2014 increase in the real estate transfer taxes on the French housing market. *Economie et Statistique / Economics and Statistics*, 500-501-502, 179–200. https://www.insee.fr/en/statistiques/3622039?sommaire=3622133

Bertanha, M. & Imbens, G. (2019). External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, à paraître.

Bisbee, J., Dehejia, R., Pop-Eleches, C. & Samii, C. (2017). Local Instruments, Global Extrapolation: External Validity of the Labor Supply-Fertility Local Average Treatment Effect. *Journal of Labor Economics*, 35(S1), S99–S147. https://doi.org/10.1086/691280

Bozio, A. (2011). La réforme des retraites de 1993 : l'impact de l'augmentation de la durée d'assurance. *Économie et Statistique*, 441-442, 39–53. https://www.insee.fr/fr/statistiques/1377511?sommaire=1377529

Brodaty, T., Crépon, B. & Fougère, D. (2007). Les méthodes micro-économétriques d'évaluation et leurs applications aux politiques actives de l'emploi. *Économie & prévision*, 177(1), 93–118. https://www.cairn.info/revue-economie-et-prevision-2007-1-page-93.htm

Bunel, M., Gilles, F. & L'Horty, Y. (2009). Les effets des allégements de cotisations sociales sur l'emploi et les salaires : une évaluation de la réforme de 2003. *Économie et Statistique*, 429-430, 77–105. https://www.insee.fr/fr/statistiques/1377396?sommaire=1377406

Carbonnier, C. (2009). Réduction et crédit d'impôt pour l'emploi d'un salarié à domicile, conséquences incitatives et redistributives. *Économie et Statistique*, 427-428, 67–100. https://www.insee.fr/fr/statistiques/1377124?sommaire=1377130

Chabé-Ferret, S. (2015). Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes. *Journal of Econometrics*, 185(1), 110–123. https://www.sciencedirect.com/science/article/pii/S0304407614002437

Chabé-Ferret, S., Dupont-Courtade, L. & Treich, N. (2017). Évaluation des politiques publiques : expérimentation randomisée et méthodes quasi-expérimentales. *Economie & prévision*, 211-212(2), 1–34. https://www.cairn.info/revue-economie-et-prevision-2017-2-page-1.htm.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N. & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: a meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37(2), 403–429. https://doi.org/10.1002/pam.22051

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Christensen, G. & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3), 920–980. https://doi.org/10.1257/jel.20171350

Clemens, M. A. (2017). The Meaning of Failed Replications: A Review and Proposal. *Journal of Economic Surveys*, 31(1), 326–342. https://doi.org/10.1111/joes.12139

Connolly, P., Keenan, C. & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomized controlled trials in education research 1980–2016. *Educational Research,* 60(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Crépon, B. & Desplatz, R. (2001). Une nouvelle évaluation des effets des allégements de charges sociales sur les bas salaires. *Économie et statistique*, 348, 3–24. https://www.insee.fr/fr/statistiques/1376044?sommaire=1376054

Crépon, B., Devoto, F., Duflo, E. & Parienté, W. (2015). Estimating the impact of microcredit on those who take it up: evidence from a randomized experiment in Morocco. *American Economic Journal: Applied Economics*, 7(1), 123–150. https://doi.org/10.1080/00131881.2018.1493353

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R. & Zamora, P. (2013). Do labor market policies have displacement effects: evidence from a clustered randomized experiment. *The Quarterly Journal of Economics*, 128(2), 531-580.
https://doi.org/10.1093/qje/qjt001

Crépon, B. & Jacquemet, N. (2018). *Econométrie : Méthodes et Applications*, *2ème édition*. Louvain-la-Neuve : De Boeck Universités.

Crépon, B., Leclair, M. & Roux, S. (2004). RTT, productivité et emploi : nouvelles estimations sur données d'entreprises. *Économie et Statistique*, 376-377, 55–89.
https://www.insee.fr/fr/statistiques/1376466?sommaire=1376476

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, 48(2), 424–55.
https://doi.org/10.1257/jel.48.2.424

Deaton, A. & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21.
https://doi.org/10.1016/j.socscimed.2017.12.005

Dehejia, R., Pop-Eleches, C. & Samii, C. (2019). From Local to Global: External Validity in a Fertility Natural Experiment. *Journal of Business & Economic Statistics,* à paraître.

Dong, Y. & Lewbel, A. (2015). Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models. *Review of Economics and Statistics*, 97(5), 1081–1092.
http://dx.doi.org/10.1162/REST_a_00510

Duflo, E. & Banerjee, A. (2017). *Handbook of Field Experiments, Vol. 1 & 2*. Amsterdam: North-Holland.

Edo, A. & Jacquemet, N. (2013). Discrimination à l'embauche selon l'origine et le genre : défiance indifférenciée ou ciblée sur certains groupes ? *Économie et Statistique*, 464-466, 155–172.
https://www.insee.fr/fr/statistiques/1378023?sommaire=1378033

Even, K. & Klein, T. (2007). Les contrats et stages aidés : un profit à moyen terme pour les participants ? Les exemples du CIE, du CES et du Sife. *Économie et Statistique*, 408-409, 3–32.
https://www.insee.fr/fr/statistiques/1377206?sommaire=1377217

Fack, G. & Landais, C. (2009). Les incitations fiscales aux dons sont-elles efficaces ? *Économie et Statistique*, 427-428, 101–121.
https://www.insee.fr/fr/statistiques/1377126?sommaire=1377130

Farrell, M. H. (2015). Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations. *Journal of Econometrics*, 189(1), 1–23.
https://dx.doi.org/10.2139/ssrn.2324292

Fougère, D. & Poulhès, M. (2014). La propriété immobilière : quelle influence sur le portefeuille financier des ménages ? *Économie et Statistique*, 472-473, 213–231.
https://www.insee.fr/fr/statistiques/1377779?sommaire=1377781

Frölich, M. & Sperlich, S. (2019). *Impact Evaluation: Treatment Effects and Causal Analysis*. Cambridge : Cambridge University Press.

Geniaux, G. & Napoléone, C. (2011). Évaluation des effets des zonages environnementaux sur la croissance urbaine et l'activité agricole. *Économie et Statistique*, 444-445, 181–199.
https://www.insee.fr/fr/statistiques/1377857?sommaire=1377863

Givord, P. (2014). Méthodes économétriques pour l'évaluation de politiques publiques. *Économie & prévision,* 204-205(1), 1–28.
https://www.cairn.info/revue-economie-et-prevision-2014-1-page-1.htm

Glazerman, S., Levy, D. M. & Myers, D. (2003). Nonexperimental Versus Experimental Estimates of Earnings Impacts. *The Annals of the American Academy of Political and Social Science*, 589(1), 63–93.
https://doi.org/10.1177%2F0002716203254879

Goux, D., Gurgand, M. & Maurin, E. (2017). Adjusting Your Dreams? Highschool Plans and Dropout Behaviour. *Economic Journal*, 127(602), 1025–1046.
https://dx.doi.org/10.1111/ecoj.12317

Hahn, J. & Shi, R. (2017). Synthetic Control and Inference. *Econometrics*, 54(2), 52.
https://doi.org/10.3390/econometrics5040052

Hill, J. (2008). Comment. *Journal of the American Statistical Association*, 103(484), 1346–1350.
https://doi.org/10.1198/016214508000001002

Heckman, J. J., Lalonde, R. & Smith, J. (1999). The economics and econometrics of active labor market programs. In: Ashenfelter, O. C. & Card, D. (Eds.), *Handbook of Labor Economics*, vol. 3, part A, chapter 3, pp. 865–2097. Amsterdam: Elsevier.

Heckman, J. J. & Vytlacil, E. J. (2007a). Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In: Heckman, J. J. & Leamer, E. (Eds.), *Handbook of Econometrics*, vol. 6, part B, chapter 70, pp. 4779–4874.

Heckman, J. J. & Vytlacil, E. J. (2007b). Econometric evaluation of social programs, Part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In: Heckman, J. J. & Leamer, E. (Eds.), *Handbook of Econometrics*, vol. 6, part B, chapter 71, pp. 4875–5143. Amsterdam: Elsevier.

Hotz, V. J., Imbens, G. & Mortimer, J. H. (2005). Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations. *Journal of Econometrics*, 125(1–2), 241–70. http://dx.doi.org/10.1016/j.jeconom.2004.04.009

Hudgens, M. & Halloran, E. (2008). Towards Causal inference With Interference. *Journal of the American Statistical Association*, 103(482), 832–842. https://doi.org/10.1198/016214508000000292

Imai, K., King, G. & Stuart, E. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, 171(2), 481–502. http://dx.doi.org/10.1111/j.1467-985X.2007.00527.x

Imbens, G. (2010). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2), 399–423. https://doi.org/10.1257/jel.48.2.399

Imbens, G. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A review. *Review of Economics and Statistics*, 86(1), 4–29. http://dx.doi.org/10.1162/003465304323023651

Imbens, G. & Angrist, J. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2), 467–475. https://doi.org/10.2307/2951620

Imbens, G. & Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge : Cambridge University Press.

Imbens, G. & Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5–86. http://dx.doi.org/10.1257/jel.47.1.5

Ioannidis, J. P. A. (2005). Why Most Published Research Findings are False. *PLoS Med*, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A., Stanley, T. D. & Doucouliagos, H. (2017). The Power of Bias in Economics Research. *Economic Journal*, 127(605), F236–F265. https://doi.org/10.1111/ecoj.12461

Jacquemet, N. & L'Haridon, O. (2018). *Experimental Economics: Method and Applications*. Cambridge : Cambridge University Press.

Knaus, M. C., Lechner, M. & Strittmatter, A. (2017). Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach. IZA *Discussion Paper* N° 10961. https://ssrn.com/abstract=3029832

Knaus, M. C., Lechner, M. & Strittmatter, A. (2018). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. IZA *Discussion Paper* N° 12039. https://ssrn.com/abstract=3318814

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4), 604–620. https://www.jstor.org/stable/1806062

Leclair, M. & Roux, S. (2007). Productivité relative et utilisation des emplois de courte durée dans les entreprises. *Économie et Statistique*, 405-406, 47–76. https://www.insee.fr/fr/statistiques/1376937?sommaire=1376947

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73(1), 31–43. https://doi.org/10.1257/jep.24.2.3

Lechner, M. (2018). Modified Causal Forests for Estimating Heterogeneous Causal Effects. IZA *Discussion Paper* N° 12040. https://www.iza.org/publications/dp/12040/modified-causal-forests-for-estimating-heterogeneous-causal-effects

Lee, M. J. (2016). *Matching, Regression Discontinuity, Difference in Differences, and Beyond*. Oxford : Oxford University Press.

List, J. A., Bailey, C., Euzent, P. & Martin, T. (2001). Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior. *Economic Inquiry*, 39(1), 162–170. https://doi.org/10.1111/j.1465-7295.2001.tb00058.x

Liu, L. & Hudgens, M. (2014). Large Sample Randomization Inference of Causal Effects in the Presence of Interference. *Journal of the American Statistical Association*, 109(505), 288–301. http://dx.doi.org/10.1080/01621459.2013.844698

Lorenceau, A. (2009). L'impact d'exonérations fiscales sur la création d'établissements et l'emploi en France rurale : une approche par discontinuité de la régression. *Économie et Statistique,* 427-428, 27–62. https://www.insee.fr/fr/statistiques/1377120?sommaire=1377130

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1), S1–S23.
https://doi.org/10.1111/j.1368-423X.2012.00368.x

McCaffrey, D. F., Ridgeway, G. & Morral, A. R. (2004). Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, 9(4), 403–425.
https://psycnet.apa.org/doi/10.1037/1082-989X.9.4.403

McCloskey, D. N. & Ziliak, S. T. (1996). The Standard Error of Regressions. *Journal of Economic Literature*, 34(1), 97–114.
https://www.jstor.org/stable/2729411

Meager, R. (2019). Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, 11(1), 57–91.
https://doi.org/10.1257/app.20170299

Pearl, J. & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Peters, J., Janzing, D. & Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge : The MIT Press.

Petit, P., Duguet, E., L'Horty, Y., du Parquet, L. & Sari, F. (2013). Discrimination à l'embauche : les effets du genre et de l'origine se cumulent-ils systématiquement ? *Économie et Statistique*, 464-466, 141–153.
https://doi.org/10.3406/estat.2013.10234

Petit, P., Sari, F., L'Horty, Y., Duguet, E. & du Parquet, L. (2011). Les effets du lieu de résidence sur l'accès à l'emploi : un test de discrimination auprès des jeunes qualifiés. *Économie et Statistique*, 447, 71–95.
https://doi.org/10.3406/estat.2011.9711

Rathelot, R. & Sillard, P. (2008). Zones Franches Urbaines : quels effets sur l'emploi salarié et les créations d'établissements ? *Économie et Statistique*, 415-416, 81–96.
https://doi.org/10.3406/estat.2008.7021

Roth, A. E. (1994). Let's Keep the Con Out of Experimental Econ.: A Methodological Note. *Empirical Economics*, 19(2), 279–289.
https://econpapers.repec.org/RePEc:spr:empeco:v:19:y:1994:i:2:p:279-89

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology,* 66(5), 688–701.
https://psycnet.apa.org/doi/10.1037/h0037350

Simonnet, V. & Danzin, E. (2014). L'effet du RSA sur le taux de retour à l'emploi des allocataires. Une analyse en double différence selon le nombre et l'âge des enfants. *Économie et Statistique*, 467-468, 91–116.
https://doi.org/10.3406/estat.2014.10248

Stuart, E. A., Bradshaw, C. P. & Leaf, P. J. (2015). Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science*, 16(3), 475–485.
http://dx.doi.org/10.1007/s11121-014-0513-z

Stuart, E. A., Cole, S. R., Bradshaw, C. P. & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A*, 174(2), 369–386.
https://doi.org/10.1111/j.1467-985X.2010.00673.x

Thompson, W. C. & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials. *Law and Human Behavior*, 11(3), 167–187.
https://psycnet.apa.org/doi/10.1007/BF01044641

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1), 267–288.
https://www.jstor.org/stable/2346178

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. (2004). Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies. *Journal of the National Cancer Institute*, 96(6), 434–442.
https://doi.org/10.1093/jnci/djh075

Wager, S. & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.
https://doi.org/10.1080/01621459.2017.1319839

Wong, V. C., Valentine, J. C. & Miller-Bains, K. (2017). Empirical Performance of Covariates in Education Observational Studies. *Journal of Research on Educational Effectiveness*, 10(1), 207–236.
https://doi.org/10.1080/19345747.2016.1164781

Wyss, R., Ellis, A., Brookhart, A., Girman, C., Jonsson Funk, M., LoCasale, R. & Stürmer, T. (2014). The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score. *American Journal of Epidemiology*, 180(6), 645–655.
https://dx.doi.org/10.1093%2Faje%2Fkwu181

Ziliak, S. T. & McCloskey, D. N. (2004). Size matters: the standard error of regressions in the American Economic Review. *Journal of Socioeconomics*, 33(5), 527–546.
https://doi.org/10.1016/j.socec.2004.09.024