Méthode des doubles différences (difference-in-differences)

Denis Fougère, Nicolas Jacquemet

SciencesPo
LABORATORY FOR INTERDISCIPLINARY EVALUATION OF PUBLIC POLICIES

LIEPP FICHE MÉTHODOLOGIQUE n°9

2023

Cette fiche fait partie d'un ensemble de fiches méthodologiques publiées par le LIEPP. A ce titre, elle bénéficie du soutien apporté par l'ANR et l'État au titre du programme d'Investissements d'avenir dans le cadre de l'IdEx Université Paris Cité (ANR-18-IDEX-0001).

Méthode des doubles différences (difference-in-differences)

Denis FOUGÈRE (Sciences Po, CRIS, CNRS, LIEPP, CEPR et IZA)
denis.fougere@sciencespo.fr

Nicolas JACQUEMET (Université Paris 1, CES, PSE)
nicolas.jacquemet@univ-paris1.fr

www.sciencespo.fr/liepp

Comment citer cette publication :
FOUGERE, Denis, et JACQUEMET Nicolas, **Méthode des doubles différences (difference-in-differences)**, *LIEPP Fiche méthodologique* n°9, 2023-05

This publication is also available in English:
FOUGERE, Denis, and JACQUEMET Nicolas, **Difference-in-differences**, *LIEPP Methods Brief* n°10, 2023-05

EN BREF :

La méthode des doubles différences est une méthode quantitative quasi-expérimentale permettant d'évaluer l'impact d'une intervention grâce à la constitution de groupes de comparaison et à la mesure de l'évolution d'un résultat entre un moment initial pré-intervention et un moment ultérieur où seulement un des deux groupes a reçu l'intervention. Cette méthode est très utile pour l'évaluation *ex post* de l'impact d'une intervention.

Mots-clés : Méthodes quantitatives, méthodes quasi-expérimentales, doubles/triples différences, dimension longitudinale des données, tendances parallèles, équilibrage par entropie, groupe de contrôle synthétique/artificiel

I. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques ?

Bien que l'évaluation des politiques publiques recouvre un ensemble de problématiques et d'outils très large, qui va bien au-delà de la seule quantification de leurs effets, la question de l'efficacité des politiques mises en place dans le passé est évidemment primordiale, car elle constitue un guide utile pour envisager leur pérennisation, leur évolution, leur généralisation, voire leur abandon.

Une telle évaluation nécessite de définir clairement les objectifs poursuivis. S'ils sont mesurables, l'évaluation revient à essayer d'observer quels ont été les effets de l'intervention sur les différents types d'agents (ménages, chômeuses ou chômeurs, entreprises, régions, etc.) qui en ont bénéficié : par exemple, l'effet sur l'intensité des départs en retraite de l'élévation de l'âge minimum de la retraite, l'impact sur les transitions vers l'enseignement supérieur de la mise en place d'un système de bourses universitaires, ou encore les conséquences sur le recours au système de santé de l'introduction d'aides financières pour faciliter l'accès aux soins. Un réflexe naturel, qui apparaît (trop) souvent dans le débat public, consiste à comparer la situation des personnes qui ont bénéficié des interventions mises en place à celle d'autres personnes qui n'en ont pas bénéficié. Pour évaluer l'efficacité d'une réforme de l'assurance chômage offrant une aide personnalisée à la recherche d'emploi, on pourrait ainsi comparer les personnes qui ont bénéficié de cette aide à celles qui n'ont bénéficié d'aucun accompagnement. Comme l'illustre l'étude de Fougère, Kamionka et Prieto (2010, voir en particulier la Figure 3), une telle comparaison montre sans aucune ambiguïté que les programmes d'aide à la recherche d'emploi conduisent à un retour à l'emploi beaucoup plus lent pour les personnes qui en ont bénéficié. Faut-il en conclure que les services proposés nuisent à la capacité des personnes en recherche d'emploi à s'insérer sur le marché du travail ?

Bien évidemment non. Il faut en conclure, selon le vieil adage, que comparaison n'est pas raison, et que les personnes à qui l'on propose un accompagnement sont précisément celles qui éprouvent les difficultés les plus grandes à trouver un emploi. Lorsque l'on compare leur situation à celle des chômeuses et chômeurs qui n'ont pas bénéficié d'une aide, on fait l'hypothèse implicite que le retour à l'emploi observé dans cette catégorie peut servir de point de référence (de « *contrefactuel* ») à la situation qu'auraient connu les bénéficiaires en l'absence d'accompagnement. Or les bénéficiaires le sont précisément parce que leur situation eût été particulièrement difficile en l'absence d'accompagnement. Pour éviter de telles confusions, la méthode des doubles différences consiste à définir les groupes de comparaison de sorte que l'écart observé fournisse une mesure plus convaincante de l'effet de l'intervention.

II. En quoi consiste cette méthode ?

Supposons que l'on observe les variations entre deux dates d'une variable de résultat (également appelée variable-réponse ou variable dépendante) 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 donnée (désignée comme le traitement) ; le second, appelé groupe témoin ou groupe de contrôle, n'en bénéficie pas. La politique est mise en place entre les deux dates considérées. La mesure de l'effet de l'intervention repose exclusivement sur la variation au cours du temps de la variable de résultat entre ces deux dates. Cette variation diffère dans les deux groupes, en général à partir du moment où le traitement entre en vigueur. C'est cette inflexion dans l'écart entre les deux groupes que l'on interprète ici comme l'effet moyen de la politique sur la variable de résultat.

Pourquoi appelle-t-on ce procédé « méthode des doubles différences » ? La première différence correspond à la différence entre la valeur moyenne de la variable de résultat dans le groupe de traitement à la seconde date (après mise en œuvre de la politique à évaluer) et la valeur moyenne de cette même variable dans le même groupe à la date initiale (avant mise en œuvre de la politique à évaluer). À cette première différence, on soustrait ensuite la différence analogue pour le groupe de contrôle. La méthode des doubles différences exploite donc la dimension longitudinale des données (ou pseudo-longitudinale, car les individus qui appartiennent à chacun des groupes peuvent ne pas rester les mêmes au cours du temps) afin de fournir une évaluation *ex post* des politiques publiques mises en œuvre.

La capacité de cette méthode à mesurer l'effet moyen de l'intervention ne repose pas sur l'hypothèse selon laquelle les non-bénéficiaires peuvent servir de groupe de référence aux bénéficiaires en l'absence d'intervention, mais uniquement sur le fait qu'en l'absence d'intervention, l'évolution moyenne de la variable de résultat pour les individus du groupe traité aurait été la même que celle observée au sein du groupe de contrôle (hypothèse de tendances parallèles, « *parallel trends* »). La validité de cette hypothèse, non vérifiable, peut être confortée par le fait qu'avant la mise en place de la politique, la variable de résultat a évolué de la même façon dans les deux groupes (hypothèse de tendances communes, « *common pre-trends* »). A l'inverse de la précédente, cette seconde hypothèse peut être testée grâce aux données observées préalablement à la mise en place de l'intervention, à condition de disposer d'une profondeur d'observation suffisante au cours de cette période – par exemple d'au moins cinq observations dans les deux groupes avant la mise en œuvre de la politique évaluée (ces observations sont appelées « *leads* » dans la littérature académique). L'hypothèse de tendances parallèles équivaut à supposer que l'écart préexistant entre les deux groupes, qui provient des divers facteurs conduisant à des niveaux différents de la variable de résultat au sein de ces groupes, serait resté le même en l'absence d'intervention, de sorte que l'évolution constatée de cet écart peut être interprétée comme l'effet moyen de l'intervention.

Cette approche n'est donc valide qu'à condition que l'intervention laisse inchangée la variable de résultat dans le groupe de contrôle (hypothèse appelée SUTVA, « *Stable Unit Treatment Value Assumption* »). En effet, tout effet indirect de l'intervention sur ce groupe (si, par exemple, la difficulté à trouver un emploi augmente parce que l'accélération du retour à l'emploi dans le groupe de traitement accroît la tension sur le marché du travail) remet en cause l'hypothèse de tendances parallèles. De la même façon, cette hypothèse pourrait être remise en cause si le groupe de traitement anticipait la mise en place de l'intervention (ce qui aurait pour conséquence, par exemple, de ralentir sa recherche d'emploi en raison de la perspective des mesures d'accompagnement à venir). Cette violation de l'hypothèse de tendances parallèles est connue sous le terme d'« *Ashenfelter gap* ».

Compte tenu des nombreux facteurs susceptibles d'affecter la validité de l'approche, les développements récents de la méthode des doubles différences visent notamment à affiner la constitution des groupes de

manière à augmenter leur comparabilité (voir Roth et alii, 2022, pour une description détaillée). Il est ainsi possible de recourir aux méthodes de « *matching* » (voir fiche séparée) qui associent, à l'aide d'un critère statistique, à chaque personne bénéficiaire de l'intervention la ou les personnes du groupe de contrôle dont les caractéristiques observables sont proches – de sorte que la comparaison est réalisée entre « *plus proche voisins* » statistiques – ou bien encore à la méthode de l'équilibrage par entropie (*entropy balancing*) qui permet d'égaliser les premiers moments (moyenne, variance, coefficient d'asymétrie, etc.) des distributions des variables explicatives observables au sein des deux groupes. Une démarche similaire peut-être appliquée à la variable de résultat elle-même plutôt qu'à la distribution des caractéristiques observables. C'est l'objectif du *groupe de contrôle synthétique*, qui consiste à créer par un système adéquat de pondérations un *groupe de contrôle artificiel* à partir des observations du groupe de contrôle. Ce groupe de contrôle synthétique est construit de telle sorte que l'évolution passée de la variable de résultat en son sein soit identique à celle de la même variable dans le groupe de traitement en minimisant, à l'aide d'un système de pondération des observations du groupe de contrôle, la distance relative à la variable de résultat entre le groupe traitement avant intervention et ce groupe de contrôle synthétique. 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. Plusieurs contributions récentes ont proposé des solutions permettant de résoudre cette difficulté. Parmi celles-ci, certaines suggèrent d'utiliser des techniques de complétion de matrices, d'autres proposent des méthodes d'inférence fondées sur les échantillons (*sampling-based inferential methods*).

L'une des extensions les plus populaires destinées à prendre en compte l'existence d'interactions non observables entre les caractéristiques de groupe et de temps que la méthode des doubles différences pourrait omettre est la méthode des « triples différences » (*difference in difference in differences*). Cette méthode repose sur l'observation de deux groupes supplémentaires, un « faux » groupe de traitement et un « faux » groupe de contrôle. Imaginons par exemple une politique de santé qui s'applique dans une région *A* aux personnes de plus de 65 ans. Pour évaluer les effets de cette politique sur le recours aux soins et sur l'état de santé des personnes concernées, il est possible de considérer comme groupe de traitement les personnes âgées de 65 à 69 ans de la région *A*, et utiliser la situation de celles qui sont âgées de 60 à 64 ans dans cette même région comme groupe de contrôle. Une première double différence appliquée à ces deux groupes doit en principe produire une estimation de l'effet moyen de l'intervention sur le recours aux soins et l'état de santé des personnes de plus de 65 ans dans la région *A*. Mais on peut reprocher à cette approche qu'elle compare des populations qui ne sont pas tout à fait les mêmes du point de vue de leur état de santé : les personnes de 68 ou 69 ans sont probablement en moins bonne santé que celles âgées de 60 ou 61 ans, et donc exposées à des risques de dégradation de leur santé au cours du temps qui sont plus élevés. Pour répondre à cette critique, il est possible de considérer les mêmes groupes d'âge dans une seconde région, la région *B*, dans laquelle la même politique n'est pas mise en œuvre, puis de calculer un second estimateur des doubles différences dans cette région *B*. On peut ensuite soustraire cette seconde double différence dans la région *B* à celle calculée dans la région *A*. La seconde double différence appliquée aux deux groupes de personnes de la région *B* élimine les écarts de santé entre groupes d'âge qui prévalent naturellement dans l'ensemble de la population (l'hypothèse de tendances parallèles est donc affaiblie, et porte ici sur la différence relative entre les deux catégories de population dans chacune des deux régions).

Outre la qualité de la comparaison entre les groupes, une seconde limite de la méthode des doubles différences est que l'effet de l'intervention n'est pas toujours identique au sein de différents sous-groupes de bénéficiaires, ou au cours du temps (l'effet de l'intervention est dit « hétérogène »). En s'appuyant sur l'évolution de l'écart entre deux groupes seulement, cette méthode mesure uniquement un effet moyen, qui n'est compatible qu'avec de très fortes variations de l'effet de l'intervention entre différents sous-groupes. Afin d'étudier les variations de l'effet au cours du temps, il est utile de disposer d'observations de la variable de résultat dans les deux groupes bien au-delà de la date qui suit la mise en œuvre de

cette intervention (observations qui sont parfois appelées « *lags* »). Il est ainsi possible de s'assurer que la politique évaluée a bien des effets significatifs à moyen terme, voire à long terme si le suivi statistique est suffisamment long.

Une telle hétérogénéité de l'effet de l'intervention soulève également des difficultés importantes lorsque sa diffusion dans le groupe de traitement est graduelle. La méthode habituelle, qui consiste à intégrer les observations au groupe des bénéficiaires au fur et à mesure de leur éligibilité à l'intervention conduit en effet à des conclusions infondées (qui peuvent aller jusqu'à conclure à l'inefficacité d'une intervention dont les effets sont positifs pour l'ensemble des bénéficiaires). Les études récentes recommandent de se concentrer uniquement sur les observations qui correspondent à un changement de situation, ce qui revient à combiner de multiples estimateurs des doubles différences calculés à toutes les dates auxquelles le périmètre du groupe de bénéficiaires évolue (voir de Chaisemartin et d'Haultfoeuille, 2022, pour une présentation complète).

III. Un exemple d'utilisation de cette méthode dans le domaine de l'emploi

Comme la plupart des instruments d'intervention sur le marché du travail, l'introduction d'un salaire minimum et la fixation de son niveau résultent d'un arbitrage délicat : sur un marché du travail où les employeurs ont un pouvoir de négociation élevé qui leur permet de comprimer les salaires, le salaire minimum constitue une protection pour les salarié-e-s et permet de répartir les bénéfices de la production de manière plus équitable. Mais l'existence d'un salaire minimum implique également que toutes les activités dont la rentabilité est inférieure au salaire minimum ne pourront pas avoir lieu car elles ne permettent pas de créer une valeur suffisante pour couvrir le coût des salaires. Toute la difficulté est donc de fixer un salaire minimum qui rééquilibre la négociation salariale sans nuire de manière excessive à l'efficacité économique.

L'article de Card et Krueger (1994) sur l'augmentation du salaire minimum décrétée dans le New Jersey en avril 1992, est l'une des études les plus célèbres de la mise en œuvre de la méthode des doubles différences. Dans cette étude, Card et Krueger comparent le niveau d'emploi dans le secteur de la restauration rapide (très intensive en emploi peu qualifié qui est en général rémunéré au salaire minimum) dans le New Jersey et en Pennsylvanie, en février 1992 et en novembre 1992. Ces dates encadrent une augmentation du salaire minimum horaire de 4,25 US dollars à 5,05 US dollars intervenue en avril 1992 dans le New Jersey, alors qu'à la même date, ce salaire restait constant et égal à 4,25 US dollars en Pennsylvanie. Observer une évolution de l'emploi dans le New Jersey entre février et novembre 1992 au moyen d'une première différence ne permet pas d'attribuer cette évolution à la seule hausse du salaire minimum dans cet état, notamment parce que d'autres facteurs concomitants, tels que les conditions météorologiques ou macroéconomiques, pourraient également y contribuer. Par ailleurs, l'écart dans les niveaux d'emploi entre les deux états postérieurement à l'élévation du salaire minimum reflète non seulement l'effet de cette politique mais aussi l'ensemble des différences de fonctionnement de ce secteur d'activité entre le New Jersey et la Pennsylvanie.

En incluant à la fois les fast-foods du New Jersey (le groupe de traitement) et de Pennsylvanie (le groupe de contrôle), situés des deux côtés de la frontière de ces états, Card et Krueger peuvent limiter au moyen d'une seconde différence les effets de ces deux types de facteurs. Sous l'hypothèse de tendances parallèles, l'évolution de l'emploi dans le secteur de la restauration rapide en Pennsylvanie peut être interprétée comme l'évolution de l'emploi qu'aurait connu le secteur de la restauration rapide dans le New Jersey si le salaire minimum horaire n'avait pas augmenté dans cet état. Les estimations réalisées par Card et Krueger suggèrent que l'augmentation du salaire minimum ne s'est pas accompagnée d'une diminution de l'emploi dans le New Jersey. Plus précisément, Card et Krueger estiment que

l'augmentation de 0,80 US dollars du salaire minimum horaire dans le New Jersey a entraîné (a “causé”) une augmentation de 2,75 emplois à temps plein en moyenne dans chaque fast-food de cet état.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode ?

L'estimateur obtenu sera d'autant plus informatif (et l'hypothèse de tendances parallèles sera d'autant plus crédible) que le groupe de contrôle est semblable au groupe de traitement du point de vue des caractéristiques explicatives observables (en évitant de surinterpréter de telles comparaisons, puisque l'hétérogénéité inobservable peut varier considérablement entre les groupes sans que cela soit détectable). A moins que la constitution des groupes ne suive une procédure qui impose une telle condition, il convient pour s'en assurer de mettre en regard la distribution des caractéristiques observables entre les groupes (par exemple, dans un échantillon de salarié-e-s, la proportion de femmes, les différents groupes d'âge et de niveaux d'éducation, etc.) puis de réaliser un ensemble de tests statistiques permettant de vérifier l'absence de différences significatives entre les groupes (la procédure est connue sous le nom de « *balancing test* »). Une bonne pratique consiste à conditionner l'analyse statistique à toute caractéristique observable dont la distribution varie entre les groupes afin de prendre en compte de possibles interactions entre cette caractéristique et les variations au cours du temps.

Afin de vérifier la robustesse des résultats, il est possible de recourir à des groupes dits « placebos » de manière à répliquer l'analyse sur un groupe d'observations n'ayant pas été exposé à l'intervention évaluée. Une première façon de procéder est d'utiliser un « faux » groupe de traitement, qui peut être le même groupe de traitement mais observé à au moins deux dates antérieures à la mise en place de la politique publique évaluée, ou bien encore un troisième groupe qui est supposé ne pas être concerné par la politique mise en œuvre. La robustesse de l'analyse est confortée si cette procédure permet de conclure à l'absence d'effet. Une seconde pratique consiste à utiliser un autre groupe de contrôle, semblable au premier groupe de contrôle utilisé. En ce cas, l'estimation de l'effet moyen du traitement doit être approximativement égale à celle obtenue avec le groupe de contrôle initial.

Si l'utilisation de données répétées dans le temps permet d'améliorer la qualité des comparaisons qui sont faites, elles conduisent à travailler avec des observations qui sont liées entre elles au cours du temps. Cette propriété des données a longtemps été négligée dans l'application de la méthode des doubles différences, ceci conduisant à des mesures de la significativité statistique des effets observés qui sont erronées. Il est donc important de prendre en compte la structure de corrélation des données dans l'analyse statistique (Bertrand et alii, 2004).

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres ?

La méthode des doubles différences est une méthode quasi-expérimentale, dans le sens où elle est principalement utilisée pour étudier des changements qui surviennent spontanément et selon des modalités qui ne sont pas directement liées à l'objectif d'évaluation, mais qui produisent des observations qui permettent de se rapprocher d'une situation expérimentale. Comme toutes les méthodes quasi-expérimentales, les effets qu'elle mesure correspondent aux effets de la politique sur la sous-population qui a effectivement été ciblée et a de fait bénéficié de la politique (dans les termes du modèle d'évaluation causale des politiques publiques, elle mesure un effet de traitement sur les traités, ATT « *Average Treatment on the Treated* »). Dès lors que l'intervention a été ciblée à dessein sur des catégories de population particulières (qui sont particulièrement sensibles à l'intervention mise en œuvre, ou qui en ont particulièrement besoin) cette approche ne permet pas de mesurer l'effet moyen de la politique (ATE, « *Average Treatment Effect* »), c'est-à-dire l'effet qu'elle produirait si elle était généralisée à l'ensemble de la population, ni même les variations de l'effet entre différents individus traités. Athey et Imbens (2006)

proposent une approche alternative à la méthode des doubles différences qui fournit une estimation de la totalité de la distribution contrefactuelle de la variable de résultat et permet de mesurer plus finement les variations de l'effet de l'intervention entre différents types de bénéficiaires.

Il n'en reste pas moins que cette méthode permet de mesurer un effet moyen dans une sous-population plus large que la plupart des méthodes quasi-expérimentales existantes. À ce titre, elle se distingue en particulier de la régression sur discontinuité (voir fiche séparée) et de l'estimation de l'effet local moyen du traitement (*local average treatment effect*, ou LATE) qui permettent seulement d'estimer les effets moyens du traitement pour une sous-population particulière, à savoir pour le sous-groupe de personnes (appelées « compliers » en anglais) dont l'accès au traitement est uniquement dû à leur proximité d'un seuil fixé de manière exogène (par exemple, un seuil d'âge ou de revenu) dans le premier cas, et celles et ceux qui en bénéficient en raison de la variable d'instrumentation dans le second.

Quelques références bibliographiques pour aller plus loin

Athey, Susan, et Imbens, Guido. W.. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models", *Econometrica*, 74(2), 431–97. https://doi.org/10.1111/j.1468-0262.2006.00668.x

Bertrand, Marianne, et Duflo, Esther, et Mullainathan Sendhil. 2004. "How Much Should We Trust Differences-In-Differences Estimates?", *Quarterly Journal of Economics*, 119(1), 249–275. https://doi.org/10.1162/003355304772839588

Card, David, et Krueger, Alan B.. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania", *American Economic Review*, 84(4), 772-793. https://www.jstor.org/stable/2118030

De Chaisemartin, Clément, et D'Haultfoeuille, Xavier. 2022. "Difference-in-Differences Estimators of Intertemporal Treatment Effects", *NBER Working Paper* No. 29873. DOI 10.3386/w29873

Fougère, Denis, et Kamionka, Thierry, et Prieto, Ana. 2010. « L'efficacité des mesures d'accompagnement sur le retour à l'emploi », *Revue Economique*, 61(3), 599–612. http://dx.doi.org/10.3917/reco.613.0599

Roth, Jonathan, et Sant'Anna, Pedo H. C., et Bilinski, Alyssa, et Poe John. 2022. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature", arXiv:2201.01194, https://doi.org/10.48550/arXiv.2201.01194

Des ressources pour mettre en œuvre cette méthode avec les logiciels Stata et R

Cunningham, Scott. 2021. Causal Inference: The Mixtape. *Yale University Press*: New Haven and London. Disponible en libre accès sur le site https://mixtape.scunning.com/index.html

Huntington-Klein, Nick. 2022. The Effect: An Introduction to Research Design and Causality, Chapitre 18. Chapman and Hall/*CRC Press*: Boca Raton, Florida. Disponible en libre accès sur le site https://theeffectbook.net/ch-DifferenceinDifference.html