Guide pratique pour la mise en œuvre des forêts causales modifiées (FCM) dans l'estimation des impacts causaux des programmes du marché du travail

Atelier d'évaluation quantitative des impacts (EQI)

16 février 2026

Direction de l'évaluation – Direction générale des politiques stratégiques et des services

Aperçu de la présentation

Objectif: Présenter la procédure étape par étape pour mettre en œuvre l'approche des forêts causales modifiées (FCM) dans l'analyse d'impact des programmes du marché du travail

Aperçu de l'analyse d'impact

- Approche traditionnelle
- Apprentissage automatique causal (forêts causales modifiées)

Cadre méthodologique

- 10 étapes pour mettre en œuvre FCM à l'aide de données synthétiques

Contexte

Problème fondamental de l'inférence causale

- On ne peut jamais observer simultanément les deux résultats potentiels pour un même individu. Autrement dit, une fois qu'un individu reçoit un traitement, on ne peut pas observer ce qui se serait produit en l'absence de traitement.

Cadre des résultats potentiels

- Elle définit et estime les effets causaux en abordant le problème fondamental de l'inférence causale.
- Dans ce cadre, les effets causaux sont estimés comme la différence entre les résultats moyens* observés avec et sans traitement.

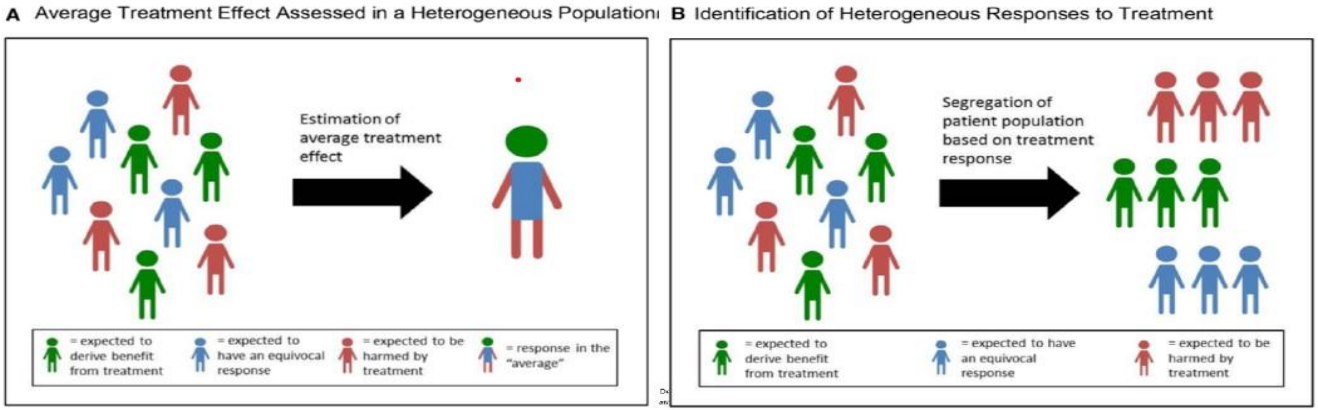
Dans les études observationnelles, on utilise généralement un cadre non- expérimental pour construire un contrefactuel à partir d'un groupe de comparaison similaire.

- Estimer les effets causaux moyens en construisant un scénario contrefactuel à l'aide d'un groupe de comparaison dont les caractéristiques observables sont, en moyenne, équivalentes à celles des participants.
- La méthode bien établie pour estimer les effets causaux repose sur l'appariement par score de propension et la méthode des doubles différences (connue comme « méthodes traditionnelles »).

* Note : nous ne pouvons pas calculer directement les effets causaux individuels

Contexte (suite)

- L’hétérogénéité des effets du traitement renvoie à la variation de l’impact des programmes du marché du travail selon les différents groupes sociodémographiques.
- Les méthodes traditionnelles ne sont pas optimales pour identifier cette hétérogénéité des effets.
- L’apprentissage automatique causal adapte les méthodes d' apprentissage automatique afin de répondre à des questions causales bien définies. La forêt causale (FC) est une approche causale fondée sur des arbres décisionnels.
- La forêt causale modifiée (FCM) introduit par Lechner (2019), permet d’estimer efficacement les effets du traitement au niveau le plus granulaire, offrant ainsi une analyse plus robuste et fondée sur les données de l’hétérogénéité des effets, et permettant ainsi de mieux comprendre ce qui fonctionne pour qui.



Forêt causale (Notions fondamentales)

Les forêts aléatoires peuvent être transformées en Forêts causales, à condition que les hypothèses classiques d'identification des effets causals soient respectées et que certaines modifications de l'algorithme* soient apportées.

Hypothèses d'identification:

- ❑ Hypothèse d'indépendance conditionnelle (HIC)
- ❑ Support Commun (SC)
- ❑ Hypothèse de valeur de traitement unitaire stable (AVTS)
- ❑ Exogénéité des facteurs de confusion

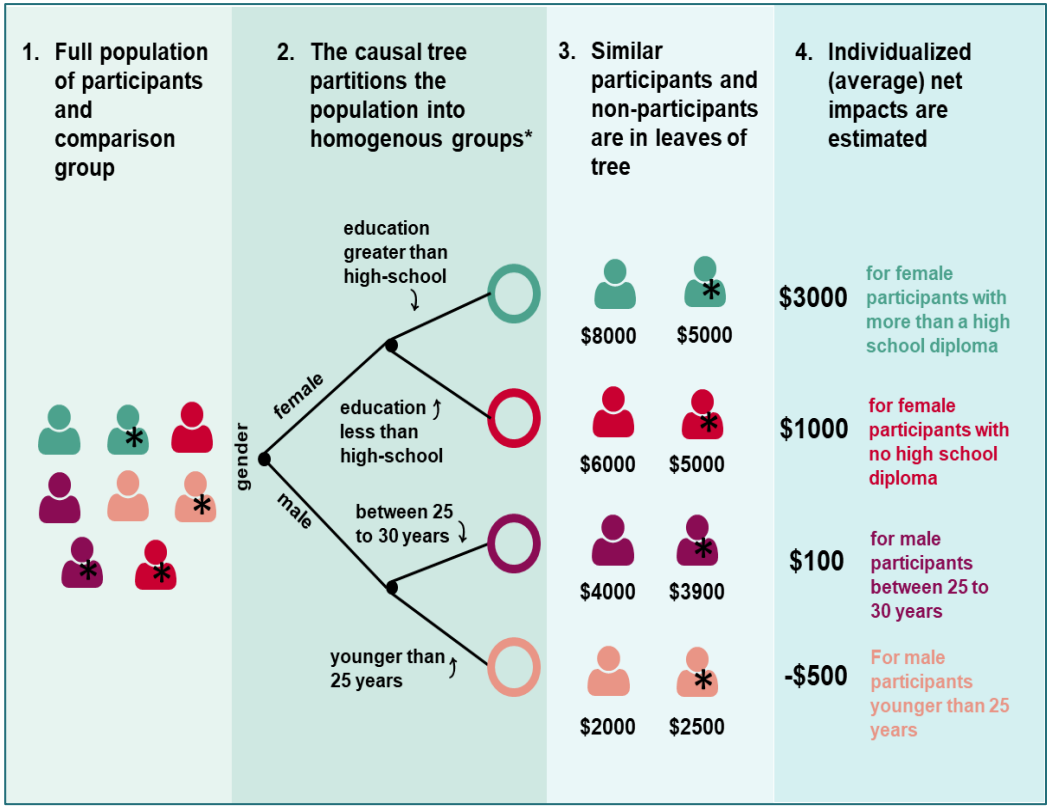
Partage de l'échantillon dans la Forêt causale :

- ❑ L'échantillon initial est divisé en deux ensembles disjoints égaux : l'ensemble entraînement (50 %) et l'ensemble honnête (équitable) (50 %).
- ❑ L'ensemble de partition est utilisé pour déterminer comment fractionner l'espace des variables (c'est-à-dire comment faire croître l'arbre), tandis que l'ensemble honnête sert à estimer les effets du traitement dans chaque feuille. Cette approche empêche d'utiliser les mêmes données à la fois pour la sélection du modèle et pour l'estimation des effets.

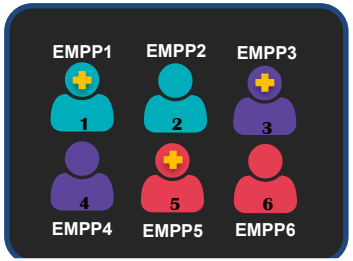
* La fonction objective de la forêt causale vise à maximiser l'hétérogénéité des effets du traitement entre les feuilles.

Visualisation d'un arbre causal

- Les arbres divisent les données de manière récursive en quartiers granulaires et ne se chevauchant pas
- À chaque nœud, la division est choisie de manière à satisfaire une fonction objective et à préserver un nombre minimal de participants et de groupe de comparaison.
- L'impact différentiel individualisé est estimé comme la différence entre les résultats moyens des participants et des non-participants dans les feuilles.

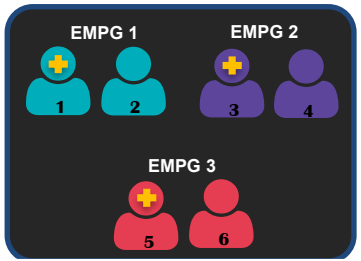


Principaux paramètres du modèle de forêt causale



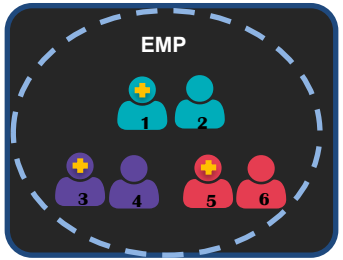
Effet moyen du programme sur la personne (EMPP)

Mesure l'effet moyen d'un programme sur des personnes ayant un ensemble de caractéristiques ou un profil donné. Représente l'impact causal du programme au niveau de granularité le plus fin.



Effet moyen du programme sur le groupe (EMPG)

L'EMPG peut être estimé en agrégeant et en pondérant les EMPP sur des sous-groupes précis. Contrairement aux analyses traditionnelles par sous-groupes, les EMPG peuvent être comparés entre les groupes.



Effet moyen du programme (EMP)

Représente l'effet moyen du programme sur la population.

Remarque : ✚ indique que la personne est un participant.

De la forêt causale à la forêt causale modifiée

La forêt causale modifiée est une extension du cadre original de la forêt causale, intégrant des modifications à la méthodologie standard. Ces modifications visent à améliorer la précision de l'estimation de l'effet du traitement.

- La première modification concerne l'amélioration du découpage de l'échantillon grâce à l'introduction d'une nouvelle règle de découpage des arbres, permettant de réduire le biais de sélection dans les études observationnelles.

- La seconde modification exploite une procédure d'inférence pondérée, autorisant une agrégation flexible des effets du traitement à différents niveaux (**Effet moyen du programme sur la personne** [**EMPP**], **Effet moyen du programme sur le groupe** [**EMPG**], **Effet moyen du programme** [**EMP**] **dans la population)** facilitant ainsi l'interprétation et l'application pratique des résultats.

Avantages de FCM comparés à l'approche traditionnelle

- L'approche traditionnelle repose généralement sur des méthodes semi-paramétriques, tandis que les approches d'apprentissage automatique telles que FCM sont non paramétriques. Elles sont donc plus flexibles.

- Comme les méthodes traditionnelles d'appariement estiment les effets de sous-groupes (EMPG) indépendamment, il est impossible d'évaluer et de comparer conjointement les EMPG et l'effet moyen du programme (EMP) sur la population. Par conséquent, il est impossible de détecter l'hétérogénéité de l'effet du traitement à l'aide de méthodes d'appariement.

  - À l'inverse, FCM surmonte cette difficulté en permettant l'estimation conjointe des EMPG. FCM peut modéliser la variation de l'effet du traitement entre les sous-groupes, permettant ainsi une analyse plus robuste axée sur les données pour examiner l'hétérogénéité de l'effet.

Mise en œuvre du cadre FCM

Étude de cas: Un pays met en œuvre un programme de formation coûteux afin de favoriser l'insertion professionnelle des chômeurs. Ce programme vise à accroître les revenus et l'emploi en rehaussant le niveau de compétences, notamment celui des personnes peu qualifiées. Le gouvernement souhaite évaluer l'impact de la participation à ce programme sur les revenus individuels des chômeurs participants. Voici quelques précisions sur le programme :

- Durée: 3-6 mois.
- Date du début de la participation: 1er trimestre 1993
- Âge : 30-50 ans
- Informations détaillées sur les variables avant le programme et les résultats.

❑ Le groupe de comparaison est composé d'individus qui étaient admissibles à participer (pendant la période considérée) mais qui n'ont pas participé et qui présentent des caractéristiques sociodémographiques similaires.

Étapes de mise en œuvre

Étape 1: Contexte institutionnel et de l'évaluation

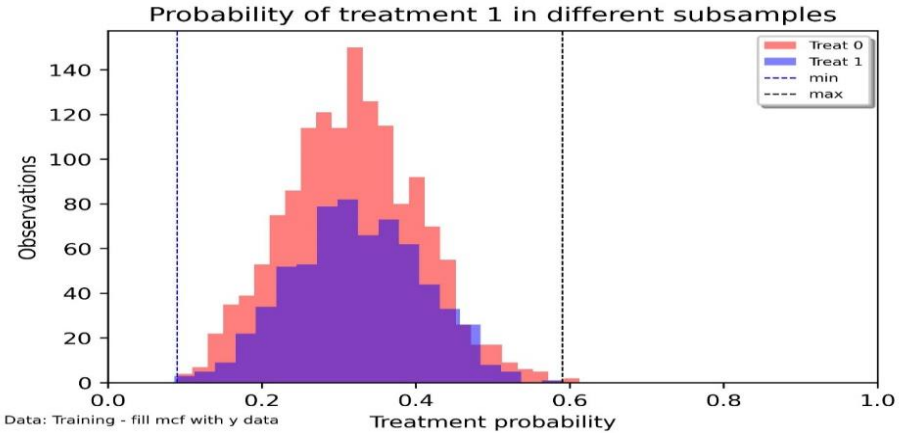
- Il s'agit d'une étude observationnelle portant sur un programme du marché du travail. Des données administratives riches sont disponibles concernant un programme de formation de courte durée offert dans le cadre du Développement des compétences.
- L'objectif principal de cette analyse d'impact est d'estimer l'effet moyen du traitement chez les participants, ainsi que l'hétérogénéité des effets du programme de formation sur un indicateur de résultats à moyen terme, en utilisant l'approche d'apprentissage automatique causal, la forêt causale modifiée.

Étape 2: Modélisation causale

- Traitement: Programme de formation sur le marché du travail.
- Résultat: Revenu gagné au cours du 4^{e} trimestre de la 6^{e} année après la participation.
- Grand volume de données sociodémographiques et d'information sur le marché du travail, permettant une modélisation détaillée. Le modèle repose sur l'hypothèse de sélection fondée sur les variables observables.

Étape 3 : Hypothèses d’identification

- **HIC** : Cette hypothèse est valable car nous disposons d'un nombre suffisant de caractéristiques sociodémographiques et d'informations sur l'historique du marché du travail.
- **SC** : On observe un très bon chevauchement entre le groupe de traitement et le groupe de comparaison.
- **AVTS** : Il est raisonnable de supposer qu'il n'y a pas d'effet de contagion, car le programme est de petite envergure par rapport au marché du travail régional concerné.
- **Exogénéité des facteurs de confusion** : Le traitement est peu susceptible d’exercer une influence sur les variables confondantes.

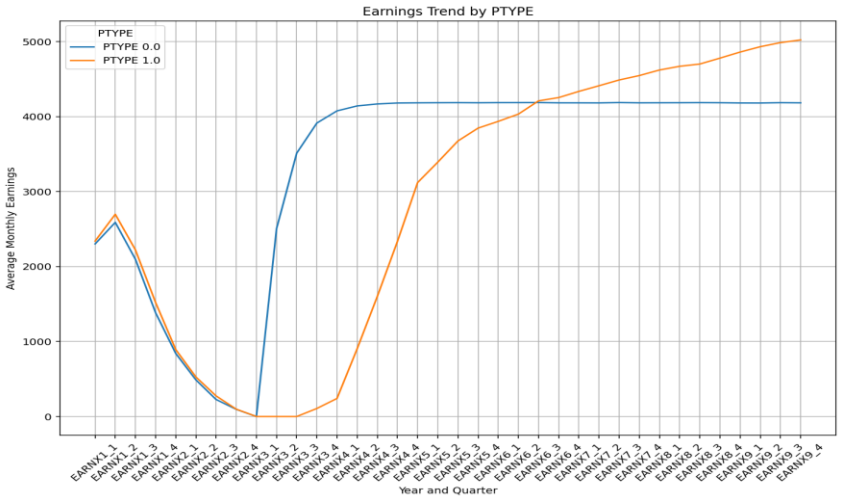


Étape 4 : Préparation des données et construction du groupe de comparaison

- **Traitement de données*:** suppression des valeurs manquantes, vérification des doublons et transformation des variables, afin de préparer le fichier analytique final.
- **Sélection du groupe de comparaison:** Le groupe de comparaison est composé de non-participants admissibles. Il faut s'assurer qu'ils disposent de la même information que les participants et que le groupe de comparaison est suffisamment large pour permettre une estimation fiable.

Étape 5 : Diagnostique avant estimation

Analyse des tendances des résultats: Nous comparons les revenus des participants et du groupe de comparaison avant et après leur participation au programme. Pour ce programme de formation de courte durée, nous observons un effet de stabilisation des revenus durant les deux premiers trimestres



*L'ensemble de données comprend 13 628 sujets, répartis entre 4 251 participants et 9 377 individus du groupe comparatif, et comporte 82 variables observées.

Étape 6 : Estimation de l'EMPP et de l'effet moyen du traitement

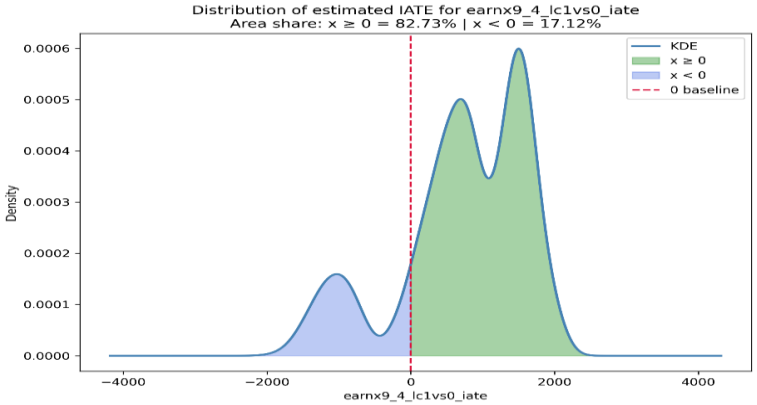
- FCM fournit l'estimation la plus fine de l'effet du traitement, appelée EMPP
- L'impact moyen du programme sur la population (EMP) et sur le participant (EMPPP*) peut être obtenu par approximation des EMPP. Les valeurs estimées de l'EMPP et de l'EMPPP sont hautement statistiquement significatives.

Paramètre	Estimation	Valeur P
EMP	723	0
EMPPP	662	0

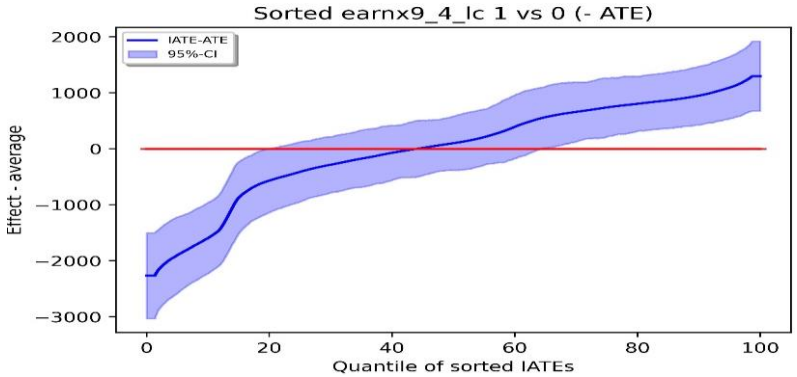
*Dans cette étude, nous nous concentrons principalement sur les estimations relatives à la population traitée (telles que l' EMPPP, le EMPGP et le EMPGPE), puisque notre principal objectif est de mesurer l'impact d'un traitement ou d'une intervention sur les personnes qui l'ont effectivement reçu.

Étape 7 : Identification de l'hétérogénéité au niveau individuel

- Distribution des EMPP estimés : L'analyse de la distribution des EMPP estimés révèle que la plupart des participants (83 %) sont avantagés par leur participation au programme. Cependant 17 % sont désavantagés par leur participation au programme, leur EMPP étant négatif.



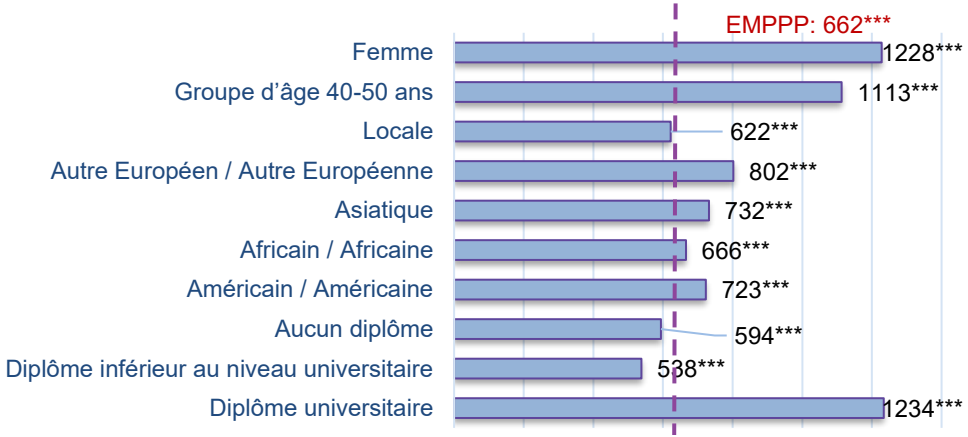
- Pour détecter l'hétérogénéité de l'effet, nous avons représenté graphiquement la différence entre les EMPP et les EMP. On observe un écart entre la droite EMPP-EMP et la ligne de référence zéro (ainsi qu'une absence de chevauchement entre les intervalles de confiance). Ceci indique une hétérogénéité de l'effet du traitement au niveau individuel.



Étape 8 : Identification de l'hétérogénéité au niveau du groupe

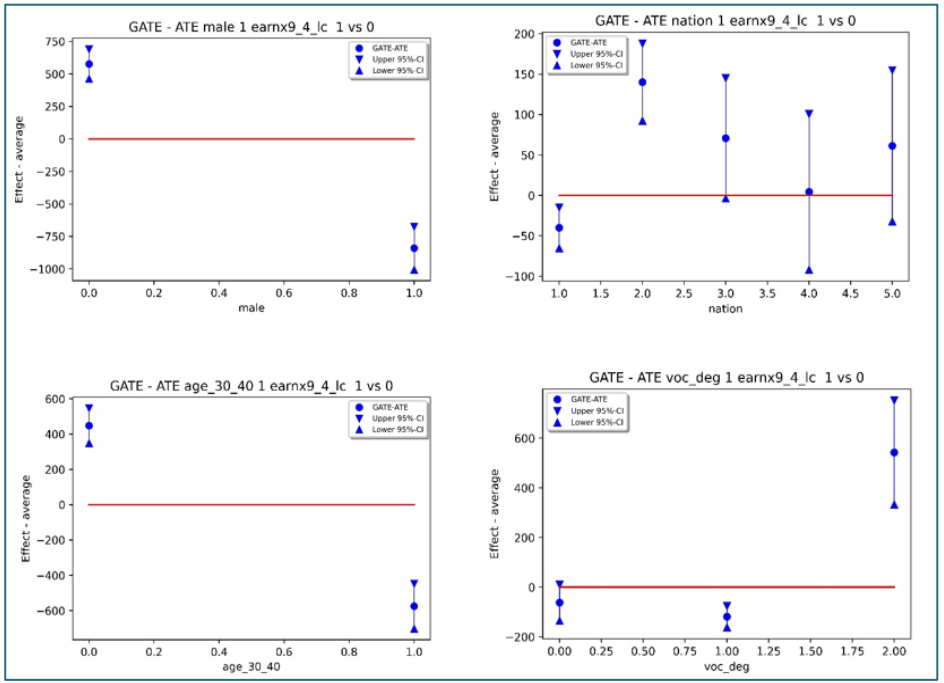
Pour identifier l'hétérogénéité de l'effet, nous estimons d'abord l'effet moyen du programme selon le groupe des participants (EMPGP) pour quatre caractéristiques sociodémographiques : le sexe, l'âge, la nationalité et le niveau d'éducation

- Tous les sous-groupes suivants présentent une EMPGP statistiquement significative sauf les hommes et le groupe d'âge 30-40 ans.



Étape 8 (suite) : Détection de l'hétérogénéité : EMPGP-EMPPP

- ❑ Pour détecter une hétérogénéité des effets, il convient d'examiner la différence entre EMPGP et EMPPP.
- ❑ On observe une hétérogénéité des effets pour certains sous-groupes, l'estimation EMPGP-EMPPP étant statistiquement significative pour ces derniers.



Étape 9 : Interprétation de l'hétérogénéité de l'effet du traitement

- Une fois l'hétérogénéité de l'effet détectée, il est important d'en identifier la source.
- Autrement dit, nous cherchons à mieux comprendre si cette hétérogénéité est due à la variable d'intérêt ou à des facteurs de confusion.
- Pour illustrer l'interprétation de cette hétérogénéité, nous nous concentrons sur les sous-groupes de participants de nationalité européenne et de sexe féminin.
- Deux approches permettent d'interpréter cette hétérogénéité :
  - Analyse du profil des sous-groupes
  - Équilibrage des effets moyens du programme selon le groupe des participants (EMPGP)

Étape 9 (suite) : Interprétation de l'hétérogénéité de l'effet du traitement

- Lorsqu'on détecte une hétérogénéité des effets à l'aide de l'estimation EMPGP-EMPPP, les covariables entre les sous-groupes (définis par une caractéristique comme le sexe) ne sont pas équilibrées.
- Si l'on équilibre les covariables des sous-groupes lors de l'estimation des EMPGP, on obtient une nouvelle estimation appelée Effets moyens du programme selon le groupe des participants équilibrés (EMPGPE). Pour détecter une hétérogénéité des effets à l'aide de l'estimation EMPGPE - EMPPP, on peut suivre la règle générale suivante :

- **Règle générale:**
  - Une fois les EMPGP équilibrés par rapport à toutes les covariables* et si l'estimation EMPGPE-EMPPP devient statistiquement non significative, on peut conclure à l'absence d'hétérogénéité des effets.
  - Toutefois, si la différence EMPGPE-EMPPP reste statistiquement significative, nous pouvons conclure qu'il existe des preuves d'hétérogénéité de l'effet pour cette variable d'intérêt.

*Les facteurs de confusion peuvent être corrélés, ce qui rend difficile d'identifier précisément quelles covariables contribuent le plus à l'hétérogénéité des effets. Comme notre intérêt stratégique est de déterminer si la principale source d'hétérogénéité provient de la variable d'intérêt ou des facteurs de confusion, nous proposons d'équilibrer l'ensemble des covariables, puis de vérifier si l'hétérogénéité des effets persiste.

Étape 9 (suite) : Interprétation de l'hétérogénéité de l'effet du traitement

- **Non Européen:** L'estimation EMPGPE – EMPPP n'étant pas statistiquement significative après ajustement de toutes les covariables, nous pouvons conclure à l'absence d'hétérogénéité d'effet liée à la nationalité pour le sous-groupe non européen.
- **Sexe Féminin:** Le EMPGPE -EMPPP étant statistiquement significatif après ajustement pour toutes les covariables, nous pouvons conclure à l'existence d'une hétérogénéité de l'effet liée au sexe dans ce sous-groupe.

Paramètre	Variable d'équilibrage	Non européen		Sexe Féminin	
		Estimation du coefficient	Valeur P	Estimation du coefficient	Valeur P
EMPGP-EMPPP	Aucun	140	0	566	0
EMPGPE – EMPPP	Informations sur le marché du travail et caractéristiques sociodémographiques (toutes les covariables)	63	**0.15**	700	**0**

Étape 10 : Analyse après estimation par regroupement

- Pour déterminer le groupe qui bénéficie le plus du programme, nous avons effectué une analyse de regroupement par la méthode des k-moyennes. Les résultats suggèrent l'existence de trois groupes.
  - Les participants du troisième groupe sont ceux qui bénéficient le plus du programme
  - Ce groupe se caractérise par une prédominance de femmes âgées de 40 à 50 ans, diplômées de l'enseignement supérieur et ayant un fort taux d'insertion professionnelle (revenus plus élevés avant leur participation)

VARIABLES	GROUPE 1	GROUPE 2	GROUPE 3
EMPP ESTIMÉ	-1034.85	542.64	**1511.85**
NOMBRE D' OBSERVATIONS	792	2382	**2264**
AGE_30_40	0.98	0.53	**0.16**
MALE	1	0.54	**0.04**
NIVEAU D'ÉDUCATION	0.85	0.96	**1.22**
EARNX1_1	2086.74	2239.39	**2468.69**
EARNX1_2	2256.3	2509.83	**2840.78**
EARNX1_3	1899.65	1995.95	**2285.88**
EARNX1_4	1267.68	1310.99	**1573.79**

Conclusion

- L'algorithme FCM est une approche alternative aux méthodes traditionnelles d'appariement pour estimer l'impact des programmes du marché du travail.

- Le FCM présente plusieurs avantages par rapport à l'appariement, notamment la détection de l'hétérogénéité des effets, impossible avec les méthodes d'appariement classiques. L'identification de cette hétérogénéité a des implications importantes pour le développement de politiques.

- L'approche FCM peut également être appliquée à d'autres domaines, tels que la santé et les entreprises, pour détecter l'hétérogénéité des effets.

Merci !

Question?

jamil.sayeed@hrsdc-rhdcc.gc.ca
andy.handouyahia@hrsdc-rhdcc.gc.ca
essolaba.aouli@hrsdc-rhdcc.gc.ca