

Méthodes de différence en différence

Université d'été en Sciences Sociales

Tany Vao 2022 - Madagascar

Florent Bédecarrats, Jeanne de Montalembert,
Marc Bouvier, Marin Ferry, Kenneth Hougbedji

Université de Toliara, Octobre 2022



Principe des méthodes de différence en différence

Est ce que les Aires Protégées augmentent le couvert forestier?

On veut connaître :

$$E[CF1|AP = 1] - E[CF0|AP = 1]$$

avec $CF1$ le couvert forestier s'il y a une AP et $CF0$ le couvert forestier s'il n'y a pas d'AP **au même moment**.

Autrement dit, $E[Y1|T = 1] - E[Y0|T = 1]$


On n'observe pas simultanément les deux: quel **contrefactuel** ?

- Différents espaces ? [▶ démo](#)
- Avant, après ? [▶ démo](#)

⇒ Combinaison des deux: double différence, différence en différence (DD, DiD)

Principe des méthodes de différence en différence

Doubles différences

Espace	Temps	Couvert forestier	D_1	D_2
Makay	Avant	$CF_{av} = M$	$T + AP$	AP 
	Après	$CF_{ap} = M + T + AP$		
Autre	Avant	$CF_{av} = A$	T	
	Après	$CF_{ap} = A + T$		

⇒ D_1 : différence avant-après *i.e* supprime les différences entre le *Makay* et *Autre*

⇒ D_2 : différence des différences pour supprimer les effets de tendance

Principe des méthodes de différence en différence

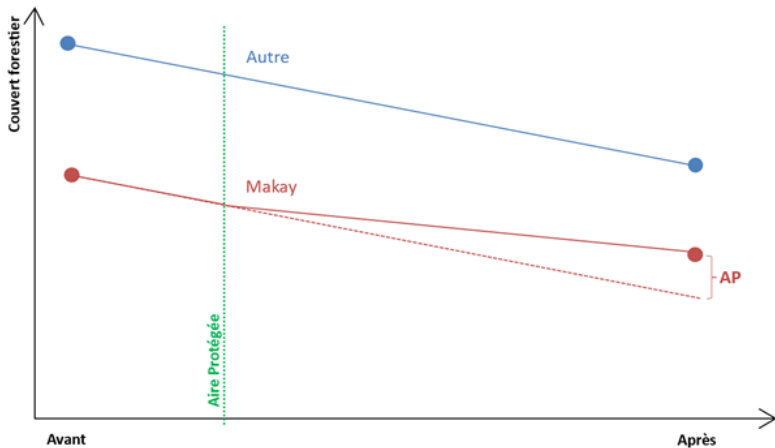


Figure 1: Illustration DiD

Hypothèses centrales

Hypothèses sous-jacente:

- Pas de facteurs non observables qui varient au cours du temps spécifique à l'espace et qui impactent la déforestation *i.e* $\gamma_{i,t}$ seulement γ_t . Autrement dit, la variation temporelle est identique entre les deux espaces au cours du temps.

⇒ **Tendances temporelles communes, tendances parallèles**

- La composition des deux groupes doit rester identique avant et après l'intervention *i.e* l'échantillon reste le même entre les différentes dates (avant et après l'intervention).

⇒ **Composition stable des deux groupes**

Tendances parallèles

L'hypothèse de la tendance parallèle

- En l'absence de toute mise en place d'Aire Protégée, la tendance du couvert forestier dans un espace non protégé aurait été celle à laquelle on aurait dû s'attendre à voir dans un espace protégé.

→ *L'évolution du couvert forestier dans un espace non protégé sur la période constitue un **contrefactuel** fiable du couvert forestier de l'espace protégé i.e le couvert forestier de l'espace protégé **aurait connu la même évolution** que celle de l'espace non protégée s'il n'y avait pas eu la mise en place la politique.*

⇒ Les tendances temporelles ont le même effet sur la variable dépendante pour les deux groupes i.e effets fixes temporels constants

Tendances parallèles

Vérification

- Impossible de validé (ou invalidé) complètement cette hypothèse.
- Tests (indirects) pour valider l'hypothèse:
 - Comparer les tendances avant la mise en place de la politique *i.e* tester si les pentes sont parallèles avant l'intervention pour les deux groupes.
 - Faire un test "placebo" *i.e* appliquer la même technique d'estimation autour d'une date quelconque. Si l'hypothèse est vérifiée, il ne devrait y avoir aucun effet significatif.
 - Estimateur de triple différence avec un deuxième groupe de contrôle (Robustesse des résultats)

Vérification graphique

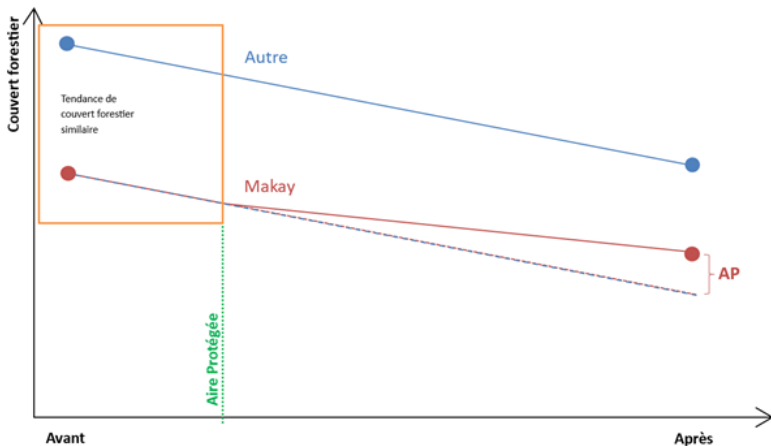


Figure 2: Validation de l'hypothèse des tendances communes

Vérification graphique

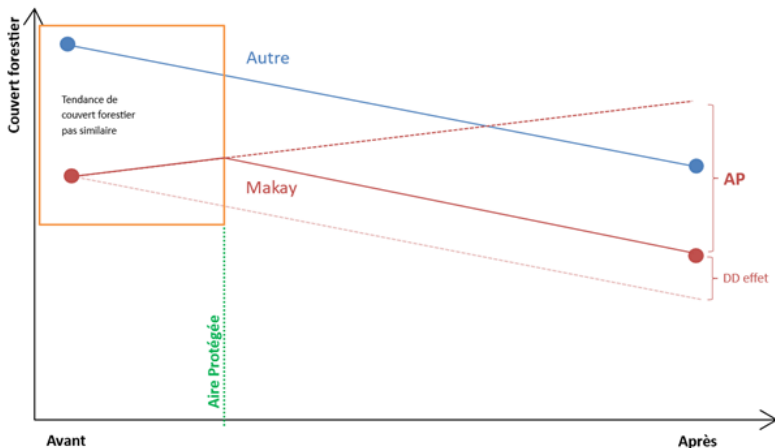


Figure 3: Invalidation de l'hypothèse des tendances communes

Composition stable

L'hypothèse de composition stable des deux groupes

- Les mêmes unités composent les traités et les contrôles aux différentes dates.

→ Il faut nécessaire deux points dans le temps et donc des données:

- longitudinales
- coupe transversale répétée

Sinon:

- Les effets fixes liés aux caractéristiques de l'espace (M et A) ne peuvent pas être supprimé avec la première différence
→ Ce risque est d'autant plus fort avec des données en coupe transversale répétées

Formalisation

Supposons:

- $CF1_{it}$: taux de couvert forestier dans l'espace i à la date t si il est protégé
- $CF0_{it}$: taux de couvert forestier dans l'espace i à la date t si il n'est pas protégé

⇒ les résultats potentiels
mais on peut en observer un à la fois.

L'hypothèse centrale des DiD est que, en l'absence du traitement, le couvert forestier dans l'espace traités i à la date t serait:

$$E[CF0_{it} | T = 1] = \beta_i + \gamma_t$$

Deux hypothèses implicites:

- Biais de sélection: lié aux caractéristiques fixes de l'espace (β)
- Tendance temporelle: identique pour le traitement et le groupe de contrôle (γ)

Estimation

- Variable de **traitement**, dichotomique:

$$TREAT_i = \begin{cases} 0 & \text{si } i = \text{Non protégé} \\ 1 & \text{si } i = \text{Protégé} \end{cases}$$

- Variable **post-traitement**, dichotomique:

$$POST_t = \begin{cases} 0 & \text{si } t < \text{Date de protection} \\ 1 & \text{si } t \geq \text{Date de protection} \end{cases}$$

- **Terme d'interaction** entre les deux *i.e* $TREAT_i \times POST_t$ où le coefficient associé correspond à l'effet causal estimé par DiD

Estimation

Et donc, on a:

$$CF_{it} = \alpha + \beta TREAT_i + \gamma POST_t + \underbrace{\lambda}_{\text{effet causal}} (TREAT_i \times POST_t) + \epsilon_{i,t}$$

Sachant que:

$$E(CF_{it} | TREAT_i = 0, POST_t = 0) = \alpha$$

$$E(CF_{it} | TREAT_i = 0, POST_t = 1) = \alpha + \gamma$$

$$E(CF_{it} | TREAT_i = 1, POST_t = 0) = \alpha + \beta$$

$$E(CF_{it} | TREAT_i = 1, POST_t = 1) = \alpha + \beta + \gamma + \lambda$$

$$\begin{aligned} & \underbrace{E(CF_{it} | TREAT_i = 1, POST_t = 1)}_{\alpha + \beta + \gamma + \lambda} - \underbrace{E(CF_{it} | TREAT_i = 1, POST_t = 0)}_{\alpha + \beta} \\ & - \underbrace{(E(CF_{it} | TREAT_i = 0, POST_t = 1))}_{\alpha + \gamma} - \underbrace{E(CF_{it} | TREAT_i = 0, POST_t = 0)}_{\alpha} = \lambda \end{aligned}$$

Estimation

Et donc, on a:

$$CF_{it} = \alpha + \beta TREAT_i + \gamma POST_t + \underbrace{\lambda}_{\text{effet causal}} (TREAT_i \times POST_t) + \epsilon_{i,t}$$

Comment choisir un groupe de contrôle?

Pour que la méthode soit valide, il faut que le groupe de contrôle présente les mêmes évolutions historiques de la variable dépendante (e.g le couvert forestier) que le groupe de traitement.

Deux possibilités:

- Identifier un groupe avec une "expérience naturelle" (variable exogène)
- Construire le groupe de manière artificielle (matching avec PSM)

Résumé

Pour mettre en oeuvre la double différence, il faut:

- Au moins 2 observations par unités (avant et après intervention): **coupe transversale répétées** ou **panel**
- Au moins 2 groupes: traité et contrôle
- Comparer les tendances temporelles communes entre les deux groupes avant le traitement

Résumé

La crédibilité des estimateurs estimés repose sur l'hypothèse des tendances temporelles communes en l'absence du traitement qui n'est pas vérifiable.

On peut seulement argumenter en faveur de cette hypothèse au moyen de suggestions graphiques et d'autres tests statistiques.

Principe des méthodes de différence en différence

Différent espace?

Espace	Couvert forestier
Makay	$CF_M = M + AP$
Autre	$CF_A = A$

Note: M représente l'effet fixe *Makay* et A l'effet fixe de l'*Autre*

Effet causal: $CF_M - CF_A = M + \underbrace{AP}_{\text{effet}} - A$

$$\rightarrow M - A$$

$M - A$ correspond aux différences de couvert forestier sans AP

⇒ **biais**

Principe des méthodes de différence en différence

Avant, Après?

Espace	Temps	Couvert forestier
Makay	Avant	$CF_{av} = M$
Makay	Après	$CF_{ap} = M + T + AP$

Note: T représente les variables temporelles qui influence CV

Principe des méthodes de différence en différence

Avant, Après?

Espace	Temps	Couvert forestier
Makay	Avant	$CF_{av} = M$
Makay	Après	$CF_{ap} = M + T + AP$

Note: T représente les variables temporelles qui influence CV

Effet causal: $CF_{ap} - CF_{av} = M + T + \underbrace{AP}_{\text{effet}} - M$

$$\rightarrow T$$

T correspond aux différentes variables omises pouvant impacter la déforestation \Rightarrow **biais**