

Recovering Causal Relations with Observational Data

(Thiết lập quan hệ nhân quả trong đánh giá tác động chính sách với
dữ liệu quan sát được)

Lê Việt Phú
Fulbright School of Public Policy and Management

13-17/3/2023

Potential outcome framework

$$ATE = \underbrace{E(Y_i^1|D=1) - E(Y_i^0|D=1)}_{ATT} + \underbrace{E(Y_i^0|D=1) - E(Y_i^0|D=0)}_{Bias}$$

$$ATE = ATT + Selection Bias$$

Đối với thử nghiệm ngẫu nhiên đảm bảo việc phân bổ vào nhóm tham gia hay đối chứng hoàn toàn độc lập với kết quả chương trình:

$$Y_i^1, Y_i^0 \perp D_i$$

thì chúng ta ước lượng được tác động can thiệp trung bình bằng sự khác biệt về kết quả của hai nhóm:

$$ATE = E[Y_i^1 - Y_i^0] = \frac{1}{N} \sum_{i=1}^N (Y_i^1 - Y_i^0)$$

Ước lượng tác động can thiệp như thế nào?

$$ATE = ATT + Bias$$

- ▶ ATE là khác biệt về mặt kết quả giữa nhóm tham gia và không tham gia chương trình.
- ▶ ATT là khác biệt giữa kết quả của nhóm tham gia với kết quả đáng lẽ đã xảy ra nếu như nhóm này không tham gia chương trình.
- ▶ Nếu $Bias = 0$ thì ATE trùng với ATT .

Thiết kế can thiệp ngẫu nhiên (RCT) tạo nhóm hưởng lợi và nhóm đối chứng hoàn toàn tương đồng về các điều kiện quan sát được và không quan sát được $\Rightarrow Bias = 0 \Rightarrow ATE = ATT$. RCT được coi là tiêu chuẩn vàng để thiết lập quan hệ nhân quả giữa can thiệp và kết quả.

Khi không thể thực hiện được RCT

- ▶ Có thể xảy ra hiện tượng lựa chọn mẫu (selection into treatment)
- ▶ Khi xác xuất phân bổ vào nhóm tham gia hay đối chứng tương quan với kết quả chương trình, $Y_i^1, Y_i^0 \sim D_i$

⇒ $ATE \neq ATT$

Chúng ta bắt buộc phải sử dụng dữ liệu quan sát được (observational data) nhằm *xây dựng một tình huống nghiên cứu tương tự như thử nghiệm ngẫu nhiên*:

- Sử dụng các thuật toán thống kê để xây dựng nhóm hưởng lợi và đối chứng tương đồng như thử nghiệm ngẫu nhiên (DiD, matching).
- Sử dụng tình huống thử nghiệm tự nhiên/bán thử nghiệm nhằm mô phỏng lại thiết kế thử nghiệm ngẫu nhiên (Regression Discontinuity, IV, regression adjustment)

⇒ Cần nắm vững lý thuyết đánh giá tác động can thiệp bằng RCT!

Đặc điểm của dữ liệu quan sát được

Đại đa số các dữ liệu điều tra thu thập thứ cấp không đảm bảo điều kiện các nhóm hưởng lợi và đối chứng hoàn toàn tương đồng về tất cả các phương diện.

- ▶ Không tương đồng về các điều kiện quan sát được (ví dụ nhóm hưởng lợi và nhóm đối chứng có các thuộc tính nhân khẩu học, kinh tế xã hội, điều kiện địa lý không giống nhau.)
- ▶ Không tương đồng về các điều kiện không quan sát được.

⇒ Bias luôn tồn tại, và không thể ước lượng được ATT trực tiếp từ mẫu mà phải dùng các thiết kế nghiên cứu hợp lý.

Quá trình phân bổ nhóm đối tượng hưởng lợi và đối chứng quyết định khi nào thì thiết lập được quan hệ nhân quả và sử dụng thiết kế gì

- ▶ Với thử nghiệm ngẫu nhiên: ngẫu nhiên hóa quá trình lựa chọn đối tượng tham gia.
- ▶ Với dữ liệu quan sát được: không đảm bảo việc tham gia là ngẫu nhiên và có thể xảy ra quá trình lựa chọn mẫu.
 - Nếu quá trình tham gia là ngẫu nhiên, không phụ thuộc ý muốn của đối tượng nghiên cứu → Tình huống thử nghiệm tự nhiên → Có thể mô phỏng gần giống nhất với thử nghiệm RCT nhất!
 - Nếu phân bổ can thiệp là ngẫu nhiên nhưng trong quá trình tham gia có thể xảy ra hiện tượng lựa chọn mẫu (self selection into treatment) → Phải có thiết kế nghiên cứu phù hợp với nguyên nhân gây ra hiện tượng tự lựa chọn mẫu.

Thế nào là tình huống thử nghiệm tự nhiên (natural experiment)/bán thử nghiệm (quasi-experiment)?

- ▶ Khi xảy ra một sự kiện hay một can thiệp chính sách mà ở đó có sự phân định ngẫu nhiên nhóm đối chứng và nhóm hưởng lợi, mặc dù không đảm bảo tất cả các thuộc tính của hai nhóm hoàn toàn tương đồng.¹
- ▶ Sự phân định ngẫu nhiên này làm giảm hoặc loại trừ vấn đề tự lựa chọn vào mẫu (self selection into treatment).

¹Học viên cần phân biệt hai khái niệm: phân bổ can thiệp, và tham gia trên thực tế.

Tại sao lại gọi là thử nghiệm tự nhiên (NE)/bán thử nghiệm (QE)?

- ▶ RCT có hai nội dung chính: chọn mẫu các cá nhân tương đồng, sau đó ngẫu nhiên hóa can thiệp trên mẫu đã chọn. Do đó, RCT đảm bảo:
 1. Mẫu tương đồng về các đặc tính quan sát được và không quan sát được (với điều kiện cỡ mẫu đủ lớn).
 2. Can thiệp là ngẫu nhiên, và không có quá trình tự lựa chọn vào nhóm hưởng lợi.
- ▶ NE/QE đảm bảo được điều kiện 2 là không có quá trình tự lựa chọn vào nhóm hưởng lợi, nhưng không đảm bảo điều kiện 1 là mẫu tương đồng.

Ví dụ tình huống thử nghiệm tự nhiên/bán thử nghiệm

Các chính sách vĩ mô cấp độ trung ương hay tỉnh, các hiện tượng thời tiết bất thường, bất khả kháng thường là các tình huống thử nghiệm tự nhiên đối với doanh nghiệp và hộ gia đình:

- Chính quyền quyết định áp thuế trước bạ lên sở hữu xe cộ ở nội thành cao hơn khu vực ngoại ô để hạn chế xe cộ. Chính sách này là thử nghiệm tự nhiên bởi người dân không được quyền lựa chọn mức thuế khi đi mua xe.
- Covid-19 là tình huống thử nghiệm tự nhiên; hạn hán, nắng nóng tác động lên năng suất cây trồng cũng là thử nghiệm tự nhiên bởi vùng bị ảnh hưởng và không bị ảnh hưởng hoàn toàn nằm ngoài khả năng chi phối của các hộ sản xuất sống trong đó.
- Các tình huống áp dụng chính sách theo các tiêu chí độc lập, không phụ thuộc vào ý chí chủ quan của người dân, ví dụ tiền lương tối thiểu vùng là thử nghiệm tự nhiên.

Đặc tính của các ví dụ thử nghiệm tự nhiên trên là gì?²

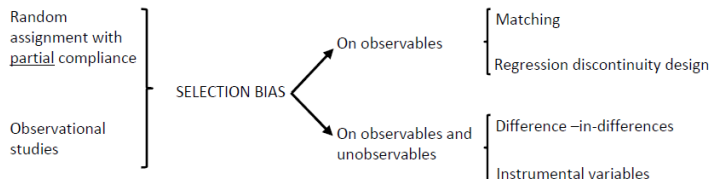
- Người bị ảnh hưởng không được tự lựa chọn mức thuế nộp. Nếu sống ở nội thành thì phải đóng thuế cao (treatment), ở ngoại thành thì thuế thấp (control). Và về mặt trung bình thì người dân ở nội thành và ngoại thành sẽ khác nhau.
- Khu vực bị tác động bởi Covid-19 hay hạn hán là ngẫu nhiên, nằm ngoài tầm kiểm soát của các cá nhân sống tại đó. Và vùng bị ảnh hưởng hay không bị ảnh hưởng có thể rất khác biệt về các đặc tính.

Người lao động hoặc chủ lao động phải trả chi phí BHXH theo vùng chứ không được tự lựa chọn. Và về mặt trung bình thì giữa các vùng cũng có sự khác biệt.

²Các giải thích này chỉ áp dụng bài toán đánh giá trong ngắn hạn. Về dài hạn, khi xảy ra vấn đề sorting thì vẫn có thể dẫn đến vấn đề tự lựa chọn. ▶

Hệ thống hóa phương pháp đánh giá tác động chính sách

Random assignment with *average* → NO SELECTION BIAS → ATT=(Average effect on treated) – (Average effect on non-treated)



Một số lựa chọn khi sử dụng dữ liệu quan sát được

- Tìm cách thiết kế nhóm đối chứng sao cho các đặc tính không quan sát được có thể cân bằng (ví dụ sử dụng thử nghiệm tự nhiên - "treatment is as-if random").
- Chấp nhận có sự khác biệt về đặc tính không quan sát được, nhưng nếu chúng không thay đổi theo thời gian (time invariant unobservables) thì có thể dùng sai phân dữ liệu để loại bỏ.
- Sử dụng phương pháp biến công cụ, với biến công cụ là cơ chế phân bổ can thiệp tương quan với tình trạng hưởng lợi nhưng không tương quan với kết quả để tính tác động dự định can thiệp ITT (tương tự như thiết kế thử nghiệm khuyến khích).
- Ghép cặp (matching) hoặc tạo nhóm đối chứng nhân tạo (synthetic controls) để xây dựng nhóm đối chứng.
- Sử dụng hồi quy gián đoạn (regression discontinuity design) để xây dựng nhóm đối chứng tại ngưỡng gián đoạn của chính sách.

Tất cả những vấn đề trên phải được thảo luận khi đề xuất một nghiên cứu sử dụng dữ liệu quan sát được.

Phân loại phương pháp đánh giá tác động với thử nghiệm tự nhiên/bán thử nghiệm

Lựa chọn phương pháp phụ thuộc vào nguyên nhân xảy ra selection bias.

- **Chệch lựa chọn quan sát được (SOO design): Regression Adjustment, Matching, SRDD**

- ▶ Matching: Sử dụng các thuật toán thống kê để xây dựng nhóm hưởng lợi và đối chứng tương đồng như thử nghiệm ngẫu nhiên. Một trường hợp không cần thiết kế mẫu ngẫu nhiên là sử dụng số liệu của cặp sinh đôi (còn gọi là ghép cặp hoàn hảo, exact matching): Các cặp sinh đôi có cùng yếu tố về năng lực hay các đặc tính sinh học nhưng bị ảnh hưởng bởi các điều kiện kinh tế, xã hội, gia đình khác nhau.

- **Chệch lựa chọn không quan sát được (SOU design): DiD, IV, Heckman sample selection, FRDD**

Thiết kế đơn giản nhất với thử nghiệm tự nhiên

- ▶ **Phương pháp điều chỉnh hồi quy (regression adjustment/control function approach)** giả định rằng sau khi đã kiểm soát tất cả các đặc tính quan sát được thì trạng thái tham gia chương trình hoàn toàn ngẫu nhiên (“treatment is as-if random after controlling for all observables”, $Y_i^1, Y_i^0 \perp D_i | X$).
- ▶ Khi này, chúng ta có thể ước lượng ATE bằng hồi quy tuyến tính đa biến, và tác động can thiệp là tham số của biến chính sách D_i :

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 x_{1i} + \beta_3 x_{2i} + \dots + \varepsilon_i$$

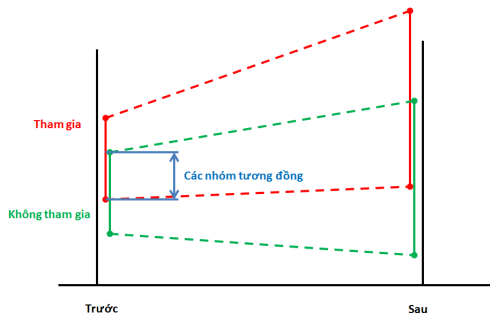
Vẫn cần mô hình thỏa các điều kiện CLRM.

- ▶ Phương pháp này đặc biệt hữu ích với các thử nghiệm tự nhiên (ví dụ đánh giá tác động của hiện tượng thời tiết cực đoan đến năng suất nông nghiệp).

Thiết lập quan hệ nhân quả bằng matching

- ▶ Matching là một thiết kế nghiên cứu dựa trên giả định quan sát được đặc tính giải thích cho vấn đề lựa chọn mẫu (selection on observables).
- ▶ Matching khác với hồi quy là không dựa trên tham số (nonparametric methods).
- ▶ Có rất nhiều phương pháp matching, tuy nhiên cốt lõi của tất cả các phương pháp là đảm bảo điều kiện cân bằng giữa hai nhóm hưởng lợi và đối chứng.
 - Chúng ta sẽ làm quen với cách ghép cặp, từ đơn giản là ghép bằng chỉ số xu hướng (propensity score matching) đến ghép bằng hàm khoảng cách tổng quát Mahalanobis distance, nhóm đối chứng nhân tạo synthetic controls, và ghép cặp bằng cách làm thô nhóm dữ liệu (coarsened exact matching).

Phương pháp ghép cặp bằng điểm xu hướng (propensity score matching-PSM)



- ▶ Bản chất của PSM là tìm nhóm hưởng lợi (treatment group) và nhóm đối chứng (control group) có tương đồng với nhau trong tổng thể mẫu dữ liệu bằng mô hình thống kê.
- ▶ Chúng ta xây dựng một chỉ số gọi là điểm xu hướng - propensity score. Điểm xu hướng là xác suất quan sát được một hộ có tham gia chính sách hay không.
- ▶ Các hộ gia đình có cùng điểm xu hướng gần nhau được ghép cặp và làm đối chứng cho nhau.

Cơ chế của phương pháp ghép cặp bằng propensity score

Các hộ gia đình có xác suất tham gia giống nhau thì có đặc tính tương đồng nhau. Do đó thay vì việc yêu cầu nhóm tham gia và hưởng lợi phải tương đồng nhau về mọi mặt (trong RCT), thì có thể dùng xác suất tham gia để tìm nhóm hưởng lợi và đối chứng.

Nhóm tham gia		Nhóm không tham gia	
HHid	prob.	HHid	prob.
A	.90	M	.10
B	.20	N	.75
C	.70	O	.15
D	.30	P	.90

- ▶ Có thể ghép hộ A (có tham gia) với hộ P (không tham gia), nhưng không thể ghép A với M, N, O.
- ▶ Ghép B với O, C với N.
- ▶ Không tìm được nhóm đối chứng với D và loại bỏ hộ này.
- ▶ Phương pháp Matching ước lượng được ATT chứ không phải là ATE.

Các bước thực hiện phương pháp PSM

1. Ước lượng mô hình xác suất tham gia hay không tham gia chính sách bằng hồi quy logit hay probit, với các biến giải thích là các đặc tính có ảnh hưởng đến khả năng tham gia chương trình:

$$\hat{P}(T = 1|X) = F(X_1, \dots, X_m)$$

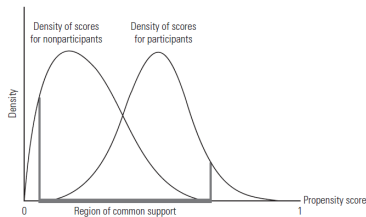
2. Xác định **vùng hỗ trợ chung - common support** và thực hiện các kiểm định đảm bảo **các điều kiện cân bằng** được thỏa mãn. Các nhóm có cùng một giá trị điểm xu hướng (hay khoảng giá trị) cần có các thuộc tính không quá khác biệt nhau.
3. So sánh nhóm hưởng lợi với nhóm đối chứng trong vùng hỗ trợ chung.

Vùng hỗ trợ chung - Common support

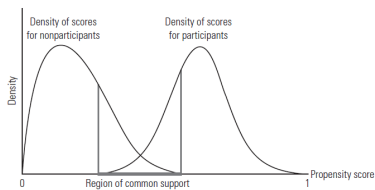
Vùng hỗ trợ chung (còn gọi là điều kiện trùng lặp - overlapping condition) là vùng có ước lượng điểm xu hướng (hoặc xác suất tham gia) của cả nhóm tham gia và nhóm kiểm soát. Hai nhóm phải có một số lượng quan sát có cùng điểm xu hướng thì mới ghép cặp được với nhau.

$$0 < P(T_i = 1|X_i) < 1$$

Vùng hỗ trợ chung tốt



Vùng hỗ trợ chung kém



Điều kiện có vùng hỗ trợ chung

- ▶ Điều kiện này đảm bảo có thể tìm được các quan sát đối chứng đối với một số đối tượng tham gia. Nếu các quan sát có điểm xu hướng khác biệt nhau quá thì không thể so sánh được với nhau.
- ▶ Có càng nhiều đối tượng tham gia và kiểm soát trong vùng hỗ trợ chung càng tốt.
- ▶ Quan sát không nằm trong vùng hỗ trợ chung sẽ bị loại do không tìm được nhóm đối chứng.

Ở ví dụ phía trên, D và M nằm ngoài vùng hỗ trợ chung và bị loại bỏ.

Tính tác động can thiệp trung bình lên đối tượng tham gia trong mô hình PSM

Nếu các điều kiện trên được thỏa mãn, thì tác động can thiệp trung bình đối với những người tham gia (ATT hay ATOT) được tính như sau:

$$ATT_{PSM} = \frac{1}{N_T} \left[\sum_{i \in T} Y_i^T - \sum_{j \in C} \omega(i, j) Y_j^C \right]$$

- ▶ ATT_{PSM} là khác biệt trung bình (có quyền số) giữa nhóm tham gia và nhóm đối chứng có cùng điểm xu hướng.
- ▶ $\omega(i, j)$ là quyền số, tùy thuộc vào phương pháp ghép cặp.
- ▶ Do tồn tại vấn đề lựa chọn mẫu nên $ATE \neq ATT$ trong phương pháp PSM.

Các hình thức xác định nhóm đối chứng và quyền số $\omega(i, j)$ dựa vào điểm xu hướng

1. Phương pháp lựa chọn n láng giềng gần nhất (n nearest neighbor matching): Với mỗi đối tượng hưởng lợi, lựa chọn n quan sát đối chứng có chỉ số xu hướng gần nhất làm phản chứng.
2. Chọn theo bán kính (caliper or radius matching): Ghép quan sát có chỉ số xu hướng nằm trong khoảng $\pm r$ cho trước.
3. Chọn phân tầng và so sánh trong từng khoảng giá trị cho trước (stratification or interval matching).
4. Chọn bằng hồi quy nội tại sử dụng phương pháp phi tham số (kernel and local linear matching).

Ví dụ ghép cặp 1-n

- ▶ Với cơ chế matching 1-1:

$$ATT = \frac{1}{N_1} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

trong đó $Y_{j(i)}$ là biến kết quả của quan sát $j(i)$ có các đặc tính quan sát được $X_{j(i)}$ gần với X_i nhất.

- ▶ Mở rộng matching với M quan sát gần nhất:

$$ATT = \frac{1}{N_1} \sum_{D_i=1} \left\{ Y_i - \left(\frac{1}{M} \sum_{m=1}^M Y_{jm(i)} \right) \right\}$$

Lưu ý đối với phương pháp PSM

- ▶ Chỉ số xu hướng hoàn toàn được xây dựng dựa trên mô hình logit/probit của các đặc tính quan sát được, do đó phương pháp PSM dựa trên giả định việc tham gia chính sách chỉ phụ thuộc vào các đặc tính có thể quan sát được (**selection on observables**). Các đặc tính không quan sát được không ảnh hưởng đến quá trình tham gia hay không tham gia chính sách. Đây gọi là giả định độc lập có điều kiện (*conditional independence hay unconfoundedness*).
- ▶ Đây là một giả định khá mạnh và khó có thể đảm bảo trong hầu hết các trường hợp.

Thực hành

Đánh giá tác động của chính sách cho vay vốn đến tổng chi tiêu bình quân hộ gia đình.

- ▶ Cài đặt chương trình cho STATA ("net install st0026_1.pkg")
- ▶ Sử dụng dữ liệu mô phỏng tín dụng của WB microfinance.dta và matching.do.

Bước 1: Ước lượng điểm xu hướng

Chạy chương trình pscore để tính điểm xu hướng và kiểm tra tính cân bằng giữa nhóm đối chứng và nhóm xử lý:

```
pscore T X1 X2... [pw=weight], pscore(myscore)  
blockid(myblock) numblo(k) comsup
```

- ▶ T là biến chính sách, có tham gia hay không
- ▶ X1 X2 ... là các biến giải thích cho việc tham gia chính sách
- ▶ weight là biến quyền số, thường sử dụng trong các bộ dữ liệu điều tra dân số hay mức sống hộ gia đình
- ▶ myscore là tên biến điểm xu hướng
- ▶ blockid là tên khoảng điểm xu hướng
- ▶ numblo là số lượng block, tùy chọn
- ▶ comsup là biến trạng thái chỉ dẫn số liệu có nằm trong vùng hỗ trợ chung hay không

- ▶ Chương trình pscore sẽ tính xác suất tham gia chương trình dựa vào các biến giải thích $[X_1, X_2, \dots]$ sử dụng hàm logit hay probit. Mặc định của pscore là probit.
- ▶ Kiểm tra điều kiện cân bằng: các biến giải thích phải cân bằng (giá trị trung bình bằng nhau) trong mỗi khoảng giá trị của điểm xu hướng. Điều kiện này đảm bảo nhóm đối chứng tương đồng với nhóm hưởng lợi.
 - Chương trình pscore sẽ kiểm tra vùng hỗ trợ chung, phân khoảng chỉ số xu hướng thành các block, và kiểm tra điều kiện cân bằng trong mỗi block. Nếu thỏa mãn chuyển sang bước 2, nếu không thì hoặc là phân lại block, hoặc là sửa hàm ước lượng ở bước 1. Lặp lại cho đến khi điều kiện cân bằng được đảm bảo.

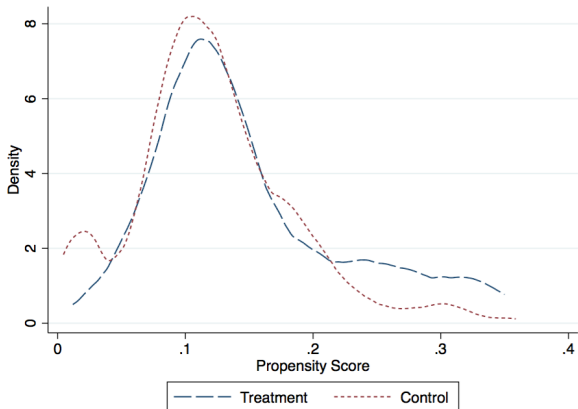
Bước 2: Ước lượng tác động can thiệp

Sau khi đã ước lượng điểm xu hướng và điều kiện cân bằng đã được đảm bảo, tính tác động can thiệp trung bình đối với người tham gia (*ATT*) theo các hình thức ghép cặp khác nhau:

Command	Method
attnd	nearest-neighbor matching
atts	stratification matching
attr	radius matching
atrk	kernel matching

- ▶ `attnd Y T [pw=weight], pscore(myscore) comsup`
- ▶ `atts Y T, pscore(myscore) blockid(myblock) comsup`
- ▶ `attr Y T, pscore(myscore) radius(xxxx) comsup`
- ▶ `atrk Y T, pscore(myscore) comsup bootstrap reps(50)`

Kiểm tra điều kiện overlapping/common support



- ▶ Dữ liệu này cho thấy vùng hỗ trợ chung khá tốt và ước lượng ATT gần với ATE.
- ▶ Nếu vùng hỗ trợ chung kém thì ATT có thể rất khác với ATE và hiệu lực ngoại vi (external validity) của ước lượng yếu.

Ước lượng phương pháp PSM - nâng cao

Tự viết chương trình để tính ra tác động thay vì sử dụng các phần mềm hay các chương trình viết sẵn là phương pháp tốt nhất để học kinh tế lượng. Sơ đồ hóa (pseudocode) của phương pháp PSM gồm các bước sau:

1. Ước lượng propensity score bằng logit hoặc probit
2. Kiểm tra vùng hỗ trợ chung, phân khoảng (block), kiểm tra điều kiện cân bằng trong mỗi block. Nếu thỏa mãn chuyển sang bước 3, nếu không thì hoặc là phân lại block, hoặc là sửa hàm ước lượng ở bước 1. Lặp lại cho đến khi điều kiện cân bằng được đảm bảo.
3. Tính $ATOT$ bằng bình quân với quyền số từ TOT của các block.