The Effects of State Coercion on Voting Outcome in Protest Movements: A Causal Forest Approach

Online Appendix

Weiwen Yin[*], Weidong Huo[†], Danyang Lin[‡]

Contents

[*]First author. Affiliation: Department of Asian and Policy Studies, The Education University of Hong Kong. Email: wwyin@eduhk.hk.

[†]Affiliation: Institute of Guangdong, Hong Kong and Macao Development Studies, Sun Yat-sen University. Email: huowd3@mail.sysu.edu.cn.

[‡]Corresponding author. Affiliation: Institute of Guangdong, Hong Kong and Macao Development Studies, Sun Yat-sen University. Email: lindy23@mail.sysu.edu.cn.

1 Technical Details of Causal Forests

Formally, the conditional average treatment effect (CATE) for a given observation i is defined as:

$$\tau(x) = E[Y_i^{W=1} - Y_i^{W=0} | X_i = x] \tag{1}$$

, where $i = 1, 2, ..., n$ represents the constituencies in the election and $W_i \in \{0, 1\}$ indicates whether the police used tear gas in constituency i. We observe the outcome of interest $Y_i^{W=1}$ if the constituency is assigned to the treatment condition (i.e., if the police used tear gas on the protesters), otherwise we observe $Y_i^{W=0}$. X_i denotes a vector of constituency characteristics.

Like causal inference problems in general, we cannot simultaneously observe both potential outcomes Y_i^0 and Y_i^1 for the same observation i, but only one (Angrist and Pischke 2008). Under unconfoundedness (i.e.,$W_i \perp \{Y_i^0 - Y_i^1\} | X_i$), however, Equation (1) can be rewritten as:

$$\tau(x) = E[Y_i | W_i = 1, X_i = x] - E[Y_i | W_i = 0, X_i = x] \tag{2}$$

This is equivalent to comparing nearby observations in the x-space that are similar to each other in terms of their observable characteristics with the only difference being that only some of them are treated. The purpose is to rule out selection into treatment on observables.

When the unconfoundedness assumption holds, we can consistently estimate the CATEs using the causal forests (CF). Formally, the CF estimates the CATE based on the following formula:

$$\widehat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)(Y_i - \widehat{g}^{(-i)}(X_i))(W_i - \widehat{e}^{(-i)}(X_i))}{\sum_{i=1}^{n} \alpha_i(x)(W_i - \widehat{e}^{(-i)}(X_i))^2} \tag{3}$$

, where Y represents the outcome variable, X represents the control variables, W represents the treatment variable, and $\widehat{g}^{(-i)}(X_i)$ and $\widehat{e}^{(-i)}(X_i)$ are estimators of $g(X_i) \equiv \mathbb{E}[Y_i | X_i]$ and $e(X_i) \equiv \mathbb{E}[W_i | X_i]$,

respectively, without using observation i.[1] Note that when $\alpha_i(x)$ is a constant, the CATE $\widehat{\tau}(x)$ is equivalent to the OLS estimator of regressing $Y_i - \widehat{g}^{(-i)}(X_i)$ on $W_i - \widehat{e}^{(-i)}(X_i)$ without an intercept, which is the effect of the treatment variable on the outcome not explained by X. In other words, the purpose of the CF, same as OLS regressions, is to partial out the confounding effect of the control variables.

$\alpha_i(x)$ is a data-adaptive weight the calculation of which is based on random forests. It is defined as:

$$\alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x),\ \alpha_{bi}(x) = \frac{\mathbb{1}[\{X_i \in L_b(x), i \in \mathcal{S}_b\}]}{|\{i : X_i \in L_b(x), i \in \mathcal{S}_b\}|}, \tag{4}$$

$\alpha_i(x)$ captures the frequency of observation i falling into the same leaf with point x in the forest consisting of B trees, where the frequency is normalized by the leaf size. The intuition is that an individual staying in the same leaf in a tree with x frequently is a "closer" individual to x, and thus receives a higher weight. Therefore, the CF can be viewed as a weighted least squares estimation with the weight being $\alpha_i(x)$. For each CATE estimated using the CF, we can also obtain a consistent estimate of its variance, allowing us to draw inference (Athey, Tibshirani and Wager 2019).

[1] Following the default option in the R package grf, we use random forests to make out-of-bag predictions in the estimation of $\widehat{g}^{(-i)}(X_i)$ and $\widehat{e}^{(-i)}(X_i)$. This process helps to avoid the over-fitting problem.

2 The Location of Tear Gas Usage

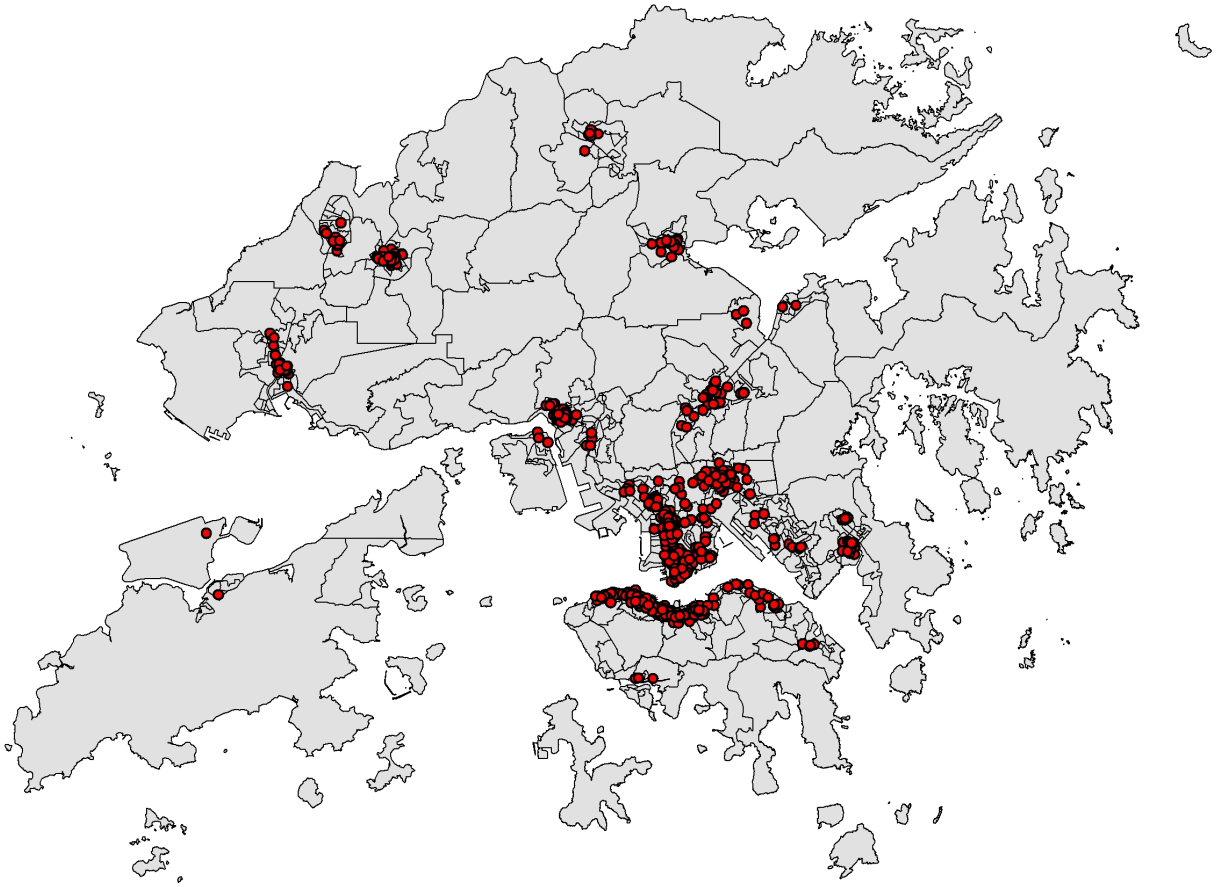

Figure 1: The Location of Tear Gas Usage

3 Definitions of Covariates

Table 1: Definitions of Covariates

Variable	Definition
Treatment	=1 if police used tear gas, 0 otherwise
Pro-democracy vote share	Vote share of pro-democracy candidate
Competing candidates	=1 if more than one pro-democracy candidate, 0 otherwise
Log constituency size	Log of constituency size
Log population density	Log of population density
First-time voters below 35	Percentage of first-time voters below 35
Non-first-time voters below 35	Percentage of non-first-time voters below 35
First-time voters 35-60	Percentage of first-time voters at 35-60
Non-first-time voters 35-60	Percentage of non-first-time voters at 35-60
First-time voters above 61	Percentage of first-time voters above 61
Non-first-time voters above 61	Percentage of non-first-time voters above 61
Female	Percentage of female registered voters
Pro-democracy incumbent	=1 if incumbent is pro-democracy, 0 otherwise
Manufacturing	Percentage of labor in the manufacturing industry
Construction	Percentage of labor in the construction industry
Trade	Percentage of labor in import/export, wholesale and retail trades
Transportation	Percentage of labor in transportation, storage, postal and courier services
Service	Percentage of labor in accommodation and food services
Information	Percentage of labor in information and communication industry
Finance	Percentage of labor in financing and insurance industry
Professional	Percentage of labor in real estate, professional and business services
Administration	Percentage of labor public administration, education, human health and social work activities
Miscellaneous	Percentage of labor in miscellaneous social and personal services
Mandarin speaker	Percentage of Mandarin speakers
Household size	Average household size (unit: person)
Household income	Median monthly domestic household income (unit: 1000 HKD)
Mortgage-to-income ratio	Median mortgage payment and loan repayment to income ratio
Rent-to-income ratio	Median rent to income ratio
Floor area	Median floor area of accommodation (unit: square metres)
Government funded housing	Percentage of population with public rental housing or subsidised home ownership
No. of confrontations	Number of confrontations between police and protesters

4 Summary Statistics of Variables

Table 2: Summary Statistics

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Pro-democracy share	435	0.567	0.069	0.300	0.527	0.610	0.895
Treatment	435	0.366	0.482	0	0	1	1
Competing candidates	435	0.074	0.261	0	0	0	1
Log constituency size	435	−0.303	0.655	−1.114	−0.747	−0.059	2.174
Log population density	435	4.517	0.653	2.151	4.256	4.961	5.373
First-time voters below 35	435	0.046	0.010	0.026	0.039	0.051	0.114
Non-first-time voters below 35	435	0.191	0.042	0.096	0.163	0.209	0.379
First-time voters 35-60	435	0.040	0.015	0.013	0.029	0.047	0.135
Non-first-time voters 35-60	435	0.402	0.058	0.290	0.357	0.443	0.624
First-time voters above 61	435	0.011	0.006	0.002	0.006	0.014	0.040
Female	435	0.516	0.014	0.436	0.509	0.525	0.556
Pro-democracy incumbent	435	0.262	0.440	0	0	1	1
Manufacturing	435	0.038	0.013	0.006	0.029	0.047	0.079
Construction	435	0.087	0.038	0.003	0.058	0.113	0.209
Trade	435	0.189	0.026	0.105	0.173	0.205	0.298
Transportation	435	0.090	0.035	0.012	0.066	0.111	0.220
Service	435	0.084	0.043	0.003	0.050	0.112	0.286
Information	435	0.036	0.013	0.007	0.026	0.043	0.082
Finance	435	0.063	0.038	0.012	0.036	0.081	0.252
Professional	435	0.144	0.026	0.056	0.126	0.163	0.225
Administration	435	0.150	0.039	0.066	0.124	0.174	0.360
Miscellaneous	435	0.113	0.066	0.028	0.070	0.139	0.494
Mandarin speaker	435	0.019	0.018	0.000	0.008	0.024	0.138
Household size	435	2.855	0.323	1.800	2.600	3.100	4.000
Household income	435	29.506	16.556	10.000	19.540	33.750	132.250
Mortgage-to-income ratio	435	0.160	0.072	0.000	0.151	0.200	0.313
Rent-to-income ratio	435	0.197	0.105	0.047	0.093	0.297	0.488
Floor area	435	44.352	17.569	17	34	50	183
Government funded housing	435	0.457	0.414	0	0	0.9	1
No. of confrontations	435	2.414	6.012	0	0	2	93

5 OLS Regression Results

Table 3: OLS Regression Results

	Dependent variable: Vote Share of Pro-democracy Candidate	
	(1)	(2)
Tear gas usage	0.01066** (0.00177, 0.01954)	−0.01722 (−2.33170, 2.29726)
Competing candidates	0.01123 (−0.00350, 0.02596)	−0.01527 (−0.03326, 0.00272)
Log constituency size	−0.02506 (−0.08143, 0.03130)	−0.00004 (−0.06826, 0.06819)
Log population density	−0.01926 (−0.07521, 0.03668)	0.00546 (−0.06252, 0.07344)
First-time voters below 35	0.18156 (−0.52235, 0.88548)	−0.34301 (−1.22008, 0.53405)
Non-first-time voters below 35	0.37179*** (0.20038, 0.54321)	0.62673*** (0.41211, 0.84135)
First-time voters 35-60	−0.13180 (−0.75569, 0.49208)	−0.10063 (−0.89455, 0.69330)
Non-first-time voters 35-60	−0.00642 (−0.12553, 0.11269)	0.13193 (−0.00890, 0.27275)
First-time voters above 61	−2.34708*** (−3.45361, −1.24056)	−1.14939 (−2.52702, 0.22824)
Female	0.12380 (−0.31272, 0.56032)	0.55093* (0.01163, 1.09023)
Pro-democracy incumbent	0.08033*** (0.07132, 0.08934)	0.08387*** (0.07251, 0.09522)
Manufacturing	0.48737 (−0.43995, 1.41470)	0.28627 (−0.74705, 1.31959)
Construction	−0.20470 (−1.12561, 0.71621)	−0.85232 (−1.87660, 0.17197)
Trade	0.03496 (−0.84976, 0.91967)	−0.76371 (−1.74800, 0.22057)
Transportation	0.14134 (−0.77640, 1.05908)	−0.23355 (−1.23909, 0.77198)
Service	0.05757 (−0.82671, 0.94185)	−0.44692 (−1.44281, 0.54896)
Information	0.32000 (−0.63956, 1.27955)	−0.07847 (−1.15795, 1.00101)
Finance	0.14033 (−0.76683, 1.04749)	−0.46255 (−1.47737, 0.55227)
Professional	−0.01823 (−0.93895, 0.90250)	−0.47513 (−1.49077, 0.54051)
Administration	0.10772 (−0.78674, 1.00219)	−0.53096 (−1.51819, 0.45626)
Miscellaneous	−0.08449 (−1.00361, 0.83462)	−0.52046 (−1.52048, 0.47956)
Mandarin speaker	−0.44314** (−0.73819, −0.14808)	−0.26464 (−0.72401, 0.19473)
Household size	0.01292 (−0.00859, 0.03443)	0.02339 (−0.00166, 0.04844)
Household income	0.00062 (−0.00016, 0.00139)	0.00054 (−0.00043, 0.00150)
Mortgage-to-income ratio	0.05042 (−0.01945, 0.12029)	0.07888 (−0.00345, 0.16120)
Rent-to-income ratio	0.07630 (−0.00395, 0.15656)	0.11144* (0.01285, 0.21002)
Floor area	−0.00148*** (−0.00216, −0.00081)	−0.00177*** (−0.00254, −0.00099)
Government funded housing	−0.00341 (−0.02797, 0.02114)	0.00515 (−0.02658, 0.03688)
No. of confrontations	0.00028 (−0.00043, 0.00098)	−0.00030 (−0.00343, 0.00283)
T*Competing candidates		0.06451*** (0.03178, 0.09724)
T*Log constituency size		−0.06559 (−0.19408, 0.06290)
T*Log population density		−0.08332 (−0.21113, 0.04448)
T*First-time voters below 35		0.33977 (−1.20918, 1.88872)
T*Non-first-time voters below 35		−0.49936** (−0.88433, −0.11439)
T*First-time voters 35-60		−0.07328 (−1.36484, 1.21827)
T*Non-first-time voters 35-60		−0.25832 (−0.52688, 0.01024)
T*First-time voters above 61		−2.88493** (−5.22575, −0.54410)
T*Female		−1.36491** (−2.33361, −0.39621)
T*Pro-democracy incumbent		−0.00667 (−0.02563, 0.01228)
T*Manufacturing		0.52907 (−1.68652, 2.74465)
T*Construction		1.47689 (−0.79607, 3.74985)
T*Trade		1.97656 (−0.20817, 4.16130)
T*Transportation		1.15325 (−1.15683, 3.46333)
T*Service		1.31687 (−0.88395, 3.51768)
T*Information		1.29734 (−1.00132, 3.59601)
T*Finance		1.59693 (−0.61277, 3.80663)
T*Professional		1.18784 (−1.10741, 3.48309)
T*Administration		1.51159 (−0.71267, 3.73585)
T*Miscellaneous		1.22085 (−1.08755, 3.52925)
T*Mandarin speaker		−0.48197 (−1.08923, 0.12529)
T*Household size		−0.01082 (−0.06061, 0.03898)
T*Household income		−0.00091 (−0.00268, 0.00087)
T*Mortgage-to-income ratio		−0.14571 (−0.30480, 0.01339)
T*Rent-to-income ratio		−0.15203 (−0.32073, 0.01666)
T*Floor area		0.00071 (−0.00089, 0.00231)
T*Government funded housing		−0.04446 (−0.09638, 0.00747)
T*No. of confrontations		0.00009 (−0.00314, 0.00331)
Constant	0.45829 (−0.43933, 1.35591)	0.56092 (−0.42332, 1.54515)
Observations	435	435

Note: *p<0.1; **p<0.05; ***p<0.01. Standard errors in parenthesis.

6 Full List of Variable Importance

	Variable Importance
Non-first-time voters below 35	0.078
Transportation	0.074
Trade	0.073
Household income	0.058
Manufacturing	0.058
Information	0.053
First-time voters 35-60	0.053
Log population density	0.050
Log constituency size	0.049
First-time voters above 61	0.044
Female	0.040
Administration	0.039
Mandarin speaker	0.039
Service	0.035
First-time voters below 3535	0.035
Professional	0.033
Miscellaneous	0.027
Rent-to-income ratio	0.026
Finance	0.025
Mortgage-to-income ratio	0.024
Construction	0.018
Non-first-time voters 35-60	0.018
Government funded housing	0.012
Household size	0.011
Floor area	0.011
No. of confrontations	0.011
Pro-democracy incumbent	0.005
Competing candidates	0.000

References

Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

Athey, Susan, Julie Tibshirani and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47(2):1148–1178.