Indirect Value of Public Infrastructure Technology

Online Appendix

Jialu Liu, Siqi Pei, Xiaoquan (Michael) Zhang

Management Science

Appendix A: Robustness Checks and Additional Analyses

In Section 4, we confirm the validity of our regression discontinuity design by showing that upstream and downstream firms were well-balanced in the CPES and the ASIF datasets. In this appendix, we conduct additional robustness checks: (1) a manipulation check, (2) alternative estimation methods, (3) alternative radius circles, (4) alternative bandwidth estimation, (5) placebo tests, (6) inclusion of covariates, (7) heterogeneity analysis, and (8) alternative samples.

Manipulation Check

A potential concern about our main analysis is that polluting firms might avoid locating above monitoring stations and instead, they moved downstream to escape regulations and fines. We test the distribution of polluting firms across monitoring stations (Table A7 in the Appendix) following the procedures proposed in Cattaneo et al. (2020). The manipulation check essentially compares the density of polluting firms around the water monitoring stations. If firms were strategically located downstream to avoid detection by the monitoring stations, we would observe fewer polluting firms upstream. However, we find no discontinuity in the distribution of polluting firms across monitoring stations, suggesting that the possibilities of firms relocating to downstream areas do not confound our main results.

Alternative Estimation Methods

In Section 5, we report non-parametric estimation results. To check whether our main result is sensitive to the use of a non-parametric approach, we also report parametric estimates (Gelman and Imbens 2019) in Table A8 in the Appendix. We estimate regression discontinuity using linear, quadratic, and cubic functions to check whether the estimates are sensitive to the order of polynomial functions. The results are consistent with those in Table 2. For private

firms in polluting industries after 2003, the regression discontinuity estimates are negative and statistically significant (Table A8 in the Appendix, column (1) – (3)), while for private firms in non-polluting industries, the estimates are not statistically significant (Table A8 in the Appendix, column (4) – (6)).

In our main result, we use the regression discontinuity method to account for the fact that the closer a firm is to the monitoring station, the more sensitive that firm will be to the complementarities between technology and organizational change. We now consider a difference-in-differences approach to investigate the research question. In Table A9, the interaction term of the upstream dummy and polluting industries dummy remains negative and statistically significant.

We also present the results of an alternative regression discontinuity estimator – Imbens and Wager (2019) estimator. This estimator is fully data-driven and calculates "optimal" weights for each observation. In addition, the estimator is defined regardless of the shape of the treatment region and is not affected by the potential discreteness of the running variable. The results are presented in Table A10 and are consistent with our main results.

Alternative Radius Circles

Recall that in our main analysis, we draw a 10-km circle around the water monitoring stations as our samples. To remove the concerns that our results are sensitive to the choice of a 10-km cutoff, we conduct additional tests with the cutoffs of 20-km (Panel A), and 30-km (Panel B) in Table A11. The results are consistent. We observe statistically significant upstream and downstream corruption gaps for polluting industry (columns (1) – (3)) but not for non-polluting industries (columns (4) – (6)).

Alternative Bandwidth Estimation

The bandwidth chosen in our main result is a common MSE-optimal bandwidth selector (Calonico et al. 2014), which minimizes the asymptotic mean squared error (MSE) of the average effect of the treatment. To check whether our main findings are sensitive to optimal bandwidth selection methods, we use five alternative bandwidth selectors suggested by Calonico et al. (2018) and report the results in Table A12 in the Appendix. They are (1) MSE-two: This method allows for a different bandwidth below and above the cutoff. It applies MSE-

optimal bandwidth selectors method on both sides of the cutoff. (2) MSE-sum: This method uses a common MSE-optimal bandwidth, but the objective function includes the mean squared-errors on both the left and the right of the cutoff point, whereas MSE and MSE-two focus on the mean-squared-error of the difference. (3) CER: This method employs a common bandwidth selector, and aims to minimize the coverage error probability (CER). It is also known as the CER-optimal bandwidth selector. Usually, CER-optimal bandwidth is smaller than the MSE-optimal bandwidth. (4) CER-two: This method allows for different bandwidths below and above the cutoff. It applies CER-optimal bandwidth selectors method on both sides of the cutoff. (5) CER-sum: This method utilizes a common CER-optimal bandwidth, but the objective function includes the sum of mean squared errors on both sides of the cutoff point, whereas in CER and CER-two, the objective function is mean-squared-error of the difference. Technical details are in Calonico et al. (2018). The results in Table A12 in the Appendix are highly consistent with our main results in Table 2, which demonstrates that our main findings are robust to various bandwidth selection methods.

Placebo Tests

We conduct placebo tests by using artificially relocating water monitoring stations. We move the original stations upstream or downstream by 2km (Table A13, Panel A), 3km (Table A13, Panel B), and 4km (Table A13, Panel C), and re-estimate the regression discontinuity models. The results show that the fake relative distance and location between firms and the placebo stations do not cause discontinuity of corruption at the fabricated cutoff. This test (Table A13 in the Appendix) confirms that the discontinuity of corruption exists only in actual monitoring stations, not placebo stations, providing additional evidence supporting our main findings.

Inclusion of covariates

Although we have checked balances between the upstream and downstream firms, we now follow Lee and Lemieux (2010)'s suggestions to include additional covariates. If our research design is valid, the additional covariates should have little effect on the estimation. As additional covariates, we include firm sales, firm value-added tax, the logarithm of the number of employees, the logarithm of one plus firm age, and the logarithm of province per

capita GDP. Table A14 in the Appendix shows results confirming our main findings: polluting firms show negative and statistically significant upstream-downstream gaps (Table A14, columns (1) – (3)), but non-polluting firms show statistically nonsignificant gaps (Table A14, columns (4) – (6)).

Heterogeneity Analysis

Given the large variance among different provinces in China in terms of their local economy, leadership, corruption level, and water quality, we conduct the difference-in-discontinuities analysis to investigate the heterogeneity effect (Table A15). Specifically, we analyze the differences in corruption discontinuity between high GDP regions and low GDP regions (Panel A), the differences in corruption discontinuity between politically motivated leaders and non-politically motivated leaders (Panel B), the differences in corruption discontinuity between centralized regions and less centralized regions (Panel C), the differences in corruption discontinuity between high corruption regions and low corruption regions (Panel D), and the differences in corruption discontinuity between high water pollution regions and low water pollution regions (Panel E). We do not find evidence that corruption discontinuity between upstream and downstream polluting firms differs in terms of their social-economic condition (Panel A), regions' centralization level (Panel C), and water quality (Panel E).

We observe that regions with politically motivated leaders experience larger corruption reduction gaps than regions with non-politically motivated leaders (Panel B). Moreover, high corruption regions show larger corruption reduction gaps than low corruption regions (Panel D). These findings offer additional corroborating evidence.

Alternative Samples

It is possible that upstream firms and downstream firms are governed by different politicians if the water monitoring stations are located at the boundary of provinces. To remove the concern, we conduct the difference-in discontinuities analysis after excluding the water monitoring stations located at the border of the provinces. The results are shown in Table A16, and are consistent with our main results.

In the main analysis, we removed ambiguous firms which are located upstream of one

water monitoring station and at the same time also located downstream of another water monitoring station. To alleviate the concern that our results are sensitive to this, we reconduct the analysis including these ambiguous firms. The results are shown in Table A17 and are consistent with our main results.

In our data cleaning process, we dropped firms with missing ETC. To alleviate concerns about the data cleaning process, we reconduct the analysis including the firms with missing ETC and treated their ETC as 0. The results (Table A18) remain consistent with our main results.

Appendix B: Tables

Table A1. Link Between ETC and Firms' Actual Misconduct

	Number of Regulation Breaches
ETC	0.014***
	(0.002)
Log Likelihood	-24811.4
Observations	32,459

Note: The dependent variable is the number of regulation breaches for a firm in a year from China Stock Market & Accounting Research Database. The independent variable is ETC (million RMB) from Wind Database. Poisson regression is employed. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A2. Covariate Balance Between Upstream and Downstream Firms

	Mean		Mean Difference
	Downstream	Upstream	
	(1)	(2)	(3)
Panel A: ASIF			
Year of Opening	1994.12	1993.02	1.100
	(11.98)	(13.90)	(1.082)
Polluting industries	0.28	0.28	-0.001
(1=Yes, 0=Others)	(0.45)	(0.45)	(0.023)
Profit	5583.67	4451.76	1131.909
(1,000 RMB)	(144712.98)	(140953.98)	(2514.313)
Value-Added Tax	3685.50	3072.04	613.459
(1,000 RMB)	(42045.42)	(34363.86)	(702.502)
# of Employees	258.11	240.56	17.541
(Male)	(1042.80)	(1078.64)	(33.022)
# of Employees	102.83	95.38	7.454
(Female)	(332.37)	(360.07)	(12.175)
Capital Stock	40620.32	26047.07	14573.255
(1,000 RMB)	(658851.78)	(155018.10)	(8121.275)
Intermediate Input	75408.55	71480.62	3927.938
(1,000 RMB)	(874074.76)	(572838.81)	(14806.010)
Panel B: CPES			
Year of Opening	1998.99	1997.99	0.998
	(5.77)	(5.24)	(0.635)
Sales	7,796.03	6,394.30	1,401.732
(10000 RMB)	(102,832.91)	(67,007.83)	(2,510.164)
Tax	268.30	277.56	-9.267
(10000 RMB)	(1,740.17)	(3,570.98)	(84.882)
Profit	279.99	354.42	-74.426
(10000 RMB)	(1,831.14)	(3,349.80)	(87.264)

Note: Columns (1)–(2) report the means and standard deviations of firm characteristics. In column (3), we test the covariate balance between upstream and downstream firms. The difference coefficients are obtained by running OLS regressions of firm characteristics on an upstream dummy. Standard errors reported in the parentheses are clustered at the water monitoring station level. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A3. Industry Balance Between Upstream and Downstream Firms

	Mean		Mean Difference
	Downstream	Upstream	
	(1)	(2)	(3)
Agricultural and Sideline Food Processing	0.03	0.02	0.011
Ind. Code: 13	(0.18)	(0.14)	(0.007)
Food Manufacturing	0.02	0.02	0.001
Ind. Code: 14	(0.15)	(0.15)	(0.007)
Beverage Manufacturing	0.01	0.01	-0.003
Ind. Code: 15	(0.10)	(0.11)	(0.004)
Textile Mills	0.09	0.07	0.017
Ind. Code: 17	(0.29)	(0.26)	(0.036)
Apparel and Clothing Accessories Manufacturing	0.05	0.06	-0.014
Ind. Code: 18	(0.21)	(0.24)	(0.022)
Leather, Fur, and Related Product Manufacturing	0.04	0.01	0.031
Ind. Code: 19	(0.19)	(0.08)	(0.016)
Wood and Bamboo Products Manufacturing	0.01	0.01	0.005
Ind. Code: 20	(0.11)	(0.08)	(0.003)
Furniture Manufacturing	0.02	0.00	0.013*
Ind. Code: 21	(0.13)	(0.07)	(0.005)
Paper Products Manufacturing	0.02	0.02	0.006
Ind. Code: 22	(0.15)	(0.13)	(0.006)
Printing and Reproduction of Recorded Media	0.03	0.03	-0.002
Ind. Code: 23	(0.18)	(0.18)	(0.009)
Education and Entertainment Articles Manufacturing	0.01	0.01	-0.001
Ind. Code: 24	(0.08)	(0.08)	(0.004)
Petrochemicals Manufacturing	0.01	0.01	0.003
Ind. Code: 25	(0.11)	(0.09)	(0.005)
Chemical Products Manufacturing	0.06	0.08	-0.027
Ind. Code: 26	(0.23)	(0.28)	(0.018)
Medical Goods Manufacturing	0.03	0.03	-0.001
Ind. Code: 27	(0.16)	(0.16)	(0.006)
Rubber Products Manufacturing	0.03	0.01	0.020
Ind. Code: 29	(0.18)	(0.11)	(0.022)
Plastic Products Manufacturing	0.05	0.04	0.003
Ind. Code: 30	(0.21)	(0.20)	(0.009)
Non-Metallic Mineral Products Manufacturing	0.04	0.05	-0.005
Ind. Code: 31	(0.21)	(0.22)	(0.012)
Basic Metal Processing	0.01	0.01	-0.002
Ind. Code: 32	(0.11)	(0.12)	(0.006)
Non-Ferrous Metal Processing	0.02	0.02	-0.004

Ind. Code: 33	(0.13)	(0.14)	(0.006)
Fabricated Metal Products Manufacturing	0.04	0.06	-0.018
Ind. Code: 34	(0.21)	(0.24)	(0.011)
General Purpose Machinery Manufacturing	0.07	0.10	-0.025
Ind. Code: 35	(0.26)	(0.30)	(0.014)
Special Purpose Machinery Manufacturing	0.05	0.07	-0.022*
Ind. Code: 36	(0.21)	(0.25)	(0.011)
Transport Equipment Manufacturing	0.06	0.06	-0.004
Ind. Code: 37	(0.24)	(0.24)	(0.012)
Electrical Equipment Manufacturing	0.08	0.07	0.014
Ind. Code: 39	(0.27)	(0.25)	(0.015)
Computers and Electronic Products Manufacturing	0.03	0.04	-0.010
Ind. Code: 40	(0.18)	(0.20)	(0.011)
General Instruments and Other Equipment Manufacturing	0.02	0.03	-0.013
Ind. Code: 41	(0.14)	(0.18)	(0.012)
Craftworks Manufacturing	0.02	0.01	0.011
Ind. Code: 42	(0.13)	(0.08)	(0.008)
Electricity and Heat Supply	0.02	0.01	0.009
Ind. Code: 44	(0.14)	(0.11)	(0.005)
Water Production and Supply	0.01	0.00	0.007**
Ind. Code: 46	(0.10)	(0.06)	(0.002)

Note: Columns (1)–(2) report the means and standard deviations of firm characteristics. In columns (3), we test the covariate balance between upstream and downstream firms within 5km of water monitoring stations. The difference coefficients are obtained by running OLS regressions of firm characteristics on an upstream dummy. Standard errors reported in the parentheses are clustered at the water monitoring station level. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A4. Summary Statistics of Entertainment and Travel Costs and Company's Location

Variable	Mean	Standard Deviation	Min	Max
CPES (10000 RMB)				
ETC	9.49	35.59	0.00	1,100
ASIF (1000 RMB)				
ETC	123.16	267.07	0.00	3,479
Distance to nearest water monitoring station (meters)	11,158.64	40,782.65	13.65	2,391,560
Distance to the second nearest water monitoring station (meters)	46,846.90	38,270.60	13.65	3,520,434

Table A5. Entertainment and Travel Hours for Upstream and Downstream Firms (CPES)

	Before 2003			Before 2003		
Method	(1)	(2)	(3)	(4)	(5)	(6)
Conventional	0.71	0.68	0.56	-0.12	-0.13	-0.07
	(0.62)	(0.63)	(0.79)	(0.14)	(0.16)	(0.17)
Bias-corrected	0.85	0.83	0.71	-0.16	-0.16	-0.10
	(0.62)	(0.63)	(0.79)	(0.14)	(0.16)	(0.17)
Robust	0.85	0.83	0.71	-0.16	-0.16	-0.10
	(0.71)	(0.71)	(0.81)	(0.17)	(0.18)	(0.18)
Observations	888	888	888	3,626	3,626	3,626
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	5.208	5.114	4.315	6.303	7.175	5.669

Note: Each cell represents a separate regression discontinuity regression. Data are from CPES (1996 – 2009). The dependent variable is entertainment and travel hours. The running variable is the distance between a firm and a monitoring station. A positive (negative) distance means the firm is located upstream (downstream). Negative coefficients indicate that upstream firms have lower entertainment and travel hours. The discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and the MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. Year fixed effects are included in each regression. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A6. Regression Discontinuity at the Age Cutoff

	Upstream Firms			Downstream Firms		
Method	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Polluting Industries						
Conventional	-192.36**	-189.71**	-190.09**	-46.58	-50.29	-6.97
	(63.98)	(60.31)	(62.41)	(44.64)	(44.24)	(46.07)
Bias-corrected	-286.91***	-282.92***	-301.35***	57.98	59.50	62.83
	(63.98)	(60.31)	(62.41)	(44.64)	(44.24)	(46.07)
Robust	-286.91*	-282.92*	-301.35*	57.98	59.50	62.83
	(138.20)	(134.38)	(138.55)	(66.58)	(66.68)	(65.27)
Observations	1,898	1,898	1,898	1,215	1,215	1,215
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	7.284	7.188	6.274	10.70	10.16	6.043
Panel B: Non-Polluting Industries						
Conventional	-55.45	-69.76	129.29	-77.88	-79.89	-79.35
	(55.59)	(62.84)	(83.95)	(114.22)	(111.87)	(111.09)
Bias-corrected	-41.38	-49.84	150.67	-45.25	-52.12	-51.93
	(55.59)	(62.84)	(83.95)	(114.22)	(111.87)	(111.09)
Robust	-41.38	-49.84	150.67	-45.25	-52.12	-51.93
	(96.87)	(105.24)	(155.35)	(160.96)	(164.30)	(164.08)
Observations	4,969	4,969	4,969	3,268	3,268	3,268
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	7.073	6.819	4.596	19.82	19.43	12.95

Note: Each cell represents a separate regression. Data are from ASIF. The running variable is the age of the secretaries. The cutoff point is 60 years old. Negative coefficients indicate that firms in provinces with secretaries below 60 years old have lower ETC than that for local officials above 60 years old. Discontinuities at age are estimated using methods proposed by Calonico et al. (2014) and MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A7. Regression Discontinuity Manipulation Tests

	(1)	(2)
T	0.06	0.42
P>|T|	0.95	0.67
Bandwidth Left	1,660.39	1,534.40
Bandwidth Right	1,534.40	1,534.40
Observations	900	900
Bandwidth Selectors	Each	Diff

Note: This table reports regression discontinuity manipulating tests using the local polynomial density estimators proposed by Cattaneo et al. (2020). The sample consists of ASIF polluting firms located within 5km of water monitoring stations. Two bandwidth selectors are used to test the density discontinuity. "Each" means we use two distinct bandwidths based on MSE of each density separately for upstream and downstream firms. "Diff" bandwidth selection is based on MSE of the difference of densities with one common bandwidth. Technical explanations are in Cattaneo et al. (2020).

Table A8. Parametric Regression Discontinuity Estimation for Private Enterprise

	Polluting industries			Non-Polluting industries		
	(1)	(2)	(3)	(4)	(5)	(6)
RD in Corruption	-72.372 *	-107.742 ***	-103.296 **	-18.099	4.851	-1.303
	(32.121)	(30.961)	(31.938)	(29.227)	(23.105)	(24.470)
Observations	2143	2143	2143	5717	5717	5717
Log Likelihood	-14731.7	-14729.5	-14729.2	-40465.7	-40464.3	-40462.5
Polynomial Function	Linear	Quadratic	Cubic	Linear	Quadratic	Cubic

Note: Each cell represents a separate regression discontinuity regression. The sample consists of ASIF private firms within 10km of water monitoring stations. The running variable is the distance between a firm and a monitoring station. Positive (negative) distance means firms are located upstream (downstream). Negative coefficients indicate that upstream firms have lower ETC. We report OLS estimates of the coefficient on an "upstream" dummy after controlling for polynomial functions in distance from the monitoring stations interacted with an upstream dummy. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A9. Difference-in-Differences Result

	(1)
Upstream x Polluting Industries	-67.858 **
	(22.857)
Upstream	-2.138
	(12.001)
Polluting Industries	42.143 *
	(19.033)
Log Likelihood	-22307.2
Observations	3,183

Note: The data are from ASIF firms. The sample consists of private firms within 5km of water monitoring stations. Upstream indicates whether the firms are located upstream of water monitoring stations. Polluting Industries indicate whether the firm belongs to polluting industries. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A10. Imbens and Wager (2019) Estimators

	Estimator	Confidence Interval	Maximum Bias	Sample Error
Polluting industries	-313.56	-313.56±288.06	0.16	146.97
Non-polluting industries	-12.72	-12.72±280.56	0.12	143.14

Note: Each row represents a separate regression discontinuity regression. Data are private firms from ASIF within 20km of water monitoring stations. The running variable is the distance between a firm and a monitoring station. A positive (negative) distance means the firm is located upstream (downstream). Negative coefficients suggest that firms located upstream exhibit lower levels of ETC compared to downstream firms. The discontinuities at monitoring stations are estimated using methods proposed by Imbens and Wager (2019). Reported are bias-adjusted 95% confidence intervals, a bound on the maximum bias, and an estimate of the sampling error.

Table A11. Alternative Radius Circles Around the Monitoring Stations

	Polluting Industries			Non-Polluting Industries		
Method	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Within 20-km Radius						
Robust	-92.04*	-99.28*	-115.30**	-19.77	-18.91	-29.27
	(41.15)	(39.47)	(38.02)	(26.01)	(26.77)	(28.97)
Observations	6,397	6,397	6,397	16,117	16,117	16,117
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	8528	8461	7548	11034	10600	9313
Panel B: Within 30-km Radius						
Robust	-95.94**	-102.51**	-122.45***	-18.67	-19.54	-5.30
	(36.36)	(36.65)	(36.64)	(31.29)	(32.12)	(29.94)
Observations	10,978	10,978	10,978	26,074	26,074	26,074
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	9164	8749	8581	12001	11203	14224

Note: Each cell represents a separate regression. Data are private firms from ASIF. The running variable is the distance between a firm and a monitoring station. Positive (negative) distance means firms are located upstream (downstream). Negative coefficients indicate that upstream firms have lower ETC. Discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A12. Alternative Bandwidths for Private Enterprise in Polluting Industries

Bandwidth Selection Method	(1)	(2)	(3)
MSE-Two	-96.65**	-96.07**	-91.62**
	(41.97)	(40.75)	(37.76)
MSE-Sum	-108.68**	-117.21***	-119.85***
	(44.45)	(42.96)	(38.62)
CER-RD	-101.01**	-107.19***	-106.72**
	(40.86)	(39.17)	(46.38)
CER-Two	-73.55	-73.15*	-84.06**
	(46.77)	(44.40)	(40.34)
CER-Sum	-85.08*	-91.12*	-126.63***
	(51.61)	(50.10)	(38.74)
Kernel	Triangular	Epanechnikov	Uniform
Observations	3,502	3,502	3,502

Note: Each cell represents a separate regression discontinuity regression. Data are from ASIF. The running variable is the distance between a firm and a monitoring station. Positive (negative) distance means firms are located upstream (downstream). Negative coefficients indicate that upstream firms have lower ETC. Discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) for different kernel weighting methods. We use alternative bandwidth selectors proposed by Calonico et al. (2014). Technical details are in Calonico et al. (2018). Robust estimates are reported. Standard errors clustered at the station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table A13. Placebo Test for Private Enterprises in Polluting Industries

	Move Monitoring Stations Downstream			Move Monitoring Stations Upstream		
Method	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: Move by 2km			
Robust	-47.51	-51.99	-57.74	5.49	-2.63	-36.15
	(32.73)	(32.96)	(33.35)	(43.94)	(42.56)	(38.67)
Observations	3,502	3,502	3,502	3,502	3,502	3,502
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	10503	9753	8371	12477	12243	13845
			Panel B: Move by 3km			
Robust	-26.92	-26.38	-29.27	12.57	12.87	-7.22
	(30.00)	(30.15)	(31.05)	(40.92)	(39.53)	(36.47)
Observations	3,502	3,502	3,502	3,502	3,502	3,502
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	10318	9427	8344	13053	12748	12929
			Panel C: Move by 4km			
Robust	28.32	24.98	16.56	34.58	43.22	31.20
	(32.81)	(33.07)	(32.88)	(33.10)	(32.71)	(30.84)
Observations	3,502	3,502	3,502	3,502	3,502	3,502
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	9582	9633	9113	11523	11025	11488

Note: Each cell represents a separate regression. Data are private enterprises in polluting industries from ASIF. The running variable is the distance between a firm and a monitoring station. Positive (negative) distance means firms are located upstream (downstream). Negative coefficients indicate that upstream firms have lower ETC. Discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A14. Inclusion of Covariates for Private Enterprises

	Polluting Industries			Non-Polluting Industries		
Method	(1)	(2)	(3)	(4)	(5)	(6)
Conventional	-88.81**	-88.93**	-93.48**	-15.62	-3.22	-25.86
	(30.11)	(29.50)	(31.84)	(24.23)	(25.54)	(25.58)
Bias-corrected	-102.64***	-103.11***	-108.46***	-22.12	-3.22	-33.00
	(30.11)	(29.50)	(31.84)	(24.23)	(25.54)	(25.58)
Robust	-102.64**	-103.11**	-108.46**	-22.12	-3.22	-33.00
	(35.33)	(34.42)	(37.07)	(27.52)	(28.66)	(29.18)
Observations	3,499	3,499	3,499	9,276	9,276	9,276
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	9065	8924	7188	12946	10286	10957

Note: Each cell represents a separate regression discontinuity regression. Data are from ASIF. The running variable is the distance between a firm and a monitoring station. Positive (negative) distance means firms are located upstream (downstream). The negative coefficients indicate that upstream firms have lower ETC. Discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Covariates include sales, value added tax, log(# of employees), log(1+firm age) and log(per capita GDP). Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A15. Heterogeneity Effect

Method	(1)	(2)	(3)
		Panel A: Social Economy	
Conventional	-26.06	-17.12	3.60
	(72.21)	(69.59)	(66.53)
Bandwidth	11,837	11,480	10,533
		Panel B: Political Structure	
Conventional	-137.38*	-138.53*	-143.97**
	(73.86)	(72.77)	(59.15)
Bandwidth	13,290	13,387	15,878
		Panel C: Centralized vs Less Centralized	
Conventional	-96.94	-85.67	0.25
	(76.22)	(72.52)	(43.95)
Bandwidth	15,811	15,401	18,933
		Panel D: Corruption	
Conventional	-169.92**	-161.65**	-165.11**
	(73.02)	(73.22)	(72.31)
Bandwidth	11,980	11,014	9,996
		Panel E: Water Quality	
Conventional	-62.15	-53.54	-36.30
	(70.35)	(67.14)	(63.30)
Bandwidth	12285	11921	11174
Observations	3,113	3,113	3,113
Kernel	Triangular	Epanechnikov	Uniform
Bandwidth	3.686	3.987	2.576

Note: Each cell represents a separate difference-in-discontinuities estimate: the differences in corruption discontinuity between high GDP regions and low GDP regions (Panel A), the differences in corruption discontinuity between politically motivated leaders and non-politically motivated leaders (Panel B), the differences in corruption discontinuity between centralized regions and less centralized regions (Panel C), the differences in corruption discontinuity between high corruption regions and low corruption regions (Panel D), and the differences in corruption discontinuity between high water pollution regions and low water pollution regions (Panel E). We define regions with GDP higher than the median GDP as high GDP regions, city officials greater than 60 years old as politically motivated leaders, regions with distance to Capital Beijing less than the median distance as centralized regions (Huang et al. 2017), regions with corruption costs greater than median corruption costs as high corruption region, and regions with COD levels higher than median COD levels as high water pollution regions. Data are from CPES. The running variable is the distance between the county center of a firm and a monitoring station. Positive (negative) distance means firms are located upstream (downstream). Discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Year-fixed effects are included in the estimation. Standard errors clustered are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table A16. Difference-in-Discontinuities Estimates Excluding Boundary Stations

Method	(1)	(2)	(3)
Conventional	-10.96**	-10.73**	-10.93*
	(5.10)	(5.01)	(5.96)
Bias-corrected	-13.09**	-13.30***	-12.98**
	(5.10)	(5.01)	(5.96)
Robust	-13.09**	-13.30**	-12.98*
	(6.52)	(6.21)	(7.18)
Observations	5,853	5,853	5,853
Kernel	Triangular	Epanechnikov	Uniform
Bandwidth	3.830	3.808	3.471

Note: Each cell represents a separate difference-in-discontinuities estimate: the differences between corruption discontinuity before and after 2003. Data are from CPES after removing water monitoring stations located at the boundary. The running variable is the distance between the county center of a firm and a monitoring station. Positive (negative) distance means firms are located upstream (downstream). Discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Year-fixed effects are included in the estimation. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table A17. Analysis Including Ambiguous Firms

	Polluting Industries			Non-Polluting Industries		
Method	(1)	(2)	(3)	(4)	(5)	(6)
Conventional	-73.76*	-76.92*	-88.07**	-16.54	-17.22	-15.40
	(34.14)	(34.28)	(33.68)	(20.40)	(21.16)	(23.80)
Bias-corrected	-85.48*	-89.98**	-100.83**	-22.30	-23.77	-24.55
	(34.14)	(34.28)	(33.68)	(20.40)	(21.16)	(23.80)
Robust	-85.48*	-89.98*	-100.83*	-22.30	-23.77	-24.55
	(39.94)	(39.90)	(39.95)	(23.21)	(23.82)	(25.50)
Observations	4,320	4,320	4,320	12,221	12,221	12,221
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	9434	9011	7644	10731	9934	8743

Note: Each cell represents a separate regression discontinuity regression. Data are private firms from ASIF. The running variable is the distance between a firm and a monitoring station. A positive (negative) distance means the firm is located upstream (downstream). The negative coefficients indicate that upstream firms have lower ETC. The discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and the MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A18. Analysis Including Firms with Missing ETC

	Polluting Industries			Non-Polluting Industries		
Method	(1)	(2)	(3)	(4)	(5)	(6)
Conventional	-92.51**	-95.51**	-107.41**	-14.61	-15.60	-20.68
	(33.56)	(32.80)	(35.12)	(23.07)	(23.94)	(27.48)
Bias-corrected	-105.63**	-109.01***	-125.40***	-21.42	-23.84	-28.29
	(33.56)	(32.80)	(35.12)	(23.07)	(23.94)	(27.48)
Robust	-105.63**	-109.01**	-125.40**	-21.42	-23.84	-28.29
	(36.83)	(35.90)	(38.73)	(25.87)	(26.61)	(30.63)
Observations	3,509	3,509	3,509	9,314	9,314	9,314
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	11246	10994	7950	11827	11891	9385

Note: Each cell represents a separate regression discontinuity regression. Data are private firms from ASIF. The running variable is the distance between a firm and a monitoring station. A positive (negative) distance means the firm is located upstream (downstream). The negative coefficients indicate that upstream firms have lower ETC. The discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and the MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 5% ** significant at 1% *** significant at 0.1%.

Table A19. The Upstream-Downstream Corruption Gap for Private Enterprises Using Different Confidence Level Cutoffs

	Polluting Industries			Non-Polluting Industries		
Method	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Confidence level greater than 20 (error within 10km)						
Conventional	-67.06*	-73.22*	-89.03**	-28.24	-27.86	-30.27
	(30.76)	(31.73)	(31.41)	(20.96)	(21.69)	(23.05)
Bias-corrected	-75.98*	-84.06**	-99.90**	-35.07	-35.02	-38.18
	(30.76)	(31.73)	(31.41)	(20.96)	(21.69)	(23.05)
Robust	-75.98*	-84.06*	-99.90**	-35.07	-35.02	-38.18
	(35.76)	(36.20)	(33.13)	(23.18)	(23.92)	(24.78)
Observations	4,351	4,351	4,351	11,012	11,012	11,012
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	8948	8480	7686	11023	10354	9663
Panel B: Confidence level greater than 50 (error within 1km)						
Conventional	-95.84*	-103.57**	-115.07**	-20.98	-20.87	-17.83
	(40.31)	(37.86)	(36.39)	(25.97)	(26.27)	(30.17)
Bias-corrected	-111.41**	-118.44**	-130.51***	-24.86	-26.71	-25.71
	(40.31)	(37.86)	(36.39)	(25.97)	(26.27)	(30.17)
Robust	-111.41*	-118.44**	-130.51**	-24.86	-26.71	-25.71
	(46.50)	(41.90)	(39.72)	(30.26)	(30.28)	(33.42)
Observations	3,178	3,178	3,178	8,326	8,326	8,326
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	8661	9229	9482	10224	10024	8712

Note: Each cell represents a separate regression discontinuity regression. Data are from ASIF. Confidence level is the output parameter from Baidu Map API, which indicates the error between street address and output coordinates. The running variable is the distance between a firm and a monitoring station. Positive (negative) distance means firms are located upstream (downstream). Negative coefficients indicate that upstream firms have lower ETC. Discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table A20. Corruption Gap for Private Enterprises After Removing Unbalanced Industries

	Polluting Industries			Non-Polluting Industries		
Method	(1)	(2)	(3)	(4)	(5)	(6)
Conventional	-93.11**	-95.93**	-107.34**	-14.75	-15.22	-7.37
	(33.75)	(32.83)	(35.41)	(23.69)	(24.68)	(27.69)
Bias-corrected	-106.41**	-109.50***	-124.71***	-19.10	-20.36	-14.69
	(33.75)	(32.83)	(35.41)	(23.69)	(24.68)	(27.69)
Robust	-106.41**	-109.50**	-124.71**	-19.10	-20.36	-14.69
	(37.09)	(35.91)	(39.16)	(26.72)	(27.65)	(30.13)
Observations	3,502	3,502	3,502	8,449	8,449	8,449
Kernel	Triangular	Epanechnikov	Uniform	Triangular	Epanechnikov	Uniform
Bandwidth	11100	10970	7834	9987	9499	8174

Note: Each cell represents a separate regression. Data are from ASIF after removing two unbalanced industries in Table A2. The running variable is the distance between a firm and a monitoring station. Positive (negative) distance means firms are located upstream (downstream). Negative coefficients indicate that upstream firms have lower ETC. Discontinuities at monitoring stations are estimated using methods proposed by Calonico et al. (2014) and MSE optimal bandwidth proposed by Calonico et al. (2014) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table A21. Covariate Balance of Data Cleaning Process

	Mean		Mean Difference
	After Data Cleaning	Before Data Cleaning	
	(1)	(2)	(3)
Panel A: ASIF			
Year of Opening	1995.21	1995.06	0.150***
	(10.83)	(11.00)	(0.029)
Polluting industries	0.33	0.33	0.000
(1=Yes, 0=Others)	(0.47)	(0.47)	(0.001)
Private Enterprise	0.87	0.86	0.010***
(1=Yes, 0=Others)	(0.33)	(0.34)	(0.001)
Profit	4,106.95	4,053.04	53.912
(1,000 RMB)	(160,262.77)	(158,948.18)	(429.102)
Value-Added Tax	2,463.55	2,435.87	27.684
(1,000 RMB)	(41,047.29)	(40,781.59)	(109.998)
# of Employees	244.14	240.63	3.510
(Male)	(1,097.35)	(1,100.13)	(2.954)
# of Employees	104.75	103.30	1.446
(Female)	(389.85)	(402.35)	(1.065)
Capital Stock	21,060.69	20,799.08	261.606
(1,000 RMB)	(467,307.80)	(463,420.54)	(1,251.140)
Intermediate Input	54,942.99	54,232.17	710.825
(1,000 RMB)	(527,172.64)	(523,939.60)	(1412.949)

Note: Columns (1)–(2) report the means and standard deviations of firm characteristics. In column (3), we conduct t-test about the sample after data cleaning process and before data cleaning process. Standard errors are reported in the parentheses. *** significant at 0.1%.

References

Calonico S, Cattaneo MD, Farrell MH (2018) On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association* 113(522):767-779.

Calonico S, Cattaneo MD, Titiunik R (2014) Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82(6):2295-2326.

Cattaneo MD, Jansson M, Ma X (2020) Simple local polynomial density estimators. *Journal of the American Statistical Association* 115(531):1449-1455.

Gelman A, Imbens G (2019) Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics* 37(3):447-456.

Huang Z, Li L, Ma G, Xu LC (2017) Hayek, local information, and commanding heights: Decentralizing state-owned enterprises in China. *American Economic Review* 107(8):2455-78.

Imbens G, Wager S (2019) Optimized regression discontinuity designs. *Review of Economics and Statistics* 101(2):264-278.

Lee DS, Lemieux T (2010) Regression discontinuity designs in economics. *Journal of economic literature* 48(2):281-355.