TESINA

Carmela Pia Senatore

1. PRESENTAZIONE DEL PROBLEMA

Il problema in analisi deriva dall'introduzione di una legge che prevede un cambiamento nel sistema elettorale per l'elezione del sindaco. In particolare , questa legge messa in atto , afferma che i comuni che hanno meno di 15000 abitanti continuano a votare con il sistema del voto unico, mentre, i comuni la cui soglia di abitanti è maggiore di 15000 mila vengono sottoposti all'introduzione del doppio turno.

Lo studio posto in essere si concentrerà nel comprendere se l'introduzione del doppio turno ha portato a una riduzione o a un aumento dell'efficienza basando l'analisi su un anno 'pre-policy' (1994) e su un anno 'post-policy'(2003).

2. PRESENTAZIONE DEI DATI

Il Dataset preso in analisi, presente nel file excel myxz_runoff.xls, è composto da 22 variabili e 17328 osservazioni che rappresentano alcuni comuni italiani in vari anni (dal 1994 al 2005).

Ho iniziato ripulendo il dataset con il metodo delle Tukey's Fences ponendo in un primo momento k=1.5 per escludere tutte le possibile anomalie, successivamente ho posto k=3 in modo da escludere solo le osservazioni che risultano essere sospetti gross-outliers. Ad ogni modo, non ho ripulito le variabili categoriali, le variabily dummy e la variabile imptot, essendo quest'ultima somma di impegni di spesa corrente per personale e non per il personale, che sono state invece ripulite. La mia analisi è stata svolta nell'appendice, in cui sono presenti passo per passo le analisi univariate delle variabili e la loro rispettiva pulizia, implementando un confronto pre e post pulizia.

Ho deciso di lavorare sul dataset nuovo ripulito con k=3 , dunque eliminando solo i sospetti gross outliers, ciò in quanto ho evinto che riducendo il k a 1.5 avrei potuto tagliare fuori dall'analisi dei comuni importanti che potrebbero essere fondamentali per spiegare i fenomeni oggetto di studio.

Inoltre, è necessario sottolineare che sono stati individuati due dataset diversi per via di una problematica riscontrata durante l'analisi dovuta al fatto che, ripulendo l'intero dataset dai gross outliers sulle variabili stabilite nel file di pulizia(quintali di rifiuti, km strade, studenti, nr punti luce , nr permessi, nr iscritti, km polizia, nr certificati, imppers, impnopers, vecchiaia), non viene riscontrato nessun comune dei trattati con popolazione>15000 abitanti. Per cui ho individuato un dataset ripulito da tutte le variabili e un dataset ripulito dalla sola variabile di quintali di rifiuti raccolti per stabilire un confronto a livello di robustezza (analizzato nel 4 punto di questo documento).

3. PRESENTAZIONE DELLE TECNICHE UTILIZZATE

Per il problema che mi è stato posto ho deciso di lavorare con l'analisi non parametrica. L'approccio parametrico, essendo di tipo stocastico, prova a scindere gli effetti imputabili all'inefficienza, da quelli dovuti all'errore; l'approccio non parametrico, viceversa, non essendo di tipo stocastico, confonde i due effetti. L'analisi non parametrica tiene conto dei seguenti metodi:

FDH: ovvero Free Disposal Use, che è costruita con una sola e unica ipotesi alla base presupponendo che la frontiera di produzione di un insieme è definita come il limite dell'insieme della libera disposizione dei dati di input e di output. Questo metodo non dispone di forti ipotesi a priori sulla convessità dell'insieme degli input, come avviene nell'approccio DEA. L'insieme di riferimento, che parte da un punto osservato , ingloba i processi produttivi inefficienti avvalendosi di determinismo , ipotizzando la non presenza di rumore stocastico. La sua rappresentazione grafica è una spezzata.

DEA-V: ovvero Data Envolopement Analisys, dove per DEA intendiamo un metodo la cui principale caratteristica è quella di "inviluppare" i dati, determinare una frontiera senza basarsi su funzioni di produzione o distribuzioni dell'errore predeterminate. La DEA-V è un FDH con l'aggiunta della convessità. Infatti l'insieme di riferimento non prende tutti processi produttivi presenti ma solo quelli che sono combinazione convessa. La sua rappresentazione grafica implicherà una forma di tipo piramide convessa , avremo infatti meno gradini rispetto all'FDH dovuto alla considerazione di combinazioni convesse.

DEA-C: è un FDH con l'aggiunta di proporzionalità e additività. Per proporzionalità intendiamo che l'insieme di riferimento contiene tutti i processi produttivi che sono proporzionali a quelli ammissibili. Per additività intendiamo che l'insieme di riferimento contiene tutti i processi produttivi che sono somme dei processi produttivi ammissibili. La sua rappresentazione grafica sarà più liscia rispetto alle altre due.

Per misurare l'impatto della policy riguardante l'inserimento del doppio turno per i paesi con più di 15000 abitanti ho utilizzato i metodi DiD e SM.

Il metodo Did (Difference-In-Difference) è possibile utilizzarlo quando si hanno informazioni sia sulle unità incluse che escluse dalla politica, ma non sul processo di selezione. Vengono quindi individuate tre variabili:

a) Dopo, vale 0 se l'anno della iesima osservazione è riferente all'anno pre-policy, 0 altrimenti;
b) Doppio, vale 0 se il comune ha meno di 15000 abitanti,1 altrimenti;
c) DiD, è il prodotto tra dopo e doppio.

Il metodo dello Statistical Matching utilizzabile se si hanno informazioni relative sia alle unità escluse sia alle modalità di assegnazione del beneficio. Ogni unità trattata infatti avrà sicuramente un corrispondente gemello con le stesse caratteristiche in questione senza però essere davvero trattata.

4. PRESENTAZIONE DEI PRINCIPALI RISULTATI

Ho deciso di utilizzare l'analisi non parametrica per il problema presentato. Ho utilizzato le tre tecniche (fdh, dea-v e dea-c) per calcolare l'efficienza dei comuni nel 1994 e nel 2003 sia per il dataset ripulito da tutte le variabili che per il dataset ripulito dalla sola variabile quintali di rifiuti raccolti. Mi concentrerò sull'analisi della dea-v pre e post- policy, le altre tabelle sono state inserite nell'appendice con i rispettivi commenti.

A) DATASET RIPULITO DA TUTTE LE VARIABILI

DEA-V ORIENTATA DAL LATO DEGLI OUTPUT, ANNO==1994					
Eff range	#	%			
F==1	185	10.4			
1<F=<1.1	234	13.1			
1.1<F=<1.2	303	17.0			
1.2<F=<1.3	292	16.4			
1.3<F=<1.5	382	21.4			
1.5<F=<2	309	17.3			
2<F=<5	77	4.3			
Min	1st Qu	Median	Mean	3rd Qu	Max
1.0	1.111	1.259	1.332	1.458	3.188

Nell'anno 1994 le unità efficienti ricoprono solo 185 comuni, ovvero il 10.4 % del totale. Il restante 90% delle osservazioni (1597 comuni) da questo punto di vista è inefficiente. Il punteggio più elevato è più inefficiente. Ha un minimo in 1 e un massimo in 2.059, ovvero l'espansione degli output per arrivare alla frontiera.

Passiamo ora all'analisi dell'anno 2003, inteso come anno post-policy.

DEA-V ORIENTATA DAL LATO DEGLI OUTPUT, ANNO==2003					
Eff range	#	%			
F==1	148	8.4			
1<F=<1.1	182	10.4			
1.1<F=<1.2	200	11.4			
1.2<F=<1.3	190	10.8			
1.3<F=<1.5	359	20.5			
1.5<F=<2	478	27.3			
2<F=<5	195	11.1			
Min	1st Qu	Median	Mean	3rd Qu	Max
1.0	1.158	1. 381	1. 470	1.688	3.990

Nel 2003 i comuni efficienti sono 148 , ovvero l'8.4% del totale. Le restanti unità risultano essere inefficienti, circa il 91.6%. I punteggi hanno un minimo in 1 e un massimo in 3.99. Genericamente l'inefficienza media aumenta con il tempo.

B) DATASET RIPULITO DALLA SOLA VARIABILE QUANTITATIVI DI RIFIUTI RACCOLTI

Nella seconda analisi svolta, il quantitativo di comuni presi in considerazione sono decisamente superiori rispetto al dataset ripulito da tutte le variabili, ma soprattutto, permettono di includere le unità statistiche su cui deve essere svolta l'analisi del problema di runoff, ovvero i comuni con più di 15000 abitanti.

DEA-V ORIENTATA DAL LATO DEGLI OUTPUT, ANNO==1994					
Eff range	#	%			
F==1	185	9.4			
1<F=<1.1	242	12.3			
1.1<F=<1.2	311	15.8			
1.2<F=<1.3	333	17.0			
1.3<F=<1.5	455	23.2			
1.5<F=<2	355	18.1			
2<F=<5	82	4.2			
Min	1st Qu	Median	Mean	3rd Qu	Max
1.0	1.121	1.274	1.340	1.470	3.188

Il numero di comuni efficienti risulta essere 185, ricoprendo il 9.4% del totale. Passiamo all'anno 2003 inteso come anno post-policy.

DEA-V ORIENTATA DAL LATO DEGLI OUTPUT, ANNO==2003					
Eff range	#	%			
F==1	151	7.9			
1<F=<1.1	190	10.0			
1.1<F=<1.2	216	11.3			
1.2<F=<1.3	203	10.6			
1.3<F=<1.5	413	21.7			
1.5<F=<2	529	27.7			
2<F=<5	205	10.7			
Min	1st Qu	Median	Mean	3rd Qu	Max
1.0	1.163	1.384	1. 470	1.681	3.990

Nel 2003 i comuni efficienti sono 151 , ovvero l'8.4 per cento del totale.

Per l'anno pre-policy: per la fdh il numero di comuni efficienti è maggiore nel caso del dataset ripulito da una sola variabile rispetto al dataset ripulito dai gross-outliers. Per la dea-v il numero di comuni efficienti è identico, mentre per la dea-c il numero di comuni efficienti è maggiore per il dataset ripulito dai gross outliers che per il dataset ripulito da una sola variabile.

Per l'anno post-policy: per la fdh e dea-v, il numero di comuni efficienti è maggiore nel caso del dataset ripulito da una sola variabile rispetto al dataset ripulito dai gross-outliers. Mentre per la dea-c il numero di comuni efficienti è quasi simile, si differenziano per un solo comune.

Siccome ci concentriamo sul secondo dataset , andiamo a indagare meglio individuando una matrice contenente i valori di efficienza dell'anno 1994 e 2003, usandola come variabile dipendente nell'analisi DiD. Creo quindi le variabili dopo, doppio e did (come spiegato precedentemente) individuando la seguente stima del modello di regressione:

	Estimate	Std. Error	T value	Pr(>|t|)	
(Intercept)	1.344654	0.008115	165.702	< 2e-16	***
dopo	0.132968	0.011549	11.514	<2e-16	***
doppio	-0.143614	0.047209	-3.042	0.00237	**
did	-0.169525	0.069688	-2.433	0.01503	*

Considerando un alpha pari a 0.05, tutti i p-value associati alle variabili sono prossimi allo zero per l'intercetta e la dummy dopo, leggermente più elevato per doppio e did, ma comunque tutti statisticamente significativi, per cui

significa che hanno un impatto sul valore dell'efficienza. Il concetto da fissare è che un valore dello stimatore negativo implica un miglioramento dell'efficienza. Per cui a parità di condizioni, il valore positivo per dopo implica che per l anno dopo il 1994 l'efficienza è diminuita. Al contrario, un valore negativo per il coefficiente doppio implica che per i paesi in cui è stato introdotto il doppio turno , quindi i comuni con più di 15000 abitanti, l'efficienza è aumentata rispetto ai comuni in cui vige il singolo turno. In particolare, la variabile che ci interessa più di tutte è did, che è negativa, per cui in tutti i comuni effettivamente trattati dopo il 1994 l'efficienza è aumentata. Effettuo quindi l'analisi di did aumentata:

	Estimate	Std. Error	T value	Pr(>|t|)	
(Intercept)	1.422e+00	6.687e-02	21.260	< 2e-16	***
vecchiaia	8.217e-01	1.566e-01	5.246	1.63e-07	***
laurea	-1.160e+00	2.992e-01	-3.875	0.000108	***
Superficie	-2.029e-03	1.444e-04	-14.049	< 2e-16	***
Altitudine	1.055e-04	2.369e-05	4.453	8.70e-06	***
Litoraneità	7.266e-02	2.636e-02	2.757	0.005864	***
Gradourb2	-1.471e-01	1.566e-02	-9.394	< 2e-16	***
Gradourb3	-1.842e-01	2.556e-02	-7.209	6.76e-13	***
ruralità	7.704e-03	1.560e-02	0.494	0.621450	
dopo	1.202e-01	1.068e-02	11.249	< 2e-16	***
doppio	7.705e-02	4.551e-02	1.693	0.090535	.
did	-2.159e-01	6.417e-02	-3.364	0.000776	***

Vediamo che did continua a rimanere significativo e negativo a un livello di alpha=0.05.Conferma la presenza di un miglioramento dell'efficienza in maniera generale. L'unico dei coefficienti che non è per nulla significativo è ruralità, mentre se fissiamo un livello di significatività 0.10 anche dopo rientra nelle variabili significative. Da un analisi generale di questo modello riusciamo a dedurre che: 1) i coefficienti di vecchiaia, altitudine, litoraneirà, grado di urbanizzazione basso, ruralità=1, dopo e doppio sono positivi il che significa , che nel quadro generale, sono legate a inefficienza.2) Mentre i coefficienti associati a laurea, superficie, grado di urbanizzazione alta e did sono negativi, per cui sono legati all'efficienza. Ho quindi arricchito le informazioni con i dati di panel:

	Estimate	Std. Error	T value	Pr(>|t|)	
vecchiaia	2.9267e-01	1.7281e-01	1.6937	0.09041	.
laurea	1.0727e+00	3.3096e-01	3.2412	0.00120	**
Superficie	6.2502e-04	1.5478e-04	4.0382	5.492e-05	***
Altitudine	1.1300e-05	2.6111e-05	0.4328	0.66520	
Litoraneità	2.3747e-02	2.8693e-02	0.8276	0.40793	
Gradourb2	1.7377e-02	1.7250e-02	1.0074	0.31383	
Gradourb3	4.1028e-02	2.7962e-02	1.4673	0.14239	
Ruralità	9.6246e-04	1.7128e-02	0.0562	0.95519	
did	-2.5619e-01	5.1971e-02	-4.9294	8.595e-07	***

Did continua a essere statisticamente significativo e negativo.

1994	2003
1.1614	1.1716

Impostando gli effetti fissi sul tempo ci rendiamo conto che entra in contrasto con ciò che abbiamo appena detto, per cui sembra che ci sia stato un lieve peggioramento dell' efficienza dopo l'applicazione del doppio turno.

Successivamente sono passata allo statistical matching utilizzando la tecnica del nearest neighbor prima per il 1994 e poi per il 2003, ulteriori analisi sul matching sono presenti nell'appendice e sullo script in R, qui riporterò solo i dati essenziali:

STATISTICAL MATCHING DOPPIO SUL 1994					
	Estimate	Std. Error	T value	Pr(>|t|)	
(Intercept)	1.34686	0.00719	187.3	<2e-16	***
Doppio_1994	-0.18092	0.04761	-3.8	0.000149	***

STATISTICAL MATCHING DOPPIO SUL 2003

	Estimate	Std. Error	T value	Pr(>|t|)	
(Intercept)	1.479796	0.009258	159.839	<2e-16	***
Doppio_2003	-0.318905	0.057422	-5.554	3.19e-08	***

Doppio continua a essere statisticamente significativo e negativo, quindi implica che i comuni trattati sono più efficienti, sia prima che dopo la policy.

Svolgendo il test T sulle differenze per vedere se c' è una differenza significativa tra l'intercetta di doppio prima e dopo il 94 , ha dato come valore '11.04' , avendo come differenza tra anno pre e post policy un aumento di efficienza di 0.13. Quindi, questo significa che la legge introdotta ha avuto un impatto a livello di miglioramento dell'efficienza significativa al 0.01 per cento. Mi sento di confermare questa intuizione sulla base dello statistical matching che ha fatto un buon abbinamento visto che: a) Il numero di unità di controlli è decisamente maggiore rispetto ai trattati per cui l'implementazione è attendibile; b) Dopo il bilanciamento per i dati matchati riusciamo a visualizzare che i dati sono pressoché simili il che conferma un buon abbinamento.

Infine, sono passata a fare un controllo della robustezza dei risultati con l'analisi difference-in-differences con i comandi "lm" e "plm" usando ImpTot come variabile dipendente. In un primo momento con il metodo lm ho individuato i seguenti risultati:

	Estimate	Std. Error	T value	Pr(>|t|)	
Rifiuti	-0.01797	0.045292	-8.032	1.03e-15	***
Km Strade	0.038958	0.004681	8.322	< 2e-16	***
Studenti	-0.195941	0.008981	-21.818	<2e-16	***
NrPuntiLuce	0.116196	0.010005	11.613	<2e-16	***
NrPermessi	-0.070480	0.006256	-11.266	<2e-16	***
NrIscritti	1.089199	0.016025	67.967	<2e-16	***
KmPolizia	-0.013058	0.008784	-1.487	0.1371	
Nrcertificati	-0.011528	0.006164	-1.870	0.0615	.
Dopo	0.169052	0.005624	30.057	< 2e-16	***
Doppio	0.322106	0.031083	10.363	< 2e-16	***
Did	-0.055114	0.032681	-1.686	0.0917	.

Dal modello di regressione visualizziamo che tutte e tre le variabili sono significative a un livello alpha pari a 0.10, did è negativa mentre dopo e doppio sono positive. Il che significa che la spesa per i comuni effettivamente trattati è diminuita leggermente. Persino eliminando l'intercetta doppio risulta essere statisticamente significativa a livello alpha 0.05 per cento.

Successivamente con 'plm', ho quindi individuato il modello di regressione, ma in questo caso Did è non significativa con i dati panel , quindi non possiamo dare una spiegazione generale, è come se la politica non avesse avuto impatto.

	Estimate	Std. Error	T value	Pr(>|t|)	
Rifiuti	-0.0267	0.002	-12.49	< 2e-16	***
Km Strade	0.03289	0.0044	7.41	< 2e-16	***
Studenti	-0.2115	0.0056	-24.6934	<2e-16	***
NrPuntiLuce	0.0912558	0.0096247	9.4815	<2e-16	***
NrPermessi	-0.1016734	0.0058803	-17.2904	<2e-16	***
NrIscritti	1.1620651	0.0150792	77.0642	<2e-16	***
KmPolizia	-0.0158415	0.0084186	-1.8817	0.05989	.
Nrcertificati	0.0109598	0.0059435	1.8440	0.0652	.
Did	-0.0016813	0.0117798	-0.1427	0.88651	

5. CONSIDERAZIONI CONCLUSIVE

Per quanto riguarda le considerazioni conclusive–, nonostante gli ultimi problemi riscontrati nel controllo della robustezza con i dati panel in cui la variabile did risulta essere non statisticamente significativa, tenendo conto dei risultati precedentemente analizzati :

1) Risultati relativi modello di regressione costruito con i valori delle efficienze e le variabili-: dopo, doppio e did,
2) I risultati derivanti da un ottimo matching nello statistical matching

Posso concludere, secondo il mio parere, che l’inserimento del doppio turno per i comuni con più di 15000 abitanti ha ridotto l’inefficienza e gli impegni di spesa corrente.