Il metodo delle proporzioni crescenti

Ulisse Di Corpo

Abstract

I modelli predittivi, che utilizzano le regressioni multiple, partono dall'assunto che le relazioni tra le variabili sono lineari o traducibili nella forma lineare. La pratica dimostra che questo assunto è in genere errato e che le relazioni seguono andamenti non lineari che non giustificano l'utilizzo delle regressioni multiple. In queste pagine viene proposta una tecnica alternativa che consente di individuare relazioni non lineari altamente predittive.

Premessa

Quando si vogliono stimare i valori delle micro-zone territoriali (ad esempio i comuni) per un indicatore noto solo a livello di macro-zone (ad esempio le regioni) si utilizza, in genere, il metodo delle regressioni multiple. Ma, quando poi si verificano i risultati, sommando i valori quantitativi così ottenuti, si osservano quasi sempre risultati estremamente diversi da quelli iniziali e si scopre che il modello delle regressioni multiple non consente di giungere ad una stima attendibile.

Questo perché:

- la natura delle variabili è complessa e gli andamenti delle correlazioni in genere non sono lineari: ad esempio l'andamento delle correlazioni nel Sud Italia è spesso diverso

dall'andamento delle correlazioni nel Nord Italia. L'indice di correlazione, dal quale deriva il metodo delle regressioni multiple, presuppone invece correlazioni lineari o riconducibili alla forma lineare.

- Spesso esistono delle variabili, banali, che correlano con tutte. Ad esempio in Italia tutti gli indicatori correlano con il Nord e Sud Italia. Le forti correlazioni che si osservano tra gli indicatori sono quindi semplicemente dovute al fatto che esiste una terza variabile (la variabile Nord/Sud) che correla in modo forte con tutte.
- utilizzando poche macro-unità (ad esempio le 20 regioni italiane) il calcolo della significatività statistica (probabilità di errore) è molto approssimativo e non si riesce a distinguere tra correlazioni in grado di effettuare la stima e variabili correlate in modo spurio (tramite una terza variabile, ad esempio la variabile Nord/Sud).

Il metodo delle proporzioni crescenti

Si è deciso quindi di testare un metodo in grado di individuare relazioni lineari e non lineari, la cui probabilità (rischio di errore) sia estremamente bassa e il cui potere predittivo è perciò estremamente elevato.

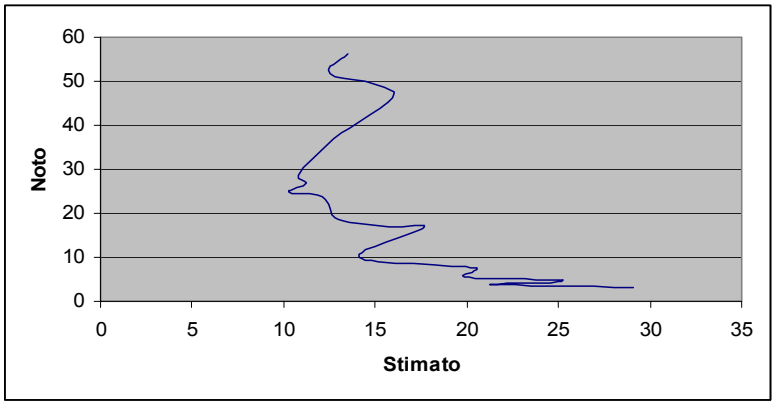
Il criterio è il seguente:

1. Ogni indicatore viene ordinato per valori crescenti
2. Vengono calcolate le proporzioni tra l'indicatore da stimare e l'indicatore ordinato per valori crescenti.
3. Vengono selezionati solo quegli indicatori che presentano sempre proporzioni crescenti.

Come si vede nell'esempio della tabella riportata di seguito, al crescere dell'indicatore noto cresce anche la proporzione del rapporto tra l'indicatore da stimare e l'indicatore noto.

Tab. 1 Esempio di proporzioni crescenti

Indicatore da stimare	Indicatore noto	Proporzione
29.1	3.2209347572312	0.1106850432038
21.4	3.7472576341536	0.1751054969231
24.0	4.2306504791205	0.1762771032967
25.2	4.6657799308346	0.1851499972553
20.8	5.2952271608988	0.2545782288894
19.8	5.9854907244274	0.3022975113347
20.5	7.6421077174166	0.3727857423130
14.1	10.1008595988539	0.7163730211953
17.7	16.9204114240064	0.9559554476840
15.8	16.9357366771160	1.0718820681719
13.0	18.7267234701782	1.4405171900137
12.3	23.0345854828162	1.8727305270582
11.4	24.3379344343111	2.1349065293255
10.3	24.7761336866902	2.4054498724942
11.2	26.9976775283271	2.4105069221721
10.8	28.7036518210801	2.6577455389889
12.7	36.7435590173757	2.8931936234154
16.0	47.1063829787234	2.9441489361702
12.5	51.7574171029668	4.1405933682373
13.5	56.3550834597876	4.1744506266509



La probabilità che questa proporzione cresca o diminuisca è, al primo passaggio, del 50% (½) al secondo passaggio del 25% (¼). Quindi, la probabilità di osservare due proporzioni crescenti è pari a ½ x ¼ ($1/2^{2!}$). La probabilità di osservare 19 proporzioni crescenti (come nella tabella accanto) è pari a $1/2^{19!}$, cioè 1 possibilità ogni 12.164.510 miliardi, praticamente nulla.

Riportando su di un grafico i valori della tabella (indicatore da stimare e indicatore noto), si osserva che la relazione non è lineare: al crescere dell'indicatore noto non cresce o decresce egualmente l'indicatore da stimare. La relazione corrisponde perciò ad una funzione di cui sono noti solo alcuni punti. Punti che possono quindi essere utilizzati per stimare i valori dell'indicatore partendo dall'indicatore noto.

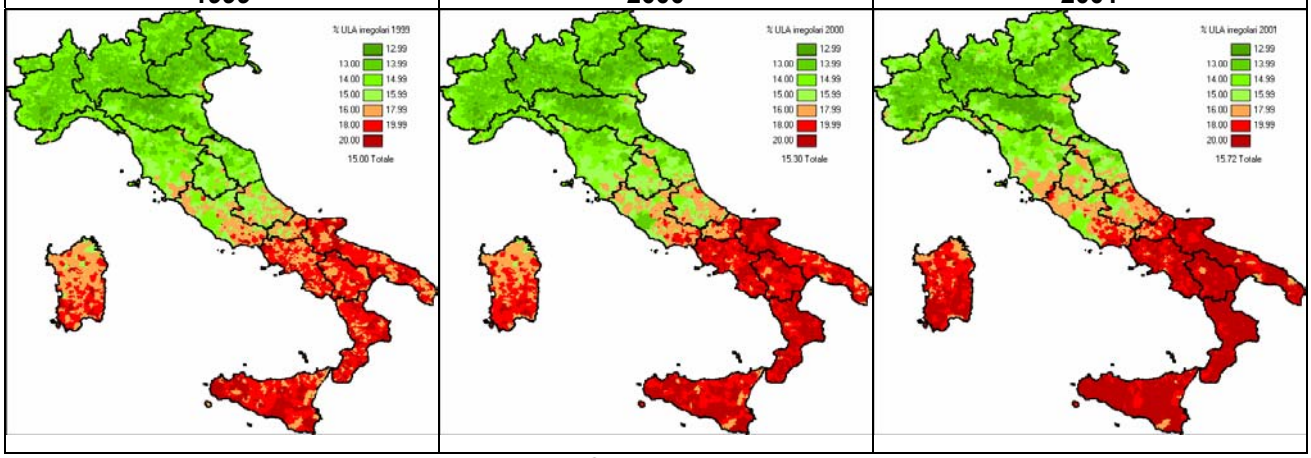
Una applicazione

Il metodo delle proporzioni crescenti è stato utilizzato per stimare a livello comunale il numero di ore di lavoro irregolare (ULA) partendo dall'indicatore regionale fornito dall'ISTAT. Utilizzando il metodo delle proporzioni crescenti sono stati individuati, tra gli indicatori prodotti dalle variabili dei censimenti, 13 indicatori che ottengono tutte e 19 le proporzioni crescenti. In questo esempio la stima dei valori intermedi tra questi punti è stata effettuata utilizzando la retta che li congiunge. Per i valori, invece, inferiori o superiori ai valori limite della distribuzione (prima e ultima riga), si è continuato ad utilizzare il valore di proporzione del punto estremo. I valori così stimati dell'indicatore hanno poi consentito di calcolare i valori quantitativi e in questo modo, sommando i valori dei comuni, calcolare i valori regionali dell'indicatore stimato; confrontando così la stima regionale con il valore fornito dall'ISTAT si osserva una differenza del 14,54% per il 1999, del 13,00% per il 2000 e del 12,54% per il 2001.

Al fine di ridurre la discrepanza e nell'ipotesi di una maggiore attendibilità dei dati stimati dall'ISTAT, i valori comunali sono stati corretti (riproporzionati) in modo da far coincidere le somme regionali con il valore fornito dall'ISTAT. Il risultato così ottenuto è il seguente.

Stima su base comunale del lavoro irregolare

1999	2000	2001



Anlisi realizzate utilizzando sintropia-AS http://www.sintropia.it/ricerca/landstat/landstat.htm

Stima di una variabile nota (Popolazione Attiva al 1991)

Per verificare il metodo appena descritto si è deciso di utilizzare una distribuzione nota, per la quale oltre ai dati regionali sono noti anche i dati comunali. A tal fine è stata selezionata la popolazione attiva al censimento del 1991.

	Stima	Dato reale	Discrepanza
Piemonte	1.810.256	1.915.651	- 5.50%
Valle d'Aosta	51.118	52.712	- 3.02%
Lombardia	3.879.296	4.020.360	- 3.51%
Trentino Alto Adige	397.782	392.729	- 1.29%
Veneto	1.927.250	1.936.915	- 0.50%
Friuli Venezia Giulia	505.731	509.894	- 0.82%
Liguria	673.259	673.315	- 0.01%
Emilia Romagna	1.707.527	1.814.770	- 5.91%
Toscana	1.488.400	1.543.354	- 3.56%
Umbria	347.504	336.412	+ 3.30%
Marche	620.276	626.172	- 0.94%
Lazio	2.183.445	2.168.728	+ 0.68%
Abruzzo	525.869	502.429	+ 4.67%
Molise	132.778	132.390	+ 0.29%
Campania	2.353.818	2.197.869	+ 7.10%
Puglia	1.648.401	1.562.468	+ 5.50%
Basilicata	245.075	245.622	- 0.22%
Calabria	823.986	800.200	+ 2.97%
Sicilia	2.041.341	1.829.059	+11.61%
Sardegna	675.876	663.589	+ 1.85%

Le discrepanze osservate possono essere attribuite alla compressione dei valori estremi per i quali si utilizza l'ultimo punto della funzione. Ciò è suggerito dal fatto che le discrepanze maggiori si osservano (di segno positivo) in Sicilia e Campania e (di segno negativo) in Emilia Romagna e Piemonte. E' da notare che il valore reale viene calcolato sempre come conseguenza di un indicatore, in questo caso l'indicatore "*popolazione attiva/popolazione residente*"; laddove i valori dell'indicatore sono più bassi – estremo inferiore – il valore reale viene sovrastimato, mentre laddove i valori dell'indicatore sono più alti – estremo superiore – il valore reale viene sottostimato.

Confronto di comuni con valori estremi: l'esempio della Liguria

Si è scelta la Liguria in quanto è la regione nella quale si osserva la percentuale di discrepanza più bassa tra i valori stimati e i valori forniti dall'ISTAT. Tabulando i valori dei comuni si osserva la tendenza di questa metodologia a sottostimare i valori più elevati e a sovrastimare i valori più bassi:

	Dato Reale	Dato Stimato
Genova	270.602	262.771
La Spezia	40.576	39.346
Savona	26.973	25.710
San Remo	24.221	21.440
Propata	51	65
Fascia	48	51
Montegrosso Pian Latte	35	59
Rondanina	34	40

Questo errore sistematico risulta secondario nel momento in cui i dati grezzi vengono utilizzati per il calcolo di indicatori in quanto, in genere, porta alla sovrastima o sottostima sia del numeratore che del denominatore, giungendo in questo modo a risultati che si discostano poco da quelli reali.

Conclusioni

Il pregio della metodologia presentata in queste pagine è indubbiamente quello di permettere di individuare relazioni non lineari altamente predittive. In questo modo è, ad esempio, possibile stimare i microdati partendo da dati aggregati a livello di macroaree.

E' importante, però, ricordare che le stime non vanno confuse con i valori reali. Si tratta sempre di "suggerimenti" e "indicazioni" utili per indirizzare una prima attività di approfondimento, ma che non possono sostituirsi ad attività più mirate di studio e di indagine.