

A faint, light-colored map of Italy serves as the background for the slide. It shows major cities, regions, and geographical features like the Gulf of Venice and Sicily. The map is oriented vertically, with the top of the page corresponding to the north of Italy.

**La regionalizzazione della spesa pubblica:
migliorare la qualità e la tempestività delle informazioni**

Roma, 16 ottobre 2003

Discussant su Metodi e tecniche

Renato Coppi
Università di Roma “La Sapienza”

- Intervento

Intervento

Essendo un metodologo di professione - insegno scienze statistiche ed analisi multivariata - il mio interesse specifico è stato stimolato dalle metodologie utilizzate dal gruppo di lavoro che ha prodotto l'Indicatore Anticipatore.

Mi ritrovo su un doppio binario con le impostazioni scelte dal gruppo e illustrate da Francisco Barbaro e Roberto Di Manno. Su un piano più generale di tipo concettuale ritengo che tutte le impostazioni descritte rientrino in una branca contemporanea di grande sviluppo delle metodologie statistiche: il *data mining*, cioè lo sfruttamento di giacimenti informativi disponibili anche se rilevati per motivi soprattutto gestionali. Questo porta allo sfruttamento a fini conoscitivi ed operativo-decisionali di dati raccolti ad altri scopi ed è collegato ad una serie di problematiche che sorgono nell'uso di questi giacimenti informativi. È necessario manipolare i dati, effettuare vere e proprie fusioni di dati, fare un *link* tra diverse fonti informative. Il *data mining* ha dunque l'esigenza di gestire dati empirici e contemporaneamente la necessità di utilizzare l'informazione teorica che è fatta di assunzioni, di modelli, che vengono assunti al fine di estrarre quell'informazione conoscitiva ed operativo-decisionale che ci interessa. Sono stati illustrati esempi in questo senso sia per quanto riguarda il modello regressivo utilizzato per la stima delle Amministrazioni Locali sia per tutti i modelli logistici usati per la stima del comparto statale. Di Manno ha offerto alcune prospettive metodologiche che condivido pienamente: l'uso delle reti neurali, l'uso di matrici di distanza o di prossimità, da analizzare successivamente.

Condivido, in generale, l'uso di strategie di analisi combinate. Diverse tecniche di analisi di dati che, nel loro insieme, riproducono un'informazione più ricca e più attendibile che non quella ottenibile utilizzando di singole tecniche di analisi.

Un aspetto che andrebbe considerato approfonditamente, anche in riferimento al lavoro presentato in questa sede, è quello dell'incertezza legata alle elaborazioni. Vi è infatti un doppio aspetto da tenere presente: da un lato l'informazione dall'altro l'incertezza connessa all'informazione.

Da questo punto di vista alcune verifiche sono state fatte circa l'attendibilità dei risultati: Barbaro per esempio ci ha parlato della sovrapposizione delle curve stimate e delle curve osservate rispetto alla variabile da predire. Sorge tuttavia il dubbio che in quel caso vi possa essere una sorta di favoreggiamento dovuto ad *autofitting*: si cerca un modello ottimale sulla base del dato da stimare e chiaramente c'è una verifica endogena che produce risultati sovrastimati. Potrebbero forse essere utilizzati anche altri metodi: dalla *cross-validation*, all'utilizzo di previsori per il futuro da verificarsi con dati successivamente disponibili.

Con riferimento alla robustezza delle tecniche utilizzate potrebbe essere utile il fornire gli intervalli di confidenza o gli errori standard delle stime effettuate. Ad esempio nella costruzione della serie delle ricostruzioni a livello territoriale intervengono stime di probabilità, di appartenenza ad un determinato territorio. Queste probabilità sono stime, quindi sono dotate di un errore standard che potrebbe essere utilizzato per dare una valutazione della incertezza legata alla attribuzione di tali somme a livello di singolo territorio. Il tema dell'incertezza è sicuramente da affrontare in questo contesto.

Un tema interessante da approfondire, proprio nel contesto del *data mining*, è l'utilizzo di tecniche esplorative. Mi sembra, per esempio, di poter suggerire, nel contesto di modelli di tipo regressivo o logistico-regressivo, l'utilità di un'esplorazione preliminare dei dati. Nel caso della relazione di Barbaro si ha la disponibilità di una variabile predicenda e di una variabile predittrice, rispettivamente ISTAT e Ragioneria Generale dello Stato. Per vedere se Regioni - o unità territoriali in generale - e tempi di rilevazione sono rilevanti si potrebbe effettuare una analisi della varianza a due vie. In questo caso la variabile risposta è il rapporto tra le due

variabili indicate, nella relazione di Barbaro, con Y ed X. Attraverso un modello ad effetti semplici si potrebbe cogliere la significatività dell'effetto della Regione e dell'effetto del tempo.

Un altro possibile percorso di analisi è dato dall'utilizzo, compattato, delle variabili con la dimensione J delle categorie di spesa e con la dimensione K delle Amministrazioni. Sostanzialmente compattando queste dimensioni si costruisce un vettore a tre vie in cui abbiamo Regioni, o unità territoriali, e tempi di osservazione. Su queste variabili X ed Y, osservate in diverse situazioni, possiamo effettuare analisi per individuare fattori latenti, per vedere se esiste un modello di collegamento tra queste variabili. Tale modello trova una sua forte giustificazione quanto più esiste un fattore latente tra il dato di fonte ISTAT e la variabile Ragioneria Generale dello Stato.

Un elemento su cui soffermare l'attenzione è rappresentato dall'utilizzo delle informazioni disponibili in fase di stima dei parametri. Per stimare i parametri di un modello, sia di tipo regressivo che di tipo logistico, ci si riferisce ad un collettivo di osservazioni. In generale il collettivo di osservazioni deve essere dotato della proprietà di ripetibilità e replicabilità. Nei modelli presentati non mi è chiaro se i collettivi di osservazioni scelti sono significativi a questo proposito. La relazione di Di Manno ha toccato questo aspetto precisando che la stima viene effettuata per ogni classe, per anno e per importo in considerazione del fatto che la classe d'importo modifica strutturalmente il modello. E' importante approfondire il tema del collettivo di osservazioni replicabili in cui vale lo stesso modello strutturale, ai fini di avere un numero sufficiente di osservazioni per dare una stima attendibile dei parametri del modello.

E' importante, per arricchire ulteriormente il percorso svolto, ampliare la strumentazione metodologica anche a livello esplorativo. Il *data mining* prevede due fasi distinte: una fase di esplorazione ed una di modellizzazione dei dati. La fase di esplorazione potrebbe essere arricchita per approfondire i limiti dei modelli finora utilizzati, che comunque offrono già risultati valutabili molto positivamente, sia nel caso del modello regressivo, sia nel caso dei modelli logistici utilizzati per ridistribuire le spese a livello territoriale.