CORONA VIRUS e STATISTICA [1]

Alcuni esempi di cattivo (quantomeno discutibile) impiego di un utile strumento di analisi e qualche interrogativo.

1. Gli esempi

In un recente articolo di Enrico Bucci e Enzo Marinari (L'evoluzione dell'epidemia da Corona Virus in Italia, pubblicato il 2/3/2020 su FACEBOOK e TWITTER, Scienza in rete, il Gruppo 2003 per la Ricerca Scientifica) sono riportati i due seguenti grafici (Fig. 1).

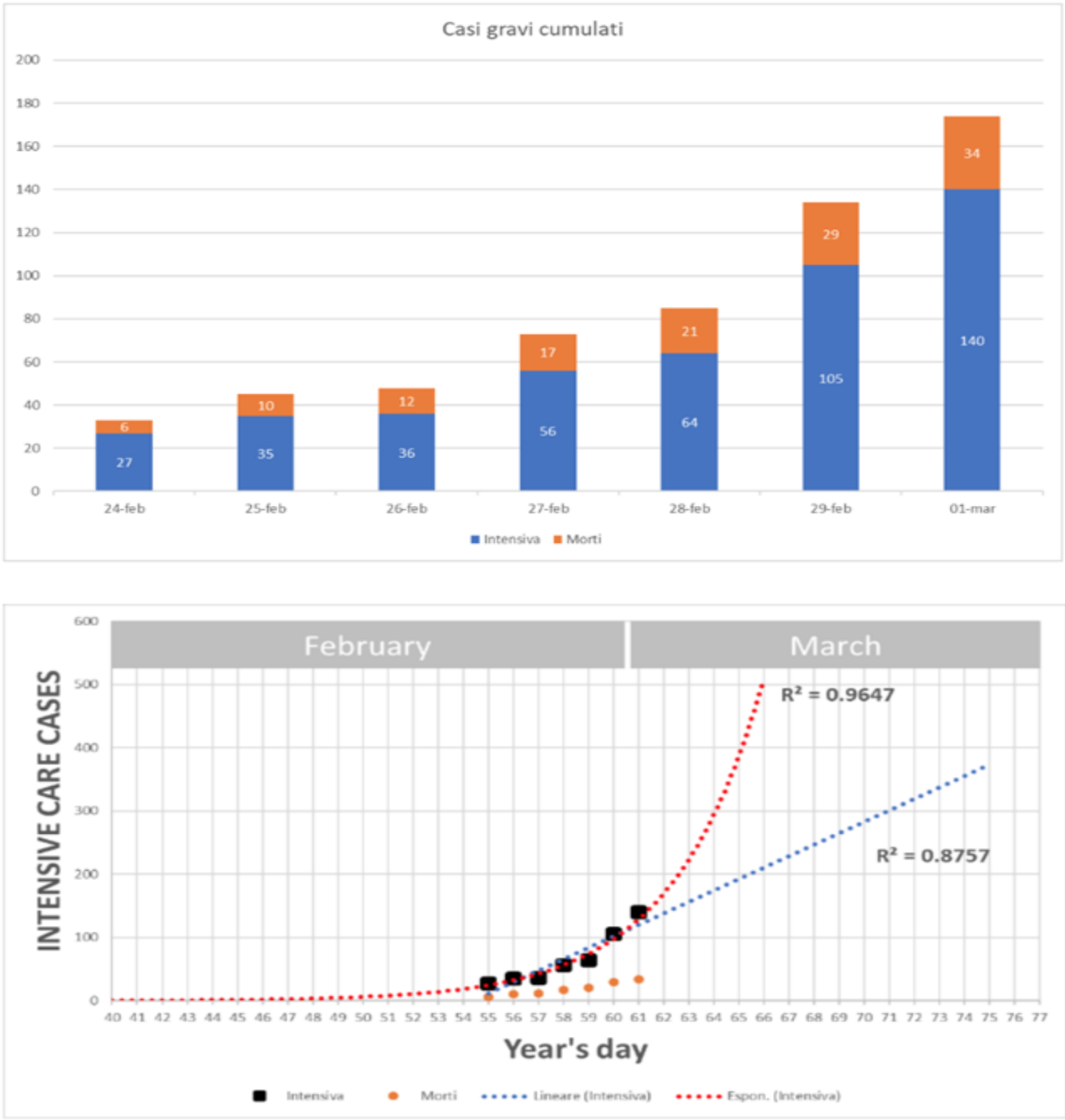
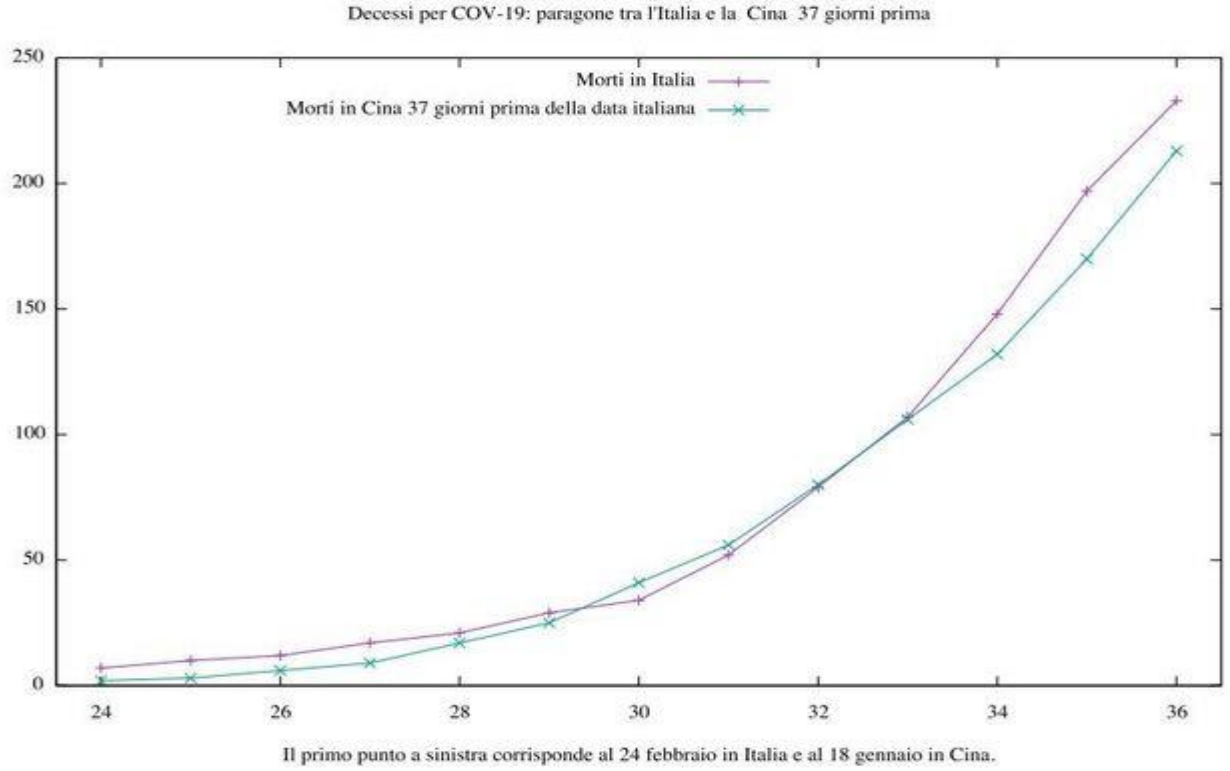


Fig. 1

[1] Nota a margine (introduzione) dell'insegnamento di Teoria Statistica delle Decisioni programmato, e rinviato, per il mese di marzo del Corso di Dottorato in Statistica dell'Università degli Studi di Firenze.

Nessuna osservazione sul primo grafico, relativamente al secondo grafico, tenendo conto della natura del fenomeno analizzato, la prima osservazione da fare riguarda l'impiego di modelli interpolativi disponendo di un numero molto limitato di casi (7). La seconda è relativa alla tipologia di modelli impiegati: il modello lineare e il modello esponenziale sono del tutto inadeguati, come verrà chiarito nelle righe seguenti, per la rappresentazione del fenomeno in esame. La terza osservazione concerne il calcolo distorcente dell'indice di adattamento R^2 che può indurre il lettore a concludere erroneamente sulla validità del modello esponenziale a ragione dell'elevato valore $R^2 = 0{,}96$ assunto dall'indice.

Nel post dell'8.3.2020 "*L'epidemia rallenterà certo prima di Pasqua ma non è una buona notizia*" scritto da Giorgio Parisi (Fisico, presidente dell'Accademia Nazionale dei Lincei) e Luca Foresti (Fisico, Chief Executive Officer del Centro Medico Santagostino) sono riportati i due grafici seguenti (Fig.2).



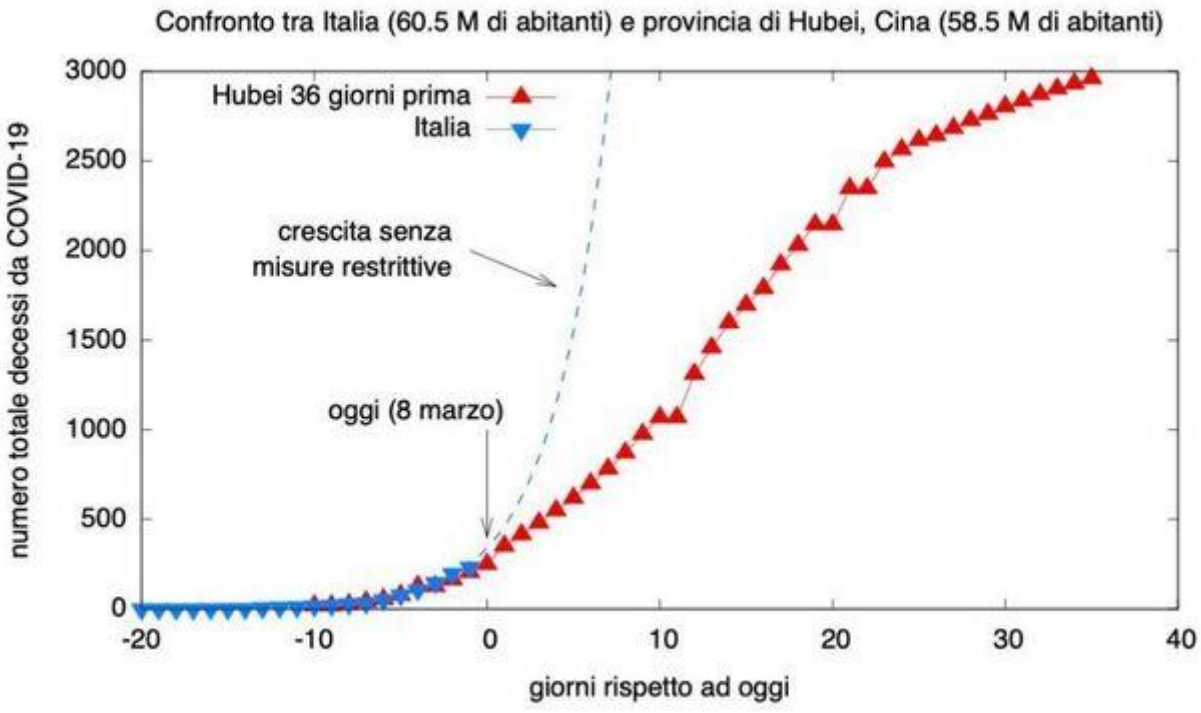


Fig. 2

Molto apprezzabile e il primo grafico e del tutto condivisibili sono le conclusioni interpretative cui giungono i due autori quale, ad esempio, che lo scostamento osservato tra gli ultimi tre punti di osservazione (morti in Cina e morti in Italia) sia da attribuire al maggior rigore delle misure di contenimento adottate nella provincia di Hubei.

Del tutto opinabile è invece il secondo grafico, in questo caso, rispetto a quanto sopra osservato riguardo le Fig. 1, il numero delle osservazioni è tale da giustificare l'impiego di un modello interpolativo ma il modello esponenziale impiegato, come già sottolineato, non è adeguato per la rappresentazione dei fenomeni epidemici di contagio. Infatti, l'evoluzione tipica dei fenomeni epidemici si caratterizza per una fase iniziale di sviluppo moderato cui segue una fase di accelerazione che si attenua con il passare del tempo con tassi di incremento che si riducono progressivamente fino ad annullarsi quando l'intera popolazione interessata risulta contagiata (saturazione del fenomeno).

Nella Fig. 3 sono riportate tre diverse tipologie evolutive dei fenomeni epidemici, nella fase iniziale di evoluzione del fenomeno sia il modello esponenziale che il modello logistico[2] forniscono un ottimo adattamento ai dati osservati. Se la finalità del modello interpolativo e di tipo previsionale il modello esponenziale si rivela del tutto inadeguato come risulta in modo molto evidente osservando anche la Fig.2. Comunque, anche indipendentemente dalla finalità di impiego, la "bontà" rappresentativa di un modello interpolativo deve essere sempre valutata tenendo conto dell'intero "ciclo di vita" del fenomeno analizzato.

[2] Il modello logistico è stato proposto da P.F. Verhlust in un articolo pubblicato nel 1838 e meglio caratterizzato in due articoli successivi del 1845. Il modello è stato riproposto (i due autori non citano Verhulst) da R. Pearl e L.J. Reed nel 1920. Successivamente il modello è stato generalizzato, soprattutto attraverso l'inserimento di variabili esplicative per meglio soddisfare specifiche esigenze di ricerca.

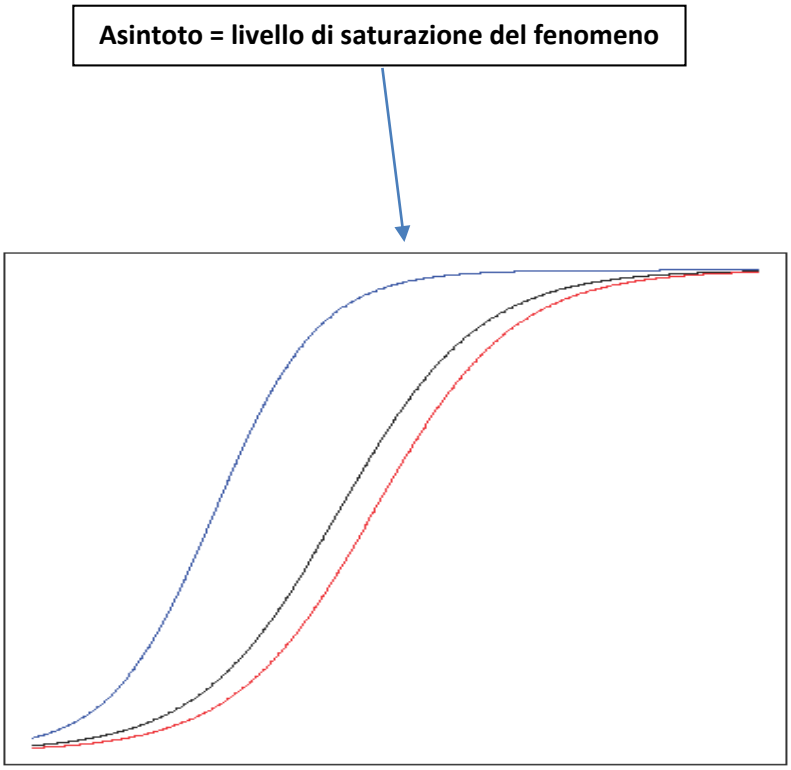


Fig. 3

L'inadeguatezza del modello esponenziale impiegato si riflette anche sulle conclusioni che dalla lettura del secondo grafico riportato nella Fig.2 conseguono; infatti, attribuire implicitamente una valenza causale alle misure restrittive poste in essere quali determinanti della variazione dell'andamento del fenomeno nella provincia di Hubei è del tutto priva di fondamento, l'unica corretta conclusione cui si può pervenire nel caso in esame è che il modello esponenziale non è in grado di rappresentare adeguatamente il fenomeno analizzato.

Volendo pervenire a delle conclusioni in merito ad interventi che si ritiene possano incidere sull'evoluzione naturale di un fenomeno epidemico, limitandosi alla osservazione di figure, il confronto non deve essere fatto tra il dato teorico derivante da un modello e il dato empirico osservato ma tra valori teorici derivanti dalla stima del modello utilizzando i dati osservati prima dell'applicazione degli interventi, e quelli derivanti dalla stima dello stesso modello sui dati osservati dopo un ragionevole lasso di tempo trascorso dall'attivazione degli interventi.

2. Gli interrogativi

Partendo dal presupposto che lo sviluppo naturale dei fenomeni di contagio può essere convenientemente rappresentato attraverso un modello logistico che ne descrive l'andamento ci si può domandare cosa accade se intervengono fattori esterni che possono inciderne sull'evoluzione naturale. Fattori che possono:

1. bloccare l'evoluzione;
2. modificare i tassi di incremento delle diverse fasi di sviluppo;
3. abbassare il livello di saturazione.

Ad oggi (13 marzo 2020), gli interventi (fattori esterni) attuati dal Governo italiano non hanno certamente bloccato la diffusione del fenomeno epidemico, sia in termini di persone infette sia in termini di decessi,

quello che si può ragionevolmente concludere è una attenuazione dei tassi di incremento (velocità di diffusione del virus); al momento, non disponendo di informazioni sufficienti, non si può avanzare alcuna ipotesi interpretativa riguardo a quanto previsto al punto 3.

Accertata la presenza di un focolaio infettivo[3] in Italia, il governo poteva decidere di non intervenire o intervenire con l'obiettivo di bloccare la diffusione del fenomeno. La scelta è stata l'intervento. Si tratta di una decisione razionale in quanto la diffusione del fenomeno infettivo provoca un incremento di persone malate che devono essere assistite, con il conseguente possibile collasso del sistema sanitario, ed incremento del numero dei decessi.

Ma la scelta, sia in termini di potenziamento delle strutture sanitarie che di imposizione di misure restrittive di comportamento imposte alla popolazione è stata effettuata valutando esclusivamente le possibili conseguenze del processo epidemico solo in termini di stato di salute della popolazione e di impatto sul sistema sanitario. Al momento, come sopra sottolineato, l'unica ragionevole conclusione cui si può pervenire è che le misure restrittive adottate non hanno avuto come conseguenza il blocco del fenomeno epidemico ma che, molto verosimilmente, hanno attenuato il suo tasso di sviluppo

Ho letto, non ricordo dove, che una rapida diffusione di una epidemia può determinare un abbassamento del livello di saturazione del fenomeno essendo i soggetti guariti immuni dal contagio. Bisogna chiedersi, pertanto (***primo interrogativo***), se la riduzione del ritmo di diffusione del contagio sia, anche dal punto di vista strettamente sanitario, davvero auspicabile.

Alla luce di quanto ad oggi accertato, tenendo anche conto del fatto che al momento il processo epidemico ha interessato solo marginalmente il centro sud e le isole, territorio che molto verosimilmente sarà soggetto a quanto già avvenuto nel nord Italia, ci si deve domandare se la decisione di intervenire con misure restrittive nei soli territori focolaio di infezione, sia stata (***secondo interrogativo***), la migliore. Ovviamente il problema non si pone se le drastiche misure restrittive adottate dal governo l'11 marzo si riveleranno tali da comportare il blocco del processo epidemico.

Come sottolineato, il processo decisionale fin qui delineato si è svolto considerando le conseguenze della scelta solo in termini sanitari (malati, decessi e impatto sul sistema sanitario), ci si deve domandare (***terzo interrogativo***) se un tal modo di procedere è stato il migliore possibile tenendo presente che è stato completamente trascurato l'impatto devastante, sia in termini economici che sociali, delle misure restrittive.

Non sarebbe stato (***quarto interrogativo***) più razionale esprimere le conseguenze dei provvedimenti restrittivi in termini di utilità sociale attraverso una funzione basata non esclusivamente sugli effetti sanitari e sulla contrapposizione effetti sanitari–effetti economici, ma considerano tutti i possibili effetti della scelta operata? In altri termini, tenendo anche conto del fatto che i provvedimenti restrittivi hanno un impatto sociale altrettanto devastante quanto quello economico e che entrambi gli effetti (economico e sociale) non hanno una durata temporale limitata nel tempo ma si protrarranno anche dopo l'esaurimento del processo epidemico.

Infine (***quinto interrogativo***), alla luce di quanto presente nella letteratura scientifica, non sarebbe risultato conveniente analizzare il problema dell'individuazione delle migliore strategia per affrontare il fenomeno epidemico inquadrando tutta la problematica in uno adeguato contesto decisionale.

[3] I primi due casi di Coronavirus in Italia, una coppia di turisti cinesi, sono stati confermati il 30 gennaio dall'Istituto Spallanzani, dove sono stati ricoverati in isolamento dal 29 gennaio. Il 26 febbraio sono stati dichiarati guariti.
Il primo caso di trasmissione secondaria si è verificato a Codogno, Comune della Lombardia in provincia di Lodi, il 18 febbraio 2020.

Qualunque processo decisionale può essere rappresentato da una tabella, come quella sotto riportata, dove le azioni alternative possibili (in numero discreto) a_i sono indicate nella prima colonna, gli stati di natura θ_j (che possono essere anche espressi mediante un modello probabilistico continuo) nella prima riga e le conseguenze, espresse in termini di utilità

$$u_{ij} = u\left(a_i, \theta_j\right)$$

all'interno della tabella.

Azione	**Stato di natura**					
	θ_1	θ_2	.....	θ_j	.....	θ_s
a_1	u_{11}	u_{12}	.....	u_{1j}	.....	u_{1s}
a_2	u_{21}	u_{22}	.....	u_{2j}	.....	u_{2s}
....	.....	.....		.....		.....
a_i	u_{i1}	u_{i2}	.....	u_{ij}	.....	u_{is}
....	.....	.....		.....		.....
a_m	u_{m1}	u_{m2}	.....	u_{mj}	.....	u_{ms}

Nel caso del COVID-19 le azioni alternative (decisioni) possibili possono essere:

a_1 = nessun intervento (il fenomeno si evolve naturalmente);

a_2 = si interviene potenziando le strutture sanitarie;

a_3 = si interviene adottando solo misure restrittive;

a_4 = si interviene potenziando le strutture sanitarie adottando, allo stesso tempo, lievi misure restrittive

a_5 = si interviene potenziando le strutture sanitarie adottando , allo stesso tempo, drastiche misure restrittive;

……

Il problema della scelta dell'azione migliore si risolve ricorrendo alla teoria decisionale più consona al caso in esame tenendo conto del patrimonio informativo disponibile:

1. Nessuna informazione disponibile ⟹ Teoria classica delle decisioni;
2. Sola informazione a priori disponibile ⟹ Teoria bayesiana delle decisioni;
3. Sola informazione campionaria disponibile ⟹ Teoria statistica classica delle decisioni;
4. Informazione a priori e informazione campionaria disponibile ⟹ Teoria statistica bayesiana delle decisioni;
5. Informazione a priori, informazione campionaria e nessi di causalità ⟹ Teoria causale bayesiana delle decisioni statistiche.

Relativamente alle informazioni disponibili e alla loro qualità si deve osservare che, nel caso specifico del COVID -19, sono state espresse molte perplessità sul numero effettivo di persone infette. Anche in questo

caso gli statistici possono contribuire efficacemente, predisponendo specifici piani di campionamento casuale sulla popolazione potenzialmente soggetta ad infezione sulla quale applicare tamponi.

***

Nel 1982, all'inizio della mia attività di consulente scientifico della Rai-Radiotelevisione Italiana, mi è stato chiesto di effettuare una previsione delle famiglie italiane abbonate nel 1983. La finalità della Rai era quella di avere una misura dei presumibili introiti provenienti da canone per programmare la propria attività.

Il problema della specificazione della funzione di utilità non si è presentato essendo l'utilità espressa dal numero delle famiglie abbonate e quindi dall'ammontare degli introiti monetari da canone. Estremamente più complesso è il problema della specificazione di una funzione di *utilità sociale* significativa per il COVID-19; specificazione che richiede il coinvolgimento di numerosi soggetti: ricercatori di vari ambiti disciplinari, istituzioni sanitarie e di governo (nazionale e locale), organizzazioni sindacali e di categoria, ecc. . Ovviamente, l'utilità può essere espressa considerando i soli aspetti sanitari (contagiati e morti) e procedere alla scelta dell'azione migliore monitorandone gli effetti positivi intervenendo di conseguenza rinviando la valutazione degli effetti negativi di ordine economico e sociale per procedere, successivamente, all'attivazione di interventi in grado di mitigarne gli effetti. Questa è la strategia adottata dal governo italiano, (***ultimo interrogativo***) è stata la scelta migliore?

Per soddisfare la richiesta della Rai ho sperimentato svariati metodi (più o meno sofisticati) di previsione, la scelta finale è stata quella di impiegare il modello logistico nella formulazione

$$y_t = e^{a+b\cdot c^t} + u_t$$

dove, *t* rappresenta il tempo e y_t , l'intensità del fenomeno (il numero di famiglie abbonate), *a*, *b* e *c* i parametri che specificano il modello e che devono essere stimati utilizzando i dati disponibili.

Il modello logistico stimato ha registrato un adattamento ai dati elevatissimo (R^2> 0,99) e un errore di previsione (calcolato l'anno successivo) inferiore all'1%.

Successivamente mi è stato chiesto di valutare l'effetto degli interventi attivati dalla Direzione abbonamenti nei confronti delle famiglie in possesso di un apparecchio TV non abbonate. Ovviamente la finalità degli interventi attivati dalla Rai si sono mossi nella direzione opposta a quella sopra richiamata, gli interventi erano, infatti, mirati alla accelerazione, e non alla eliminazione del "fenomeno infettivo" o alla riduzione dei tassi di contagio.

L'analisi causale (date le cause determinare gli effetti) è stata effettuata nel 1985 e riportata come caso di studio nel testo base di riferimento dell'insegnamento di Teoria statistica delle decisioni del dottorato: S. Bacci e B. Chiandotto, *Introduction to Statistical Decision* …..pubblicato dalla Chapman &Hall/CRC nel 2019.

Successivamente[4] ho avuto modo di verificare che gli effetti positivi degli interventi effettuati annualmente dalla Rai oltre a comportare una accelerazione del processo di diffusione hanno determinato anche un innalzamento del livello di saturazione.

[4] Le analisi, previsionale e causale, sono state ripetute, con affinamenti successivi (es. sono state inserite variabili esplicative nel modello logistico), tutti gli anni fino al 2015. Dal 2016 la Rai non si occupa più dell'acquisizione degli abbonamenti relativi alle famiglie, potenzialmente soggette a canone ma non iscritte a ruolo, a seguito dell'emanazione del provvedimento che prevede il pagamento del canone stesso tramite bolletta elettrica.
Delle osservazioni contenute in questa nota avremo modo di discutere nella penultima lezione del ciclo previsto per l'insegnamento di Teoria Statistica delle decisioni.

Chiudo la nota segnalando il lavoro di Giovanni Sebastiani "*Alcuni risultati dell'analisi dei dati epidemiologici del Coronavirus in Italia* " quale esempio di ottimo impiego della Statistica.

Bruno Chiandotto

Firenze, 14 marzo 2020 ore 12

***

Addendum inserito nella Nota il 22 marzo 2020

Fasi operative per procedere alla stima di un modello logistico

Considerando la formulazione del modello logistico

$$y_t = e^{\alpha+\beta\cdot\gamma^t} + u_t \qquad (1)$$

dove, *t =1, 2, …..T* rappresenta il tempo, ad esempio giorni, *T* il numero di osservazioni disponibili) y_t , l'intensità del fenomeno (numero cumulato di persone contagiate o numero cumulato di decessi), α , β e γ , i parametri che specificano il modello e che devono essere stimati utilizzando i dati disponibili, u_t la componente accidentale.

Inserendo stime dei parametri $\hat{\alpha}$, $\hat{\beta}$ e $\hat{\gamma}$ si ha

$$\hat{y}_t = e^{\hat{\alpha}+\hat{\beta}\cdot\hat{\gamma}^t} \; . \qquad (2)$$

La stima giornaliera del modello, ad un livello territoriale di riferimento significativo (italiano, regionale, provinciale, comunale, ecc.)[5], fornisce una misura teorica dell'intensità del fenomeno indagato (numero cumulato di infetti o di decessi)

Attribuendo a *t* i valori *T+1, T+2, … , T+h* ($h \longrightarrow +\infty$) si ottengono le stime del valori previsti dal modello.

Il modello (2) non fornisce soltanto i valori teorici ma consente anche, attraverso il calcolo della derivata rispetto a *t*, di ottenere la stima dell'intensità giornaliera del fenomeno oggetto di analisi; il massimo della derivata fornisce la stima del '***picco***'.

Se si confrontano i valori teorici (valori interpolati) forniti dal modello in due giorni diversi, adeguatamente distanziati, si perviene, in via di prima approssimazione, ad una valutazione dell'impatto degli interventi decisi dalle autorità di governo nazionale e locale. Valutazione che può essere effettuata anche attraverso il semplice confronto visivo delle curve interpolate. Qualcuno ha sostenuto che a volte un grafico (una

[5] Quando si usano dati aggregati a diversi livelli bisogna tener conto della così detta ***fallacia ecologica***. Quando, ad esempio, si stimano i dati aggregati a livello comunale, la loro somma non coincide con il dato stimato con i dati aggregati a livello provinciale.

immagine) può valere più di mille parole. La frase che è stata pronunciata da Mao Tse Tung, non so in quale occasione, dovrebbe risalire al filosofo cinese Confucio.

Un esempio molto significativo dell'uso di immagini sono le facce di Chernoff (1973) che consentono la rappresentazione di dati a K < 19 dimensioni. Successivamente sono stati pubblicati contributi che consentono la rappresentazione per un numero di caratteri notevolmente superiore a 18.

Tre osservazioni finali.

1. L'impiego del modello logistico per fornire una risposta a domande formulate dagli *stakeholders* è conforme al ***rasoio di occam***, principio metodologico che suggerisce, ai fini della risoluzione di un problema, di scegliere, tra i percorsi possibili, quello più semplice a meno che non sia necessario agire altrimenti. Come sopra sottolineato, il percorso da seguire nel caso in esame, il modello logistico può rappresentare solo il primo passo, per soddisfare esigenze operative urgenti, di un percorso molto più lungo ed impegnativo.
2. Quando si impiegano modelli probabilistici e statistici , un impatto molto rilevante è rappresentato dalla qualità dei dati utilizzati, occorre, quindi, fare riferimento anche ad un altro principio noto in letteratura: ***garbage in garbage out***.
3. Quando si procede alla misura della capacità rappresentativa di un modello non bisogna limitarsi al calcolo di indici che misurano la bontà di adattamento quali l'R^2. Per evitare i pericoli dell'***overfitting***, è la conoscenza (oggettiva e soggettiva) del fenomeno analizzato l'elemento fondamentale di riferimento.