GIUSTIZIA INSIEME

Diritto dell"emergenza Covid-19 e recovery fund

I dati non parlano da soli: l'epoca del Coronavirus smaschera l'inganno dell'algoritmo-onnipotente e rivaluta il metodo statistico di Giuseppe Arbia e Vincenzo Nardelli

di **Giuseppe Arbia**

17 marzo 2020

I dati non parlano da soli: l'epoca del Coronavirus smaschera l'inganno dell'algoritmo-onnipotente e rivaluta il metodo statistico

Giuseppe Arbia e Vincenzo Nardelli

Sommario: 1. Premesse.- 2. La differenza tra "dato" e "informazione.- 3. Il contributo della Statistica ed il metodo scientifico-induttivo.- 4. Modelli statistici e previsioni della diffusione del Coronavirus – 4.1 Verifica dell'efficacia della manovra di *lockdown* – 4.2 Stima del momento di "*picco epidemico*"- 4.3 Stima del "*tasso netto di riproduzione*" – 5. Conclusioni

1. Premesse

Una decina di anni fa Chris Anderson, l'Amministratore Delegato della società 3D Robotics, dalle colonne della rivista online "Wired", vaticinava la fine del metodo scientifico e l'avvento della tirannia dei dati affermando tra le altre cose che: "Con un sufficiente quantitativo di dati i numeri parleranno da soli"[1]. Era l'inizio della rivoluzione dei Big Data[2] e del diffondersi della illusione che la mera disponibilità di dati e di una strumentazione informatica in grado di trattarli in tempo reale sarebbe stata sufficiente a portare alla conoscenza approfondita di tutti i fenomeni empirici. Ambienti un tempo tradizionalmente impermeabili ad un approccio basato sui dati si sono andati via via convincendo che nella ricerca di una migliore conoscenza, molte operazioni potessero essere delegate ad un algoritmo di trattamento automatico dei dati. Qualora ve ne fosse bisogno, in questi giorni l'emergenza sanitaria imposta dall'epidemia di Coronavirus sta mostrando come, a fronte di un'enorme disponibilità di dati, non siamo in grado di rispondere alle domande fondamentali che davvero ci interessano, cioè: quando giungeremo al picco dell'epidemia? le misure prese sono efficaci? quando finalmente terminerà l'epidemia e torneremo alla normalità? quale sarà il prezzo che dovremo pagare in termini di vite umane? In questo articolo ci proponiamo di mostrare, facendo proprio riferimento alla emergenza legata alla diffusione del Covid-19, come è solo il metodo scientifico ed una corretta applicazione di modelli statistici a dati osservazionali e non semplici algoritmi, che possono fornire risposte adeguate alle domande e suggerire decisioni politiche scientificamente fondate.

2. La differenza tra "dato" e "informazione"

L'Italia vive in questi giorni un'emergenza sanitaria senza precedenti nella quale tutte le forze del paese sono mobilitate al fine di limitare i danni, in termini di vite umane e di costi sociali, e di ritornare al più presto alla normalità. In un momento così difficile per il nostro paese molti cittadini si rivolgono quotidianamente alla lettura di dati statistici per avere una piena coscienza di ciò che sta accadendo, oltre che per avere un conforto ed un sostegno delle volontà nello sforzo di uscire rapidamente dall'emergenza. Ogni giorno si attende con ansia il bollettino della Protezione Civile delle ore 18 nella speranza di trarne conforto. Tuttavia, molto spesso, dalla lettura dei dati a disposizione, il cittadino ne esce al contrario confuso e disorientato. Se "i dati parlassero da soli" (come afferma Chris Anderson) perché in questa emergenza, nella quale disponiamo di tantissimi dati, essi non riescono a darci risposte esaurienti?

In questo appare evidente la differenza sostanziale che sussiste tra "dato" e "informazione".

Parte dell'ambiguità nell'uso dei due termini è certamente da ascrivere all'uso che facciamo della parola *informatica* (termine introdotto da Philippe Dreyfus nel 1962 dalla contrazione dei termini francesi *informat(ion) (automat)ique)*. In realtà la nuova disciplina avrebbe dovuto essere più correttamente denominata *Datamatica* perché è ai dati che essa si riferisce e non di per sé alle informazioni in essi contenute**[3]**.

Tanti dati non rappresentano, infatti, tante informazioni. Facciamo un piccolo esempio: un documento contenente la Bibbia può essere immagazzinato in un file di circa 20 Kilobyte circa, mentre la fotografia che scatto con uno *smartphone* di media qualità al piatto di pesce che sto mangiando, ne richiede almeno 100. Se condivido poi la foto con i miei i 20 amici della *chat*, essa produce un totale di 2 Megabyte. In altri termini in un secondo ho prodotto lo stesso volume di dati che corrisponde a 100 volte la Bibbia; con un contenuto informativo a confronto che lascio giudicare al lettore!

La distinzione tra dato ed informazione è ben espressa dal popolare autore di romanzi di fantascienza, Dabiel Keys Moran quando afferma che: "*Possiamo avere dati senza informazioni, ma non informazioni senza dati".*

Non trovo mai, tuttavia, un modo migliore per chiarire questo tema della lettura della novella *La biblioteca di Babele* di Jorge Luis Borges**[4]**. In tale novella, l'autore considera un libro come un oggetto costituito da una sequenza di circa un milione di caratteri, ottenuto come il prodotto di (diciamo) 400 pagine x 40 righe a pagina x 60 caratteri per riga. Dato che tali caratteri possono assumere un numero limitato di valori (le lettere dell'alfabeto, lo spazio, il punto, la virgola, i due punti, il punto e virgola ecc.), tutte le loro possibili combinazioni nelle 400 pagine del libro danno luogo a tutti i libri che è possibile concepire: un numero non infinito, ma comunque elevatissimo **[5]**. A partire da tali considerazioni, Borges immagina che esista una biblioteca la quale possa contenere tutti i possibili libri, ma ci mette in guardia dal gioire di ciò giacché, in assenza di un catalogo veritiero e dettagliato della stessa, la probabilità di trovare il libro che ci interessa è sostanzialmente zero.

Questo racconto ci fa riflettere sul fatto che molti dati non implicano affatto conoscenza. Moltissimi volumi della biblioteca fantastica sono, infatti, costituiti solo da pagine vuote con qualche carattere sparso qui e lì. La stragrande maggioranza di essi compone parole senza un senso compiuto. Tra i pochi volumi che contengono parole con un senso compiuto moltissimi riportano realtà ingannevoli non corrispondenti alla verità.

È esattamente questa la situazione nella quale ci troviamo attualmente. Abbiamo a disposizione una quantità smisurata di dati, ma molti di questi sono inutilizzabili al fine di costruire un insieme informativo affidabile che ci porti ad una conoscenza dei fenomeni e dunque ad una base obiettiva sulla quale fondare le nostre scelte individuali e collettive.

Sono quattro, infatti, le fasi fondamentali di una conoscenza basata sui dati: (i) la raccolta e l'immagazzinamento del dato di base, (ii) l'estrazione dell'informazione da esso, (iii) la costruzione della conoscenza che dalla maggiore informazione deriva e, infine, (iv) l'uso di tale accresciuta conoscenza per prendere decisioni migliori empiricamente fondate.

Sono questi i passi che compie lo scienziato, il manager, il decisore politico così come l'uomo comune nel prendere le decisioni piccole e grandi di ogni giorno.

La lettura di dati empirici, in effetti, non è sempre immediata ed univoca e necessita che insieme ad essi vengano forniti strumenti di sintesi e chiavi di lettura per una loro corretta interpretazione ed una piena comprensione del fenomeno che essi descrivono: un compito che ha da sempre tradizionalmente svolto la Statistica.

3. Il contributo della Statistica ed il metodo scientifico-induttivo

Dopo questa lunga, ma credo utile, premessa, veniamo ai giorni nostri e al bombardamento di dati al quale veniamo sottoposti quotidianamente se vogliamo tenerci informati circa l'evoluzione della epidemia. Che aiuto possiamo attenderci dalla Statistica?

Due sono i compiti ai quali assolve la Statistica. Da un lato fornire una sintesi e una descrizione dello stato di fatto, dall'altro individuare regolarità nascoste, estendere i risultati dall'insieme necessariamente limitato che si è osservato al fenomeno nella sua interezza e, in ultima analisi, prevedere sviluppi futuri.

Da un punto di vista descrittivo possiamo misurare una serie grandezze interessanti che quantifichino l'entità del fenomeno fotografato in un determinato istante, quali, ad esempio la percentuale dei guariti tra gli infetti, la percentuale dei malati che richiedono una terapia intensiva, la letalità del virus (intesa come percentuale di decessi tra gli infetti), la sua mortalità (intesa come percentuale dei decessi sul totale della popolazione ecc.). Al momento in cui scrivo, ad esempio, leggiamo che la percentuale dei positivi al virus sulla popolazione è dello 0.03%

(ovvero 3 infetti ogni 10,000 individui), il rapporto tra il numero di ricoverati in terapia intensiva e i positivi è circa l'8%, cosi come è del 7% circa la letalità del virus e la sua mortalità raggiunge circa lo 0.002% (ovvero 2 individui su 100,000).

Possiamo, tuttavia, generalizzare questi risultati e riferirli *tout court* all'intera popolazione italiana?

La risposta è no.

Queste informazioni, infatti, pur preziosissime, si riferiscono solo a ciò che è stato osservato, e non a tutto l'osservabile. In generale, dato che di norma non possiamo osservare tutta la realtà (in tutti i suoi stati di natura passati, presenti e futuri), siamo costretti ad osservarne solo una parte (ciò che abbiamo visto fino ad oggi) e cerchiamo di trarre da queste osservazioni limitate considerazioni di tipo generale. Ciò è necessariamente vero nel caso della rilevazione del diffondersi di un fatto nuovo ed inatteso come un'epidemia influenzale. Nel linguaggio statistico ciò che osserviamo è un *campione* (ciò che i filosofi chiamano il *fenomeno* – *fainòmenon* – "ciò che appare") opposto a ciò che vorremmo osservare che lo statistico chiama popolazione (e che i filosofi chiamano invece l'universale – *noumeno* – "ciò che può essere pensato"). Nel ricavare informazioni universalmente valide da quanto osservato in un campione seguiamo dunque un procedimento statistico-induttivo.

Mentre agli albori della Statistica ci si affidava alla vaga speranza che un campione, comunque raccolto, rappresentasse bene l'intera popolazione, anche a seguito di clamorosi fallimenti, alcuni statistici (tra i quali Jerzy Neyman negli anni '30) del secolo scorso misero in guardia i ricercatori sull'uso di dati raccolti senza un rigoroso criterio di raccolta. Sempre negli anni '30 Sir Ronald Fisher (a giusta ragione ritenuto il padre della scienza statistica così come la intendiamo oggi) è tra i primi ad affermare che, affinché un'analisi statistica possa condurre ad una generalizzazione soddisfacente, gli esperimenti campionari devono essere programmati in maniera rigorosa**[6]**. La necessità che le osservazioni empiriche debbano essere raccolte seguendo un criterio rigoroso per poter costituire informazione valida alla costruzione della conoscenza è affermata anche dal matematico e filosofo della scienza Jules Henri Poincarè, nel suo celebre *"Scienza e metodo"* **[7]** *quando afferma: "Il metodo scientifico consiste nell'osservare e nello sperimentare; se lo scienziato disponesse di un tempo infinito, non ci sarebbe altro che dirgli , ma dato che gli manca il tempo di guardare tutto e tanto meno di guardare con attenzione – ed è meglio non guardare che guardare malamente – si trova nella necessità di fare una scelta. Sapere in che modo deve fare questa scelta è dunque la prima questione."*

Senza entrare in dettagli tecnici, per poter generalizzare in maniera soddisfacente i dati osservati all'intera popolazione, le osservazioni devono essere selezionate tramite un rigoroso sistema di campionamento (detto campione *casuale*) nel quale ogni unità ha la medesima probabilità di essere estratta (garantendo, per così dire, una certa qual oggettività del criterio di raccolta) in opposizione a campionamenti effettuati secondo criteri vari dettati dalla immediata disponibilità, dalla convenienza o dalla facilità di raccolta. Tale condizione non è evidentemente soddisfatta nel caso dei dati epidemiologici durante l'esplosione di un'epidemia, i quali vengono raccolti così come ci giungono senza un preciso piano degli esperimenti seguendo criteri di pura disponibilità.

Le percentuali alle quali abbiamo fatto riferimento precedentemente (percentuale dei malati sulla popolazione, la percentuale degli infetti che richiedono una terapia intensiva la letalità e la mortalità del virus) sono basati su dati che non sono stati raccolti con un criterio puramente casuale su tutto il territorio nazionale, ma al contrario sono (ovviamente) concentrati nelle regioni maggiormente colpite. Essa risulteranno per tale ragione stime inaffidabili dei veri valori che si intendono stimare a livello di tutta la popolazione.

Accanto a tali errori dovuti al criterio di raccolta, inoltre, le stesse grandezze sono affette da altri tipi di distorsione**[8]**. Ad esempio, la percentuale dell'8% degli infetti che richiedono una terapia intensiva (e, allo stesso modo la percentuale del 7% per la letalità del virus) è calcolata avendo a denominatore solo gli infetti ai quali è stato effettuato un tampone non considerando quindi gli infetti asintomatici e coloro che avendo sintomi lievi non si sottopongono al tampone. Tali percentuali rappresenteranno dunque una stima per eccesso della probabilità di quanti, affetti dal virus, necessiteranno la terapia intensiva e, rispettivamente, di coloro che avranno un esito letale.

4. Modelli statistici e previsioni della diffusione del Coronavirus**[9]**

4.1. Verifica dell'efficacia della manovra di lockdown

Se non possiamo avere stime affidabili di queste grandezze, cosa si può chiedere dunque alla Statistica in questa fase delicata?

Un obiettivo è certamente quello di suggerire e testare modelli interpretativi basati sui dati i quali riescano ad individuare regolarità nascoste e ci aiutino a prevedere, con un certo grado di

probabilità, possibili sviluppi futuri.

Un esempio molto semplice che non richiede particolari conoscenze statistiche aiuterà a comprendere cosa intendiamo. Siamo certamente tutti molto interessati a sapere se si stiano rivelando efficaci le misure di quarantena forzata imposte dal Governo l'8 marzo per contenere la diffusione del Coronavirus (il cosidetto *lockdown*).

Se osserviamo su un grafico l'andamento del numero dei contagi prima dell'introduzione di tali misure, esso segue un andamento quasi perfettamente esponenziale (quindi lineare su una scala logaritmica). Questo andamento è riportato in rosso nella Figura 1 su scala logaritmica. Se estrapoliamo tale andamento dopo l'8 marzo abbiamo un'idea di quello che sarebbe stato il verosimile sviluppo dell'epidemia in assenza del provvedimento di *lockdown*, che denominiamo il *controfattuale.* Confrontando tale andamento teorico con i dati effettivamente osservati (riportati in verde nella Figura 1), quello che si osserva è un marcato e crescente distacco dei soggetti positivi realmente osservati dall'andamento lineare precedente rivelando, dunque, l'efficacia della misura per contenere il diffondersi del contagio.

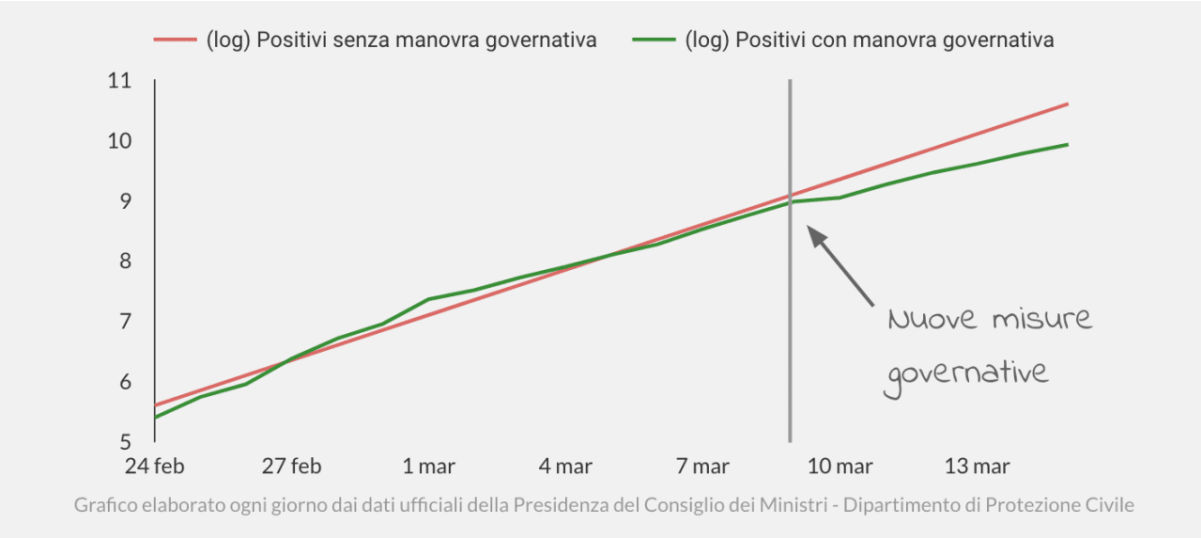


Figura 1. Andamento temporale dei positivi (su scala logaritmica) prima e dopo le misure di isolamento sociale. (Grafico elaborato sui dati ufficiali della Presidenza del Consiglio dei Ministri – Dipartimento di Protezione Civile).

4.2. Stima del momento di "picco epidemico"

Utilizzando modelli molto più sofisticati di quello precedentemente riportato a titolo esemplificativo (e senza entrare in aspetti più tecnici) si possono costruire modelli al fine di tentare di prevedere il verosimile andamento dell'epidemia.

Un'informazione estremamente rilevante in tal senso che tutti vorremmo avere è la data del cosiddetto *picco epidemico*. Con esso si intende, infatti, il punto massimo nella curva della crescita temporale del numero di infetti dove tale numero inizia a decrescere e, per così dire, si inizia a vedere la fine dell'emergenza.

Le nostre elaborazioni condotte a riguardo conducono al grafico riportato nella Figura 2 che mostra come tale picco, originalmente previsto dal nostro modello per il 30 marzo, si è oggi spostato in avanti intorno al 15 aprile (intervallo tra l'11 ed il 17 aprile) come esito delle misure di isolamento sociale introdotte l'8 di marzo.

Questa deve intendersi come una buona notizia: più il picco si sposta in avanti nel tempo, infatti, più si riduce il punto di massimo (come mostra il grafico) consentendo al sistema sanitario di fare fronte all'emergenza in maniera adeguata. Le nostre stime si allineano sostanzialmente a quelle ottenute attraverso l'uso di altri modelli da altri studiosi che rese pubbliche in questi giorni (si vedano, ad esempio, i siti web **https://statgroup-19.blogspot.com/** e

https://www.ilsussidiario.net/news/i-numeri-del-coronavirus-nuove-previsioni-e-scenario-ecco-quando-sara-il-picco/1996867/).

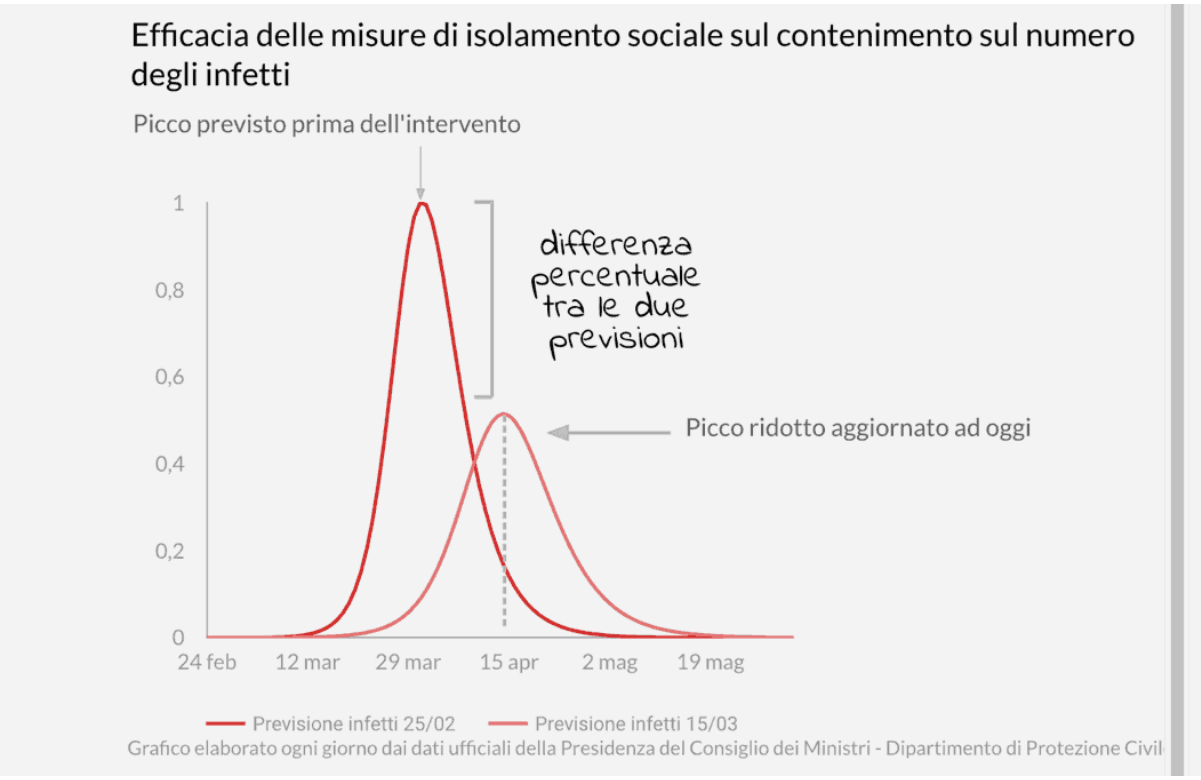


Figura 2. Curva dell'evoluzione stimata dell'epidemia nel tempo prima e dopo le misure di isolamento sociale (Grafico elaborato sui dati ufficiali della Presidenza del Consiglio dei Ministri – Dipartimento di Protezione Civile).

Tale data non va confusa con la data del picco del numero di "nuovi casi" giornalieri, che è la definizione alla quale fanno riferimento le stime del Governo, collocandola intorno al 18 marzo e che pure rappresenta un punto di svolta importante (si veda a riguardo **https://www.ilsole24ore.com/art/coronavirus-governo-stima-92mila-contagi-picco-18-marzo-ADfgS9C**)

4.3. Stima del "tasso netto di riproduzione"

Per finire, un ulteriore parametro interessante per leggere l'evoluzione del Coronavirus in Italia è il cosiddetto "tasso netto di riproduzione" (indicato con il simbolo R_0) del quale si sente spesso parlare in questi giorni. Tale indice rappresenta il numero medio di contagiati da ciascun individuo positivo nel periodo in cui è infetto. Più è alto il valore di R_0, più difficile sarà controllare l'epidemia. Se R_0 è inferiore ad 1 l'infezione tenderà ad estinguersi, mentre se R_0 è maggiore di 1 essa continuerà a diffondersi nella popolazione. Di qui l'importanza di calcolarne

il valore durante un'epidemia e di prevederne gli sviluppi futuri. Il grafico riportato nella Figura 3 mostra l'andamento di tale parametro stimato in base al nostro modello. Da un valore superiore a 2 (al 25 febbraio ogni infetto ne contagiava in media 2.04) si è passati ad oggi ad un tasso di poco superiore ad 1 ed in costante diminuzione.

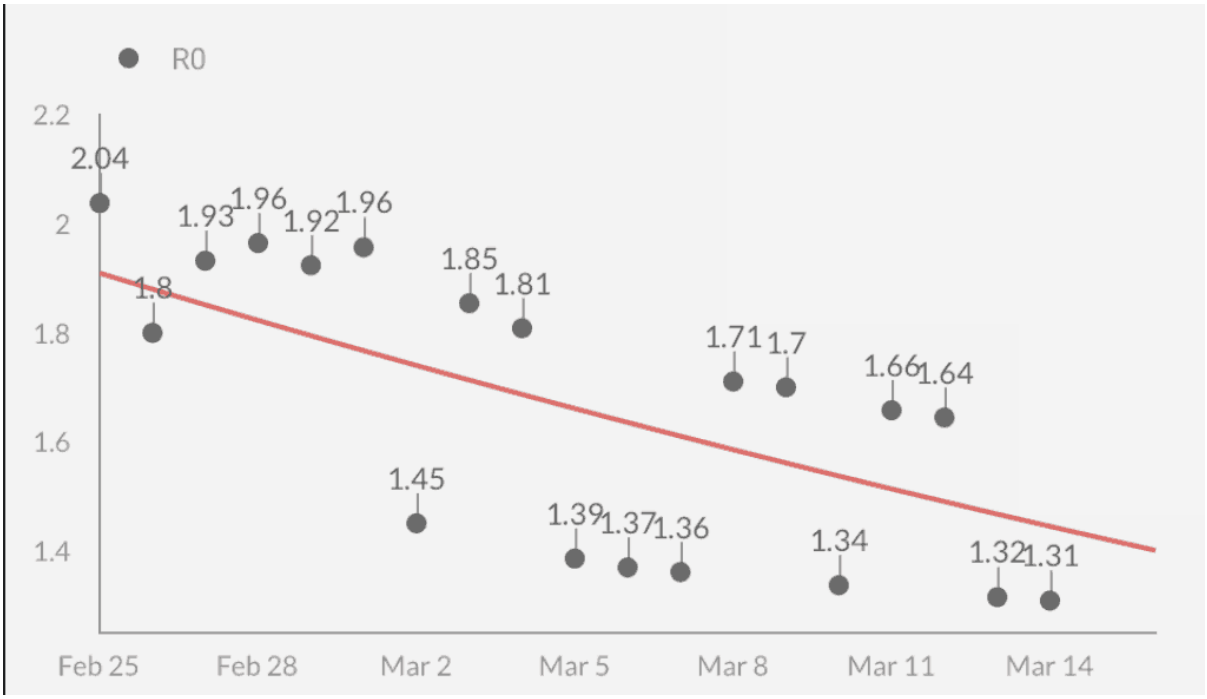


Figura 3. Curva dell'evoluzione stimata del tasso netto di riproduzione (Grafico elaborato sui dati ufficiali della Presidenza del Consiglio dei Ministri – Dipartimento di Protezione Civile).

5. Conclusioni

La Statistica ci mette in guardia circa facili generalizzazioni di quanto osserviamo. I dati da soli non dicono molto, e non dobbiamo cercare di far dire loro più di quanto possano dire. Anche nel mondo attuale, dove il diluvio di dati al quale siamo sottoposti sembra secondo alcuni aver reso obsoleto il metodo scientifico, al fine di impostare correttamente una ricerca, i dati vanno elaborati tramite rigorosi modelli al fine tradurli in informazioni utili, far crescere la conoscenza e condurci a decisioni migliori per l'utilità di tutti.

"Torneremo ad abbracciarci" ha detto il Presidente del Consiglio Giuseppe Conte nel suo discorso dell'8 marzo. «Ma fra quanto? E in quanti di meno?» Si chiedeva Carlo Verdelli nell'editoriale di Repubblica del 12 marzo. Lo sforzo di molti studiosi in Italia e nel mondo va nella direzione di provare a rispondere attraverso l'uso di modelli statistici a queste e ad altre domande che ci tengono con il fiato sospeso in questi giorni.

[1] "With enough data, the numbers speak for themselves" Chris Anderson, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, Wired Magazine, Science, 23.3.2008

[2] Sul tema dei Big Data il lettore interessato può riferirsi a Arbia, G. (2018) "Statistica, società e nuovo empirismo nell'era dei Big Data", Edizioni Nuova Cultura, Roma e Arbia, G. (2018) "Big Data: il sesto potere", La Rivista Vita e Pensiero, Milano.

[3] L'ambiguità è stata in parte corretta ai nostri giorni con l'uso sempre più diffuso del termine *Data Science* (scienza dei dati).

[4] Apparso dapprima nel 1941 nella raccolta *Il giardino dei sentieri che si biforcano* e poi nel 1944 all'interno del volume *Finzioni*.

[5] Per gli amanti del calcolo combinatorio esso è pari a 30 (il numero dei caratteri disponibili) elevato ad 1,000,000, ma se si prova a far effettuare questo calcolo ad un qualsiasi calcolatore la risposta che si ottiene è univocamente: infinito.

[6] Fisher, R. (1935) **"***The Design of Experiments***"**, Hafner Publishing Company, New York

[7] Poincarè J.-H. (1997) **"***Scienza e metodo***"**, Einaudi.

[8] Si veda a riguardo Arbia, G. (2020) "A Note on Early Epidemiological Analysis of Coronavirus Disease 2019 Outbreak using Crowdsourced Data », ***arXiv:2003.06207v1***.

[9] Le analisi qui riportate sono state elaborate dagli autori e fanno parte di un lavoro congiunto con Andrea Palladino e Luigi Atzeni che si ringraziano. Il modello utilizzato nelle nostre elaborazioni è basato sui lavori di Hamer W. H. (1906) **"**Epidemic diseases in England**",** *Lancet*, 1, 733-9 e Soper, H. E. (1929) "Interpretation of periodicity in disease prevalence, " *Journal of the Royal Statistical Society*, A, 92, 34-73. Si rimanda al sito web **https://vincnardelli.github.io/covid19-italia/** per ulteriori analisi ed approfondimenti.