Un semplice modello a due parametri per la diffusione del COVID-19 in Italia

Fabio Musmeci 24/3/2020 ore 16:00

Riassunto

Viene proposto un modello statistico molto semplice per rispondere a domande circa l'andamento del COVID-19. Secondo questo modello i nuovi contagi dovrebbero essere pressoché zero prima del 4 aprile con probabilità del 90%. I casi attualmente positivi al 3 aprile sarebbero tra i 30.000 e i 90.000 con una probabilità del 90%.

Premessa

Quando terminerà la diffusione del COVID-19? Quanti nuovi casi avremo?

Lo studio è effettuato per tentare di dare una risposta semplice a domande che si pongono a fronte di un problema complesso come quella della diffusione di un virus. Purtroppo, mentre se da un lato la semplicità aiuta la comprensione del modello adottato e delle sue ipotesi di base, non tutti gli aspetti del problema possono essere inclusi e, proprio per la complessità del fenomeno, ci si aspettano comportamenti difficilmente prevedibili.

D'altronde una sorta di previsione viene comunque effettuata anche da parte di chi non lo dice esplicitamente o che non utilizza strumenti matematici. Utilizzando strumenti matematici le assunzioni sono chiare e le conclusioni sono logica conseguenza delle assunzioni. Di seguito viene tentata una previsione per rispondere alle citate domande in termini probabilistici. Cioè si dirà il valore previsto è in questo intervallo con probabilità xx. Si è scelto di utilizzare il numero minimo di parametri per la descrizione del problema.

Materiali e metodi

Esistono vari modelli a disposizione per affrontare tematiche epidemiologiche. Il più diffuso è il SIR[1] (Sane, Infette, Recuperate). Questo modello si basa sull'effetto di contenimento dovuto al fatto che i soggetti recuperati divengono immuni, cosa non provata nel caso del COVID-19. Sue modifiche sono state adottate come, per esempio, quelle presentate sul sito del Meteo[2]. A parere di chi scrive il cuore del problema è riuscire a modellare la risposta sociale e come questa possa influenzare la diffusione della malattia.

È stato qui ipotizzato che il numero di contagi sia proporzionale al numero di "*attualmente positivi*" del giorno precedente. Il coefficiente di proporzionalità è linearmente decrescente nel tempo a descrivere la risposta sociale.

Questo permette di prevedere il numero di contagi dato quello del giorno precedente e il valore del coefficiente di proporzionalità.

Il primo giorno (24 febbraio 2020) il numero di attualmente positivi è assunto pari a quello osservato, successivamente dal giorno n.2 (25 febbraio)

[1] https://it.wikipedia.org/wiki/Immunit%C3%A0_di_gregge

[2] https://www.ilmeteo.it/notizie/coronavirus-italia-ecco-le-curve-di-previsione-matematica-aggiornate-ad-1-mese-con-contagi-picco-guariti-morti

Per ogni giorno G (G=2,…n)

Coefficiente = b X G + a

[Nuovi contagi al giorno G] = Coefficiente X [Attualmente positivi al giorno G-1]

[Attualmente positivi al giorno G] = [Attualmente positivi al giorno G-1] + [Nuovi contagi al giorno G]

Viene utilizzato, come dato di base, uno dei data base, "*open access*" disponibili su WEB[3]. Questo data base verrà aggiornato nuovamente il 25 marzo 2020 e il dato del 23 marzo è stato aggiunto da quelli disponibili dalla conferenza stampa della protezione civile.

Dati i coefficienti a e b della retta che stima l'andamento del coefficiente è possibile stimare il giorno G0 quando il coefficiente di proporzionalità sarà zero:

$$G0= -a/b$$

I dati utilizzati sono presentati nella tabella seguente.

Per la gestione dell'incertezza è stato utilizzato il metodo di "*bootstrap*"[4]. Questo metodo si basa sul riutilizzo dei dati stessi per riprodurre molte volte il campione e effettuare nuovamente le stime. Così dello stesso parametro, per esempio una media, avremo molti possibili valori è potremo dire, per esempio che la media, entro il 90% di probabilità è tra i valori x_1 e x_2. In questo lavoro si sono utilizzate 10.000 copie di bootstrap dei dati rappresentati in tabella 1.

Per organizzare e presentare i dati è stato utilizzato Excel mentre per il metodo di bootstrap è stato scritto un apposito programma in C# che restituisce dati utilizzabili in Excel.

Si noti che nel modello gli attualmente positivi rimangono tali. L'unica uscita da questo stato è dato da un'eventuale assunzione di valori negativi del coefficiente di proporzionalità dopo il raggiungimento dello zero.

[3] https://github.com/pcm-dpc/COVID-19/blob/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv
[4] https://it.wikipedia.org/wiki/Bootstrap_(statistica)

G	data	totale attualmente positivi	nuovi attualmente positivi
1	24/02	221	
2	25/02	311	90
3	26/02	385	74
4	27/02	588	203
5	28/02	821	233
6	29/02	1049	228
7	01/03	1577	528
8	02/03	1835	258
9	03/03	2263	428
10	04/03	2706	443
11	05/03	3296	590
12	06/03	3916	620
13	07/03	5061	1145
14	08/03	6387	1326
15	09/03	7985	1598
16	10/03	8514	529
17	11/03	10590	2076
18	12/03	12839	2249
19	13/03	14955	2116
20	14/03	17750	2795
21	15/03	20603	2853
22	16/03	23073	2470
23	17/03	26062	2989
24	18/03	28710	2648
25	19/03	33190	4480
26	20/03	37860	4670
27	21/03	42681	4821
28	22/03	46638	3957
29	23/03	50818	4790

Tabella 1: Dati utilizzati

Coefficiente di proporzionalità

La regressione tra i dati di tabella 1 (colonna G e *nuovi attualmente positivi*) ci porta a stimare i coefficienti come b=-0.0107 e a=0.3854 (Con R^2=0.56[5]) ossia

[Nuovi contagi al giorno G] = [-0.0107 X G + 0.03854] X [Attualmente positivi al giorno G-1]

Si noti che il b negativo comporta il fatto che il contagio diminuisca percentualmente all'aumentare dei giorni.

Si noti inoltre che il modello si basa solo sui due parametri a e b e del valore iniziale degli attualmente positivi (nel nostro caso 221 per G=1).

In figura 1 è mostrata la retta di correlazione insieme ai dati osservati e alle fasce di confidenza ottenute con il metodo di bootstrap effettuando 10.000 regressioni lineari.

Utilizzando solo questi due parametri e il valore iniziale di 221 si ottiene l'andamento degli "*nuovi attualmente positivi*" in funzione del tempo G. In figura 2 è rappresentato l'andamento, fino alla fine del mese di marzo, insieme a quello dei dati realmente osservati.

Utilizzando questo andamento è possibile anche visualizzare (figura 3) i casi "*attualmente positivi*" in funzione del tempo.

Il valore G0 di azzeramento dei contagi è:

G0 = -0.3854/-0.0107 = 36

Ossia, dato il sistema di numerazione dei giorni a partire dal 24/2/2020, G0 corrisponde al 29/3/2020, data nel quale i contagi, secondo questo modello, saranno zero.

I dati di figura 3 sono basati, come detto, sul valore 221 registrato al giorno 1.

Gestione dell'incertezza

L'insieme delle 10.000 soluzioni offerte dal bootstrap portano a diverse possibili evoluzioni dei dati.

Se si utilizza, come dato di partenza, il valore ultimo registrato al 23 marzo, invece dei 221 di inizio periodo, si ottiene la figura 4 con l'andamento dei nuovi positivi a partire dal 23 marzo 2020.

La stima effettuata precedentemente per G0 (il giorno di zero nuovi contagi) piò essere ripetuto per ognuna delle 10.000 regressioni effettuate con il bootstrap. L'insieme delle 10.000 soluzioni può essere rappresentato come una densità di probabilità, che lo zero sia in un dato giorno, in funzione dei giorni.

La figura 5 rappresenta questa probabilità. Il giorno di maggior probabilità è quello precedentemente stimato del 29 marzo con una probabilità però del 14%. I giorni prima e dopo hanno anch'essi una probabilità del 13% di essere il giorno di zero contagi. Sommando le probabilità si ottiene una probabilità cumulativa, funzione del giorno, che ci indica la probabilità che lo zero contagio sia prima o uguale al giorno G in ascissa (figura 6).

Conclusioni

Secondo questo modello i nuovi contagi dovrebbero essere pressoché zero prima del 3 aprile con probabilità del 90%. I casi attualmente positivi al 3 aprile sarebbero tra i 30.000 e i 60.000 con una probabilità del 90%.

[5] R^2 è il coefficiente di correlazione che vale 1 quando la correlazione è perfetta e vale 0 in assenza di correlazione.

Figure

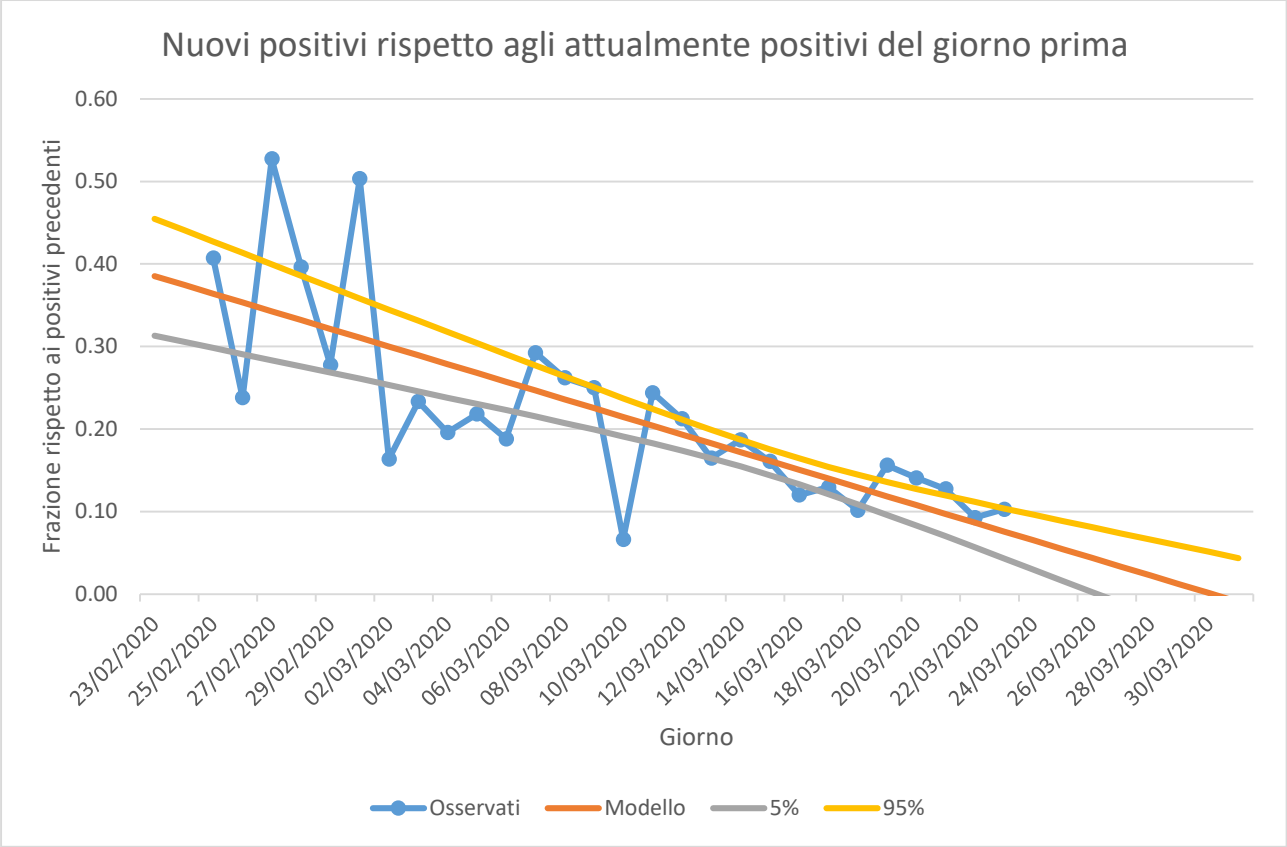

Figura 1: Andamento dei nuovi contagi rispetto ai positivi del giorno precedente, le due fasce 5% e 95% delimitano l'area entro la quale dovrebbe giacere la retta di regressione con il 90% di probabilità

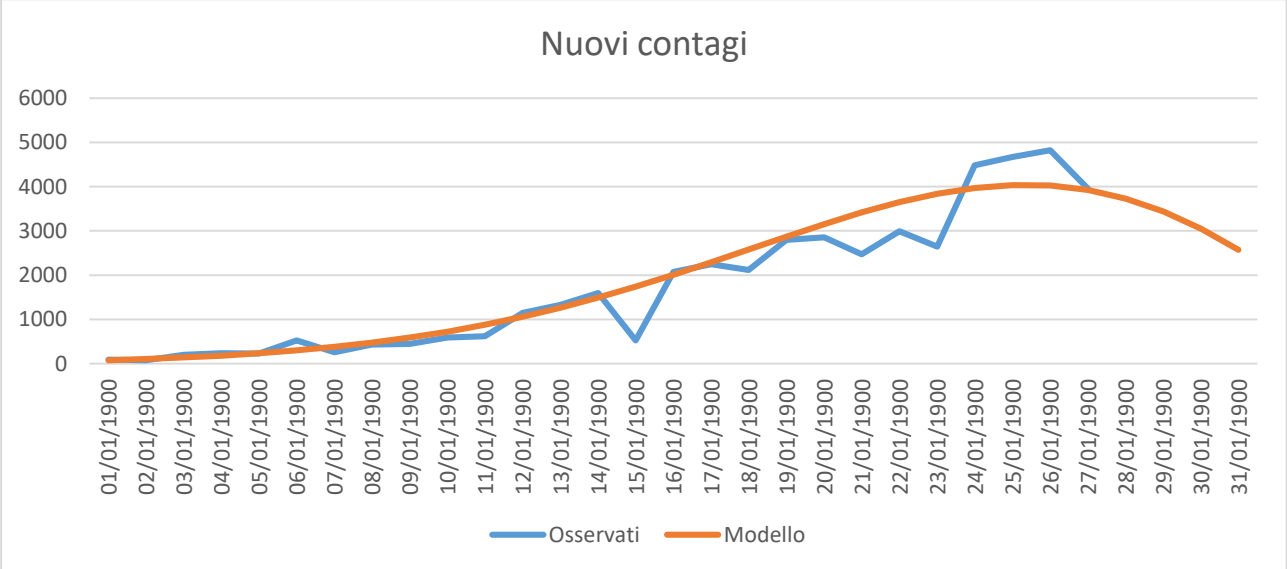

Figura 2: andamento dei casi di "*nuovi attualmente positivi*" osservati e calcolati a partire dal giorno G=2 e utilizzando solo il dato iniziale al giorno G=1 ossia 221 casi. Una migliore proiezione per il futuro è ottenibile ripartendo dall'ultimo dato osservato. Questa figura è fornita solo per illustrare graficamente la capacità del modello di *spiegare* i dati osservati anche utilizzando il solo dato iniziale.

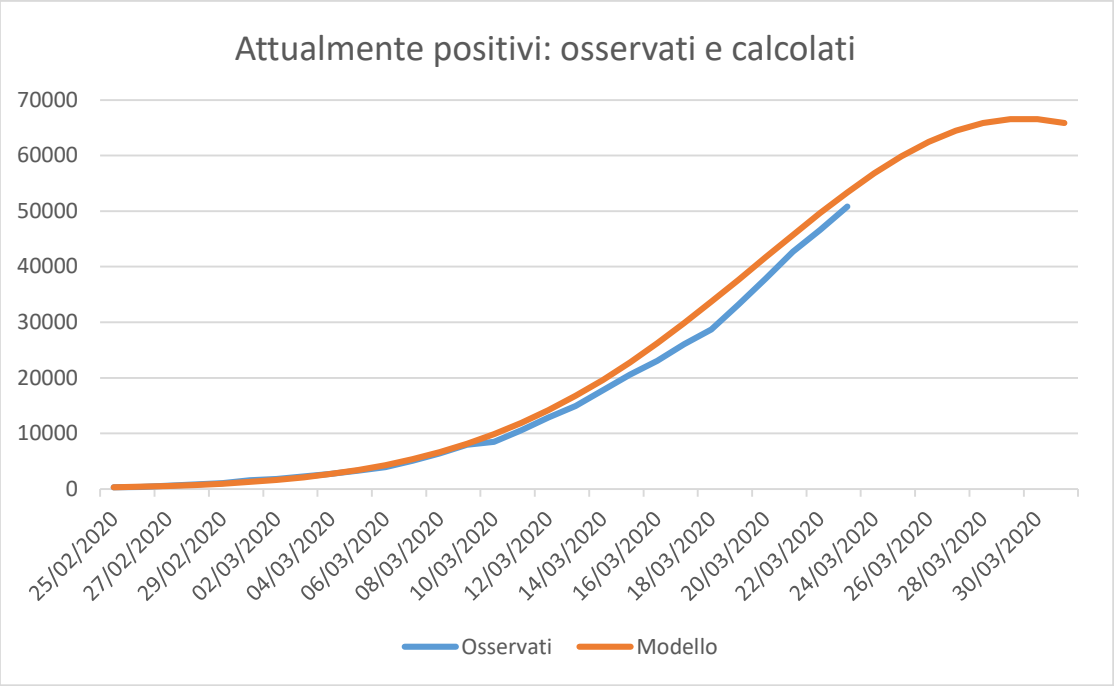

Figura 3: andamento dei casi di "attualmente positivi" osservati e calcolati a partire dal giorno G=2 e utilizzando solo il dato iniziale al giorno G=1 ossia 221 casi. Una migliore proiezione per il futuro è ottenibile ripartendo dall'ultimo dato osservato (figura successiva)

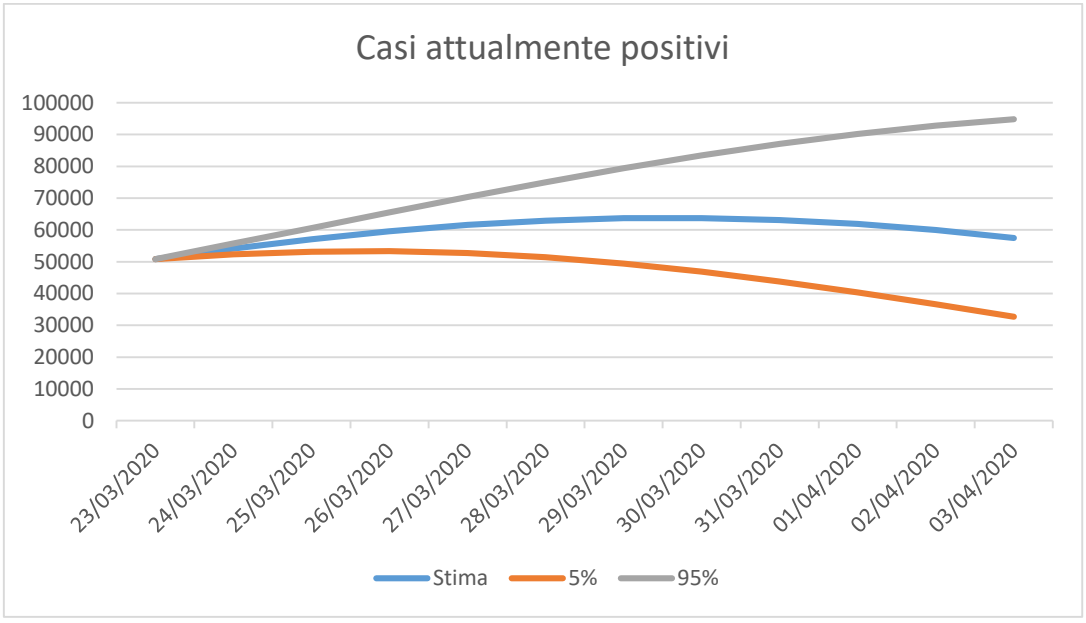

Figura 4: andamento dei casi di "attualmente positivi" osservati e calcolati a partire dal giorno G=30 (24 marzo) e utilizzando come dato iniziale il giorno G=29 (dato del 23/3) ossia 50818 casi. Insieme alla stima vi sono i limiti del 5% e 95% di probabilità. Nel caso peggiore si potrebbero avere, al 3 aprile, 93.000 casi di attualmente positivi, nel caso migliore, passato lo zero, gli attualmente positivi calano a poco più di 30.000

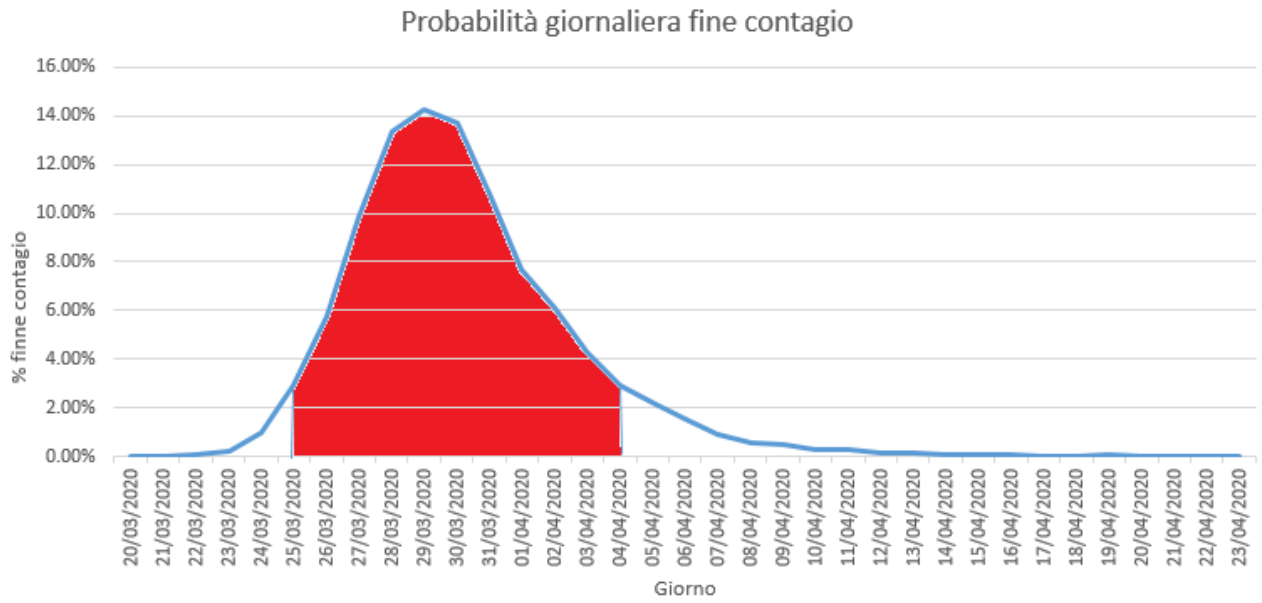

Figura 5: andamento dei nuovi contagi nei vari giorni. La probabilità che lo zero sia tra il 25/3 e il 4/4 è del 92%

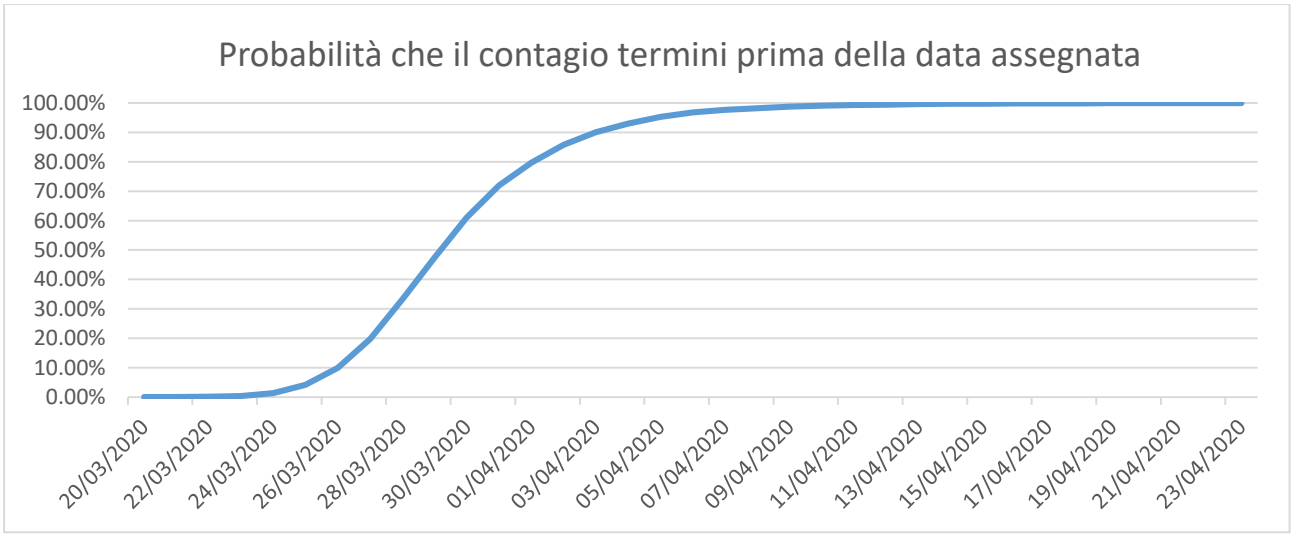

Figura 6: Probabilità il contagio termini prima della data assegnata, il 3 aprile viene superato il 90%