Calcolo dell'R_t per l'epidemia di COVID 19 in Italia

Massimo Cipressi

17/2/2023

1 Abstract

Lo scopo di quest'analisi è il calcolo dell'indice R_t per l'epidemia di COVID 19 in Italia, dalla sua diffusione nel febbraio 2020 a oggi.
Sono presentati due modelli di calcolo. Il primo si basa sul modello SIR: è di facile implementazione, richiede pochi dati, ma è più approssimativo. Risulta, infatti, avere errori anche di $\pm 0,2$ e in ritardo di circa una settimana rispetto ai dati ufficiali.
Il secondo modello si basa su una simulazione Monte Carlo; ha un errore dell'ordine di 10^{-2} e risulta più simile alla misura ufficiale. Tuttavia, richiede più dati e la misura, per un giorno fissato, varia con le misure fatte nei giorni successivi: diventando precisa solo dopo circa due settimane.

2 Calcolo basato su SIR

Il SIR è un modello epidemiologico compartimentale con tre compartimenti: i suscettibili (S), cioè coloro che non hanno mai contratto la malattia, gli infetti (I); e i rimossi (r) in cui ci sono i guariti e i morti. Non si considera demografia quindi: $N = S + I + r$ è costante.
Le equazioni differenziali che definiscono il modello sono:

$$\begin{aligned} \dot{S} &= -\beta SI\frac{c}{N} \\ \dot{I} &= \beta\frac{c}{N}SI - \gamma I \\ \dot{r} &= \gamma I \end{aligned} \tag{1}$$

β, γ e c sono assunte costanti, rispettivamente: la contagiosità, il tasso di rimozione (queste due sono caratteristiche della malattia) e il numero di incontri di una persona nell'unità di tempo (che per noi sarà sempre di un giorno).
Supponendo $\frac{S}{N} \approx 1$ (sicuramente accettabile quantomeno a inizio epidemia), si ha che βc è il tasso d'infezione, cioè quante persone si infettano nell'unità di tempo. Allora $T_c := \frac{1}{\beta c}$ è il tempo di contagio. Similmente il tempo di rimozione è $T_R := \frac{1}{\gamma}$.
Definiamo il numero di riproduzione di base (R) come il numero di infezioni secondarie che un infetto genera nel corso della sua malattia; allora:

$$R = \frac{T_R}{T_C} = \frac{\beta c}{\gamma} \tag{2}$$

Da cui si ha:

$$\frac{\dot{I}}{I} = \gamma(R - 1) \tag{3}$$

Da questo modello, si ha, dunque, che R non dipende dal tempo. Noi però, fissato γ, vogliamo usare la precedente equazione per ricavare l'R_t che generi, di giorno in giorno, i dati che abbiamo sugli infetti. L'idea più semplice è che, in realtà, sia il tempo di contagio a variare nel tempo. Per questo indice è rilevante la soglia di 1, in quanto $R > 1$ implica che gli infetti sono in crescita.
Per quanto riguarda γ, il dato pubblicato più di recente (luglio 2021)[1] per l'Italia (il valore cambia decisamente in altri paesi) è $\frac{1}{\gamma} = T_R = 9,7 \pm 2$ giorni; sarà questo la fonte della stima dell'errore sul calcolo di R_t eseguito in questo modo.
In questo caso, si considerano infette le persone positive al tampone; da notare che considerare solo il tampone molecolare o anche "rapido" o "fai da te" è una scelta che è cambiata nel corso dell'epidemia.

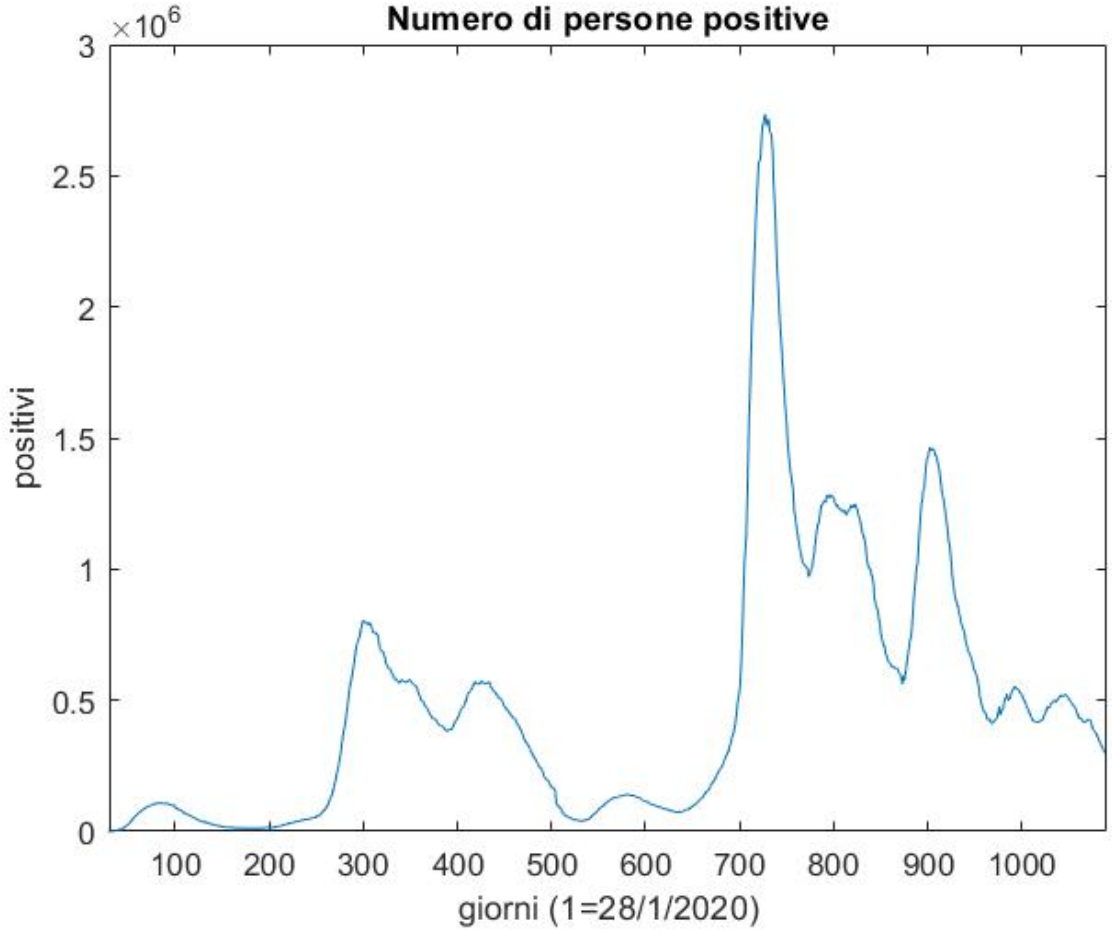


Figura 1: Il numero di positivi (dati Protezione Civile)

[1] "The SIR model towards the data" di Lazzizzera, 25/7/2021 arXiv:2106.01602v2

Riportiamo quindi il risultato per R_t calcolato in media mobile: il valore di un giorno è la media con quello dei 6 giorni precedenti. Ad esso sovrapponiamo il dato emesso ufficialmente (fonte INFN).

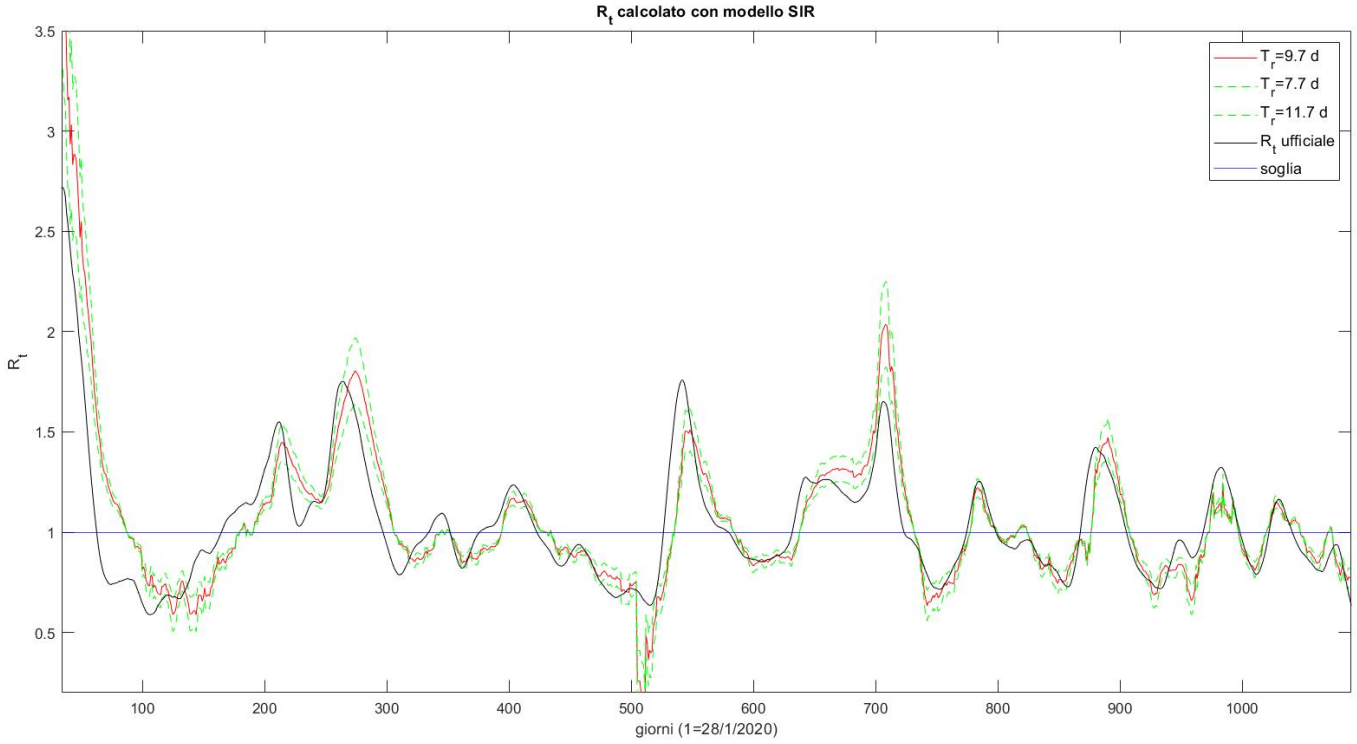


Figura 2

L'andamento qualitativo è già abbastanza corretto, ma il dato è rumoroso e notiamo un certo ritardo rispetto al dato ufficiale (circa una settimana). L'errore è abbastanza variabile, risulta massimo nei punti stazionari con picchi di circa ± 0.2 (escluse le prime settimane).

3 Calcolo con simulazione Monte Carlo

Questo è il metodo che è stato usato dall'ISS per calcolare il dato emesso ufficialmente.
Una prima differenza col metodo precedente è che, qui, come infetti, non si usano i positivi, ma i sintomatici: cioè persone che sono state seguite durante la malattia per saperne la data di inizio e fine sintomi. Questo è sicuramente un campione più piccolo (di 2 ordini di grandezza solitamente), ma è più identificativo del periodo in cui una persona è in grado di generare infezioni secondarie; la politica di raccolta dati è, inoltre, rimasta la stessa durante tutta l'epidemia. Da notare che l'aggiornamento in una data influenza solitamente anche le date precedenti (quando una persona comunica di essere già sintomatica da alcuni giorni); allora anche la stima di R_t per un giorno varierà coi dati dei giorni seguenti.

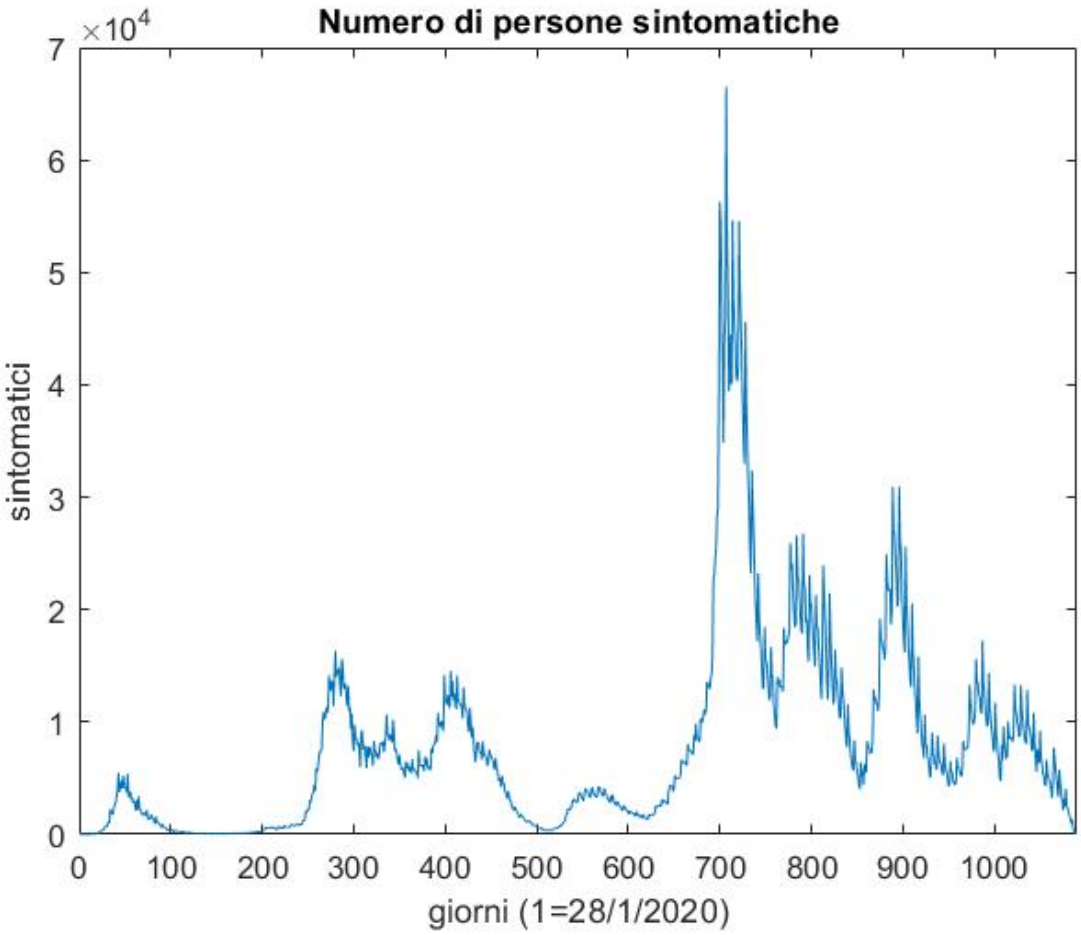


Figura 3: Il numero di sintomatici (dati ISS)

Definiamo ϕ la distribuzione dei tempi di generazione, cioè $\phi(s)$ è la densità di probabilità che passino s giorni fra un'infezione primaria e una secondaria. I dati più recenti su di essa sono del dicembre 2020[2]; si usa una distribuzione gamma:

$$\phi(s) = \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s} \tag{4}$$

con $\alpha = 1,87$, $\beta = 0,28$.
Ne riportiamo di seguito il grafico a cui è sovrapposta la discretizzazione sui giorni tagliata a 25 giorni: essendo, successivamente, il contributo inferiore a $2 * 10^{-3}$ (la discretizzazione è stata realizzata integrando con la routine quad di matlab).

[2]"Stime della trasmissibilità di SARS-CoV-2 in Italia" di Guzzetta, Merler 7/12/2020

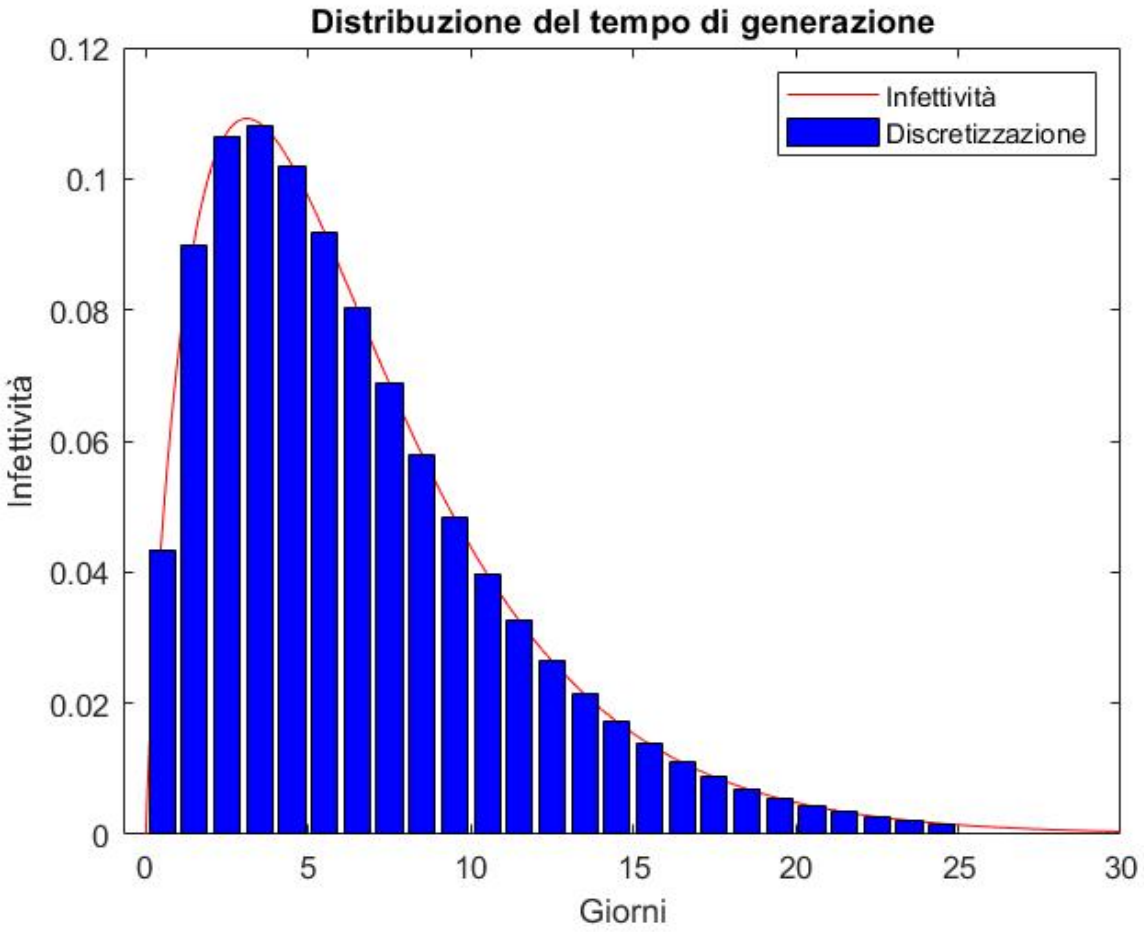


Figura 4: La barra ad ascissa 0.5 rappresenta ad es. l'integrale fra 0 e 1

Allora il valore atteso per il numero di sintomatici al giorno t è:

$$C(t) = R_t \sum_{s=1}^{25} \phi(s) C(t-s) \tag{5}$$

La generazione dei nuovi contagi può essere vista come processo poissoniano; allora usiamo $P(k,\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$ (dove k sarà il nostro dato, mentre λ la stima proveniente dalla precedente equazione) come funzione di verosimiglianza per il giorno t:

$$L = P(C(t), R_t \sum_{s=1}^{25} \phi(s) C(t-s)) \tag{6}$$

Lo scopo della simulazione Monte Carlo sarà trovare, di giorno in giorno, il valore di R_t che massimizza la funzione di verosimiglianza.
Possiamo usare un algoritmo simile a Metropolis.
Il passo di evoluzione temporale inizia con una proposta che ha distribuzione uniforme, attorno al valore precedente di R_t, con ampiezza di 1,5. La proposta è poi accettata con probabilità:

$$\frac{L(R_t(i))}{L(R_t(i-1))} \tag{7}$$

Cosicché la proposta venga accettata con certezza se fa aumentare la verosimiglianza. Di seguito è riportato il risultato dopo che è stata applicata la media mobile sulla settimana (come già discusso).

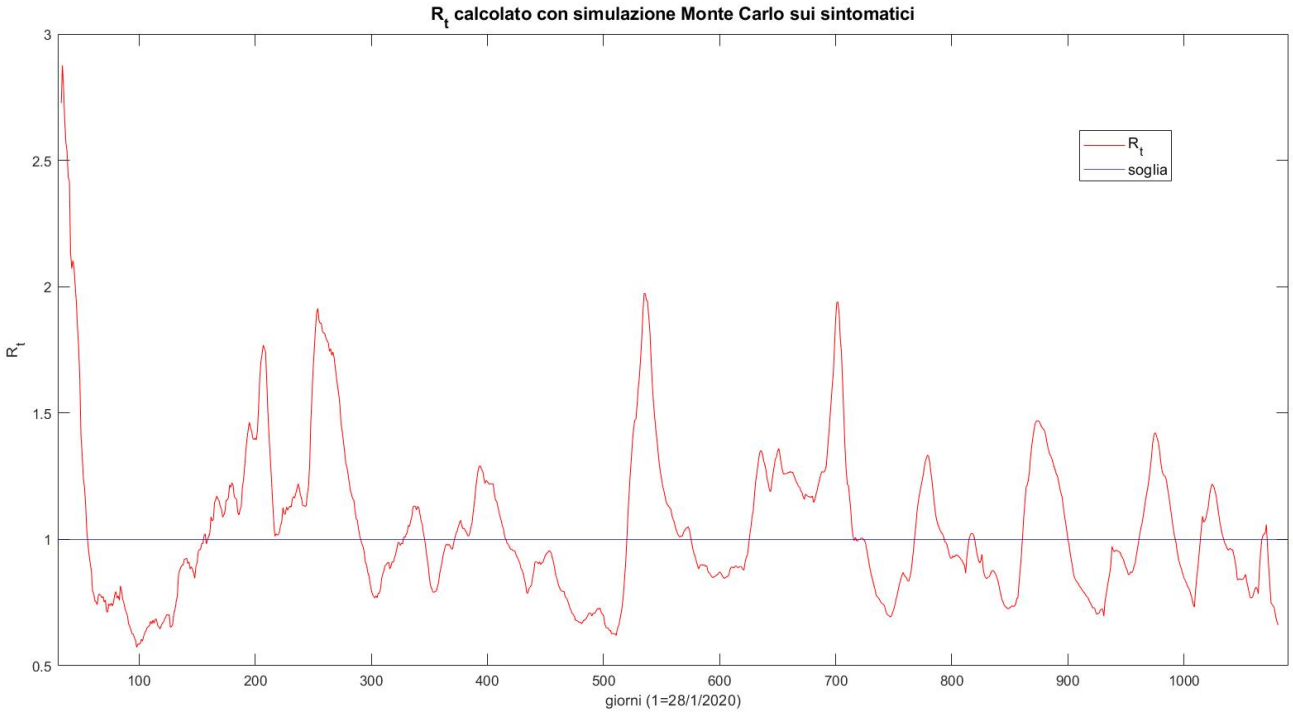


Figura 5

I dati non sembrano essere soggetti ad autocorrelazione: test a diversi valori di R_t hanno mostrato la funzione di autocorrelazione andare a 0 dopo un passo di evoluzione temporale. Possiamo aspettarcelo agendo l'algoritmo su un unico valore scalare; ed essendo la proposta random fatta su un intervallo ampio. Risulta (in modo quasi indipendente dal tempo) $\sigma \approx 3 * 10^{-3}$, l'errore non è quindi riportato sul grafico risultando (anche 3σ) difficilmente visibile.

Riportiamo quindi il confronto con il dato ufficiale:

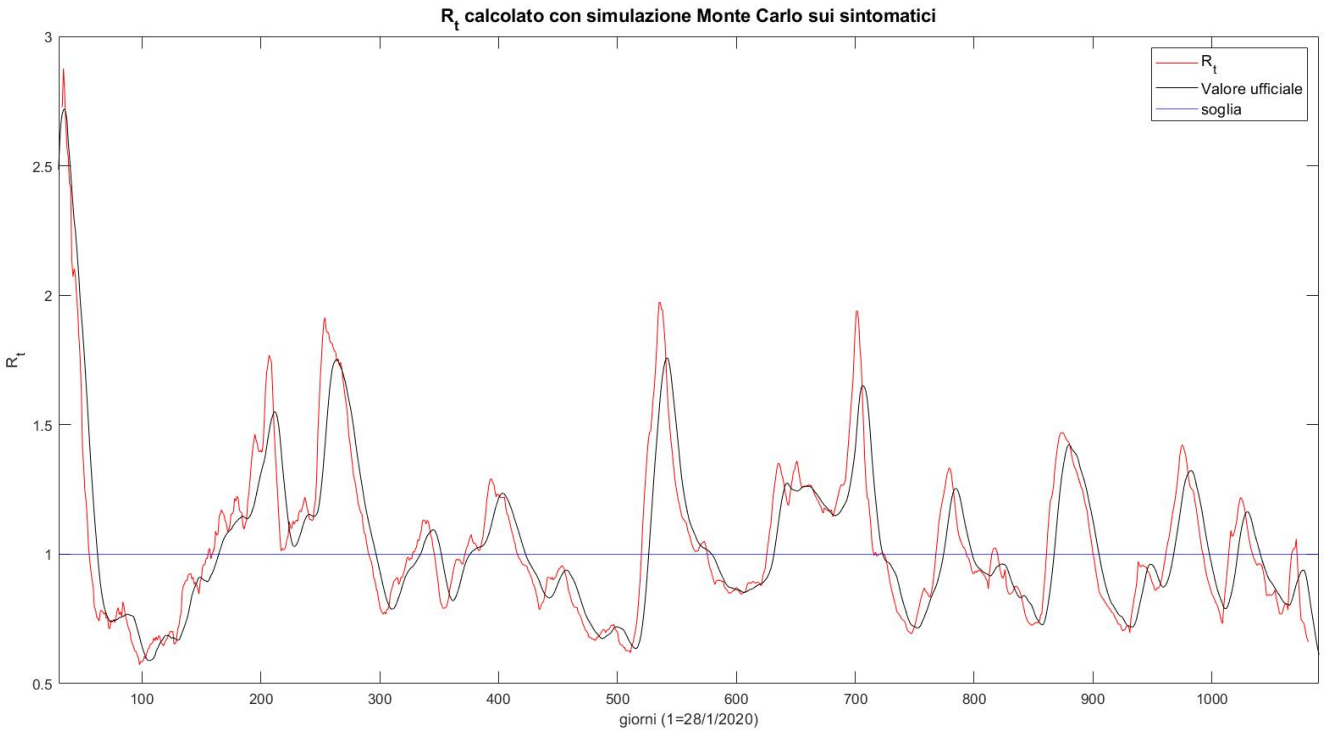
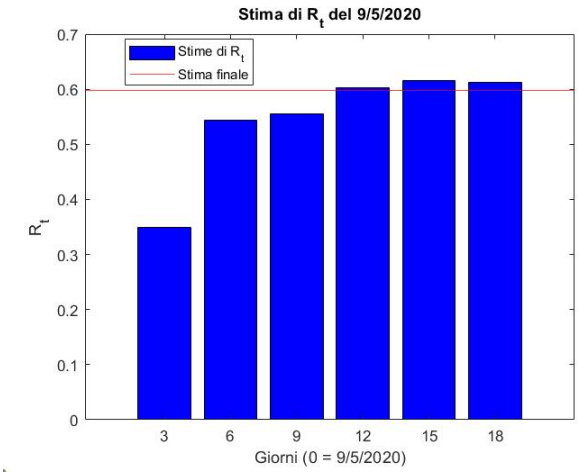


Figura 6

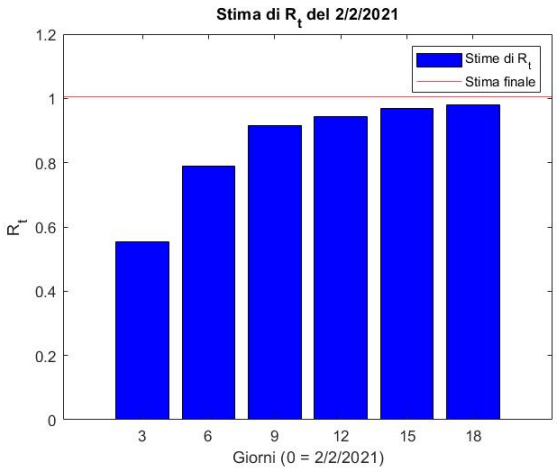
Le misure non sono comunque compatibili essendo l'errore piccolo, ma vediamo che l'andamento è rispettato in modo più fedele rispetto al risultato del metodo precedente.
Si può notare che il risultato della simulazione è in anticipo di circa una settimana rispetto al dato ufficiale.

4 Stabilità di R_t

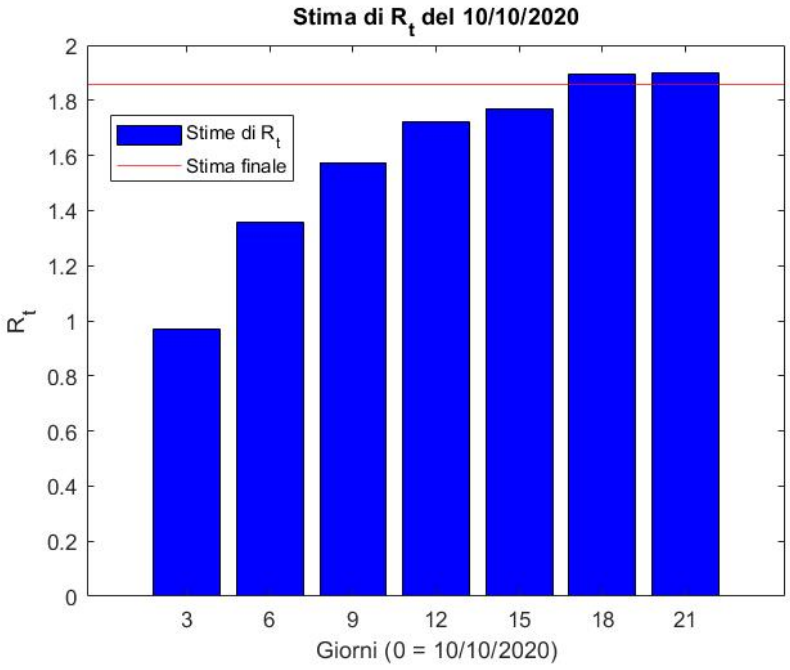
Il fatto che la misura dei sintomatici influenzi sempre anche i giorni passati, fa sì che anche la misura di R_t cambi con nuovi dati. Ci chiediamo dopo quanti giorni il valore di R_t previsto dalla simulazione per un giorno fissato sia affidabile.
Vediamo i risultati per tre giornate con diversi valori finali di R_t:



(a) 9/5/2020 $R_t = 0,60$



(b) 2/2/2021 $R_t = 1,00$



(c) 10/10/2020 $R_t = 1,86$

Figura 7

Per arrivare ad una deviazione inferiore al 5% dal valore finale ci vogliono sempre fra i 12 e i 15 giorni; possiamo però notare che, per R_t elevato, i tempi sono un po' più lunghi. La variazione di R_t è quasi sempre in crescendo: come ci aspetteremmo, essendo dovuta alla registrazione di nuovi sintomatici; discese possono essere dovute a riconteggi o all'errore della simulazione.
Queste due settimane circa sono coerenti con lo shift temporale che si trova fra l'R_t calcolato nei due metodi. Ricordando che il dato ufficiale veniva dato con una settimana di ritardo, capiamo anche perché usando il calcolo col modello SIR si abbia una settimana di ritardo rispetto ad esso, mentre col Monte Carlo una settimana di anticipo.
L'errore ad una settimana dipende più pesantemente dal valore finale di R_t: errore dell'8% con $R_t = 0,60$; errore del 27% con $R_t = 1,86$. Questo giustifica anche che le discrepanze maggiori fra valore ufficiale e simulazione si abbiano sui picchi.