Rt or RDt, that is the question!

R_t or RD_t , that is the question!

Allegato 1
Il numero di riproduzione R_t

Mentre il numero di riproduzione di base R_0 denota il numero atteso di casi secondari derivanti da un caso primario tipico in una popolazione completamente suscettibile, quando l'infezione si diffonde nella popolazione è invece opportuno considerare il numero di riproduzione effettivo R_t, ovvero il numero medio attuale di casi secondari del contagio prodotti da un singolo infettore durante il suo periodo infettivo.
Il monitoraggio dell'andamento di R_t nel tempo consente di rilevare la progressione del contagio e fornisce quindi un feedback sulle misure di contenimento. Valori di R_t inferiori a 1, o meglio ancora prossimi a 0, portano infatti a considerare che l'epidemia si possa considerare sotto controllo.

Il numero di riproduzione R_t è definito nel contesto di modelli del tipo "time-to-infection" e viene stimato sulla base di dati di sorveglianza. Si può ricorrere a diversi metodi, dalla semplice approssimazione con crescita esponenziale dei casi nel tempo ai più complessi quali il fitting di modelli meccanicistici a dati di incidenza o procedure di tipo probabilistico basate sulla ricostruzione degli alberi di trasmissione dell'epidemia. In particolare, i metodi basati sulle serie temporali di incidenza risultano facilmente implementabili e consentono di tenere conto dell'incertezza nella distribuzione degli intervalli tra l'insorgenza dei sintomi nei casi primari e secondari.

Per introdurre brevemente tali metodi è necessario definire preliminarmente gli intervalli temporali su cui si basano. Il tempo di generazione è l'intervallo che intercorre tra l'infezione di un caso primario e quella dei secondari da esso generati. L'intervallo seriale τ (detto anche impropriamente tempo di generazione in luogo del precedente) è il tempo che intercorre tra l'insorgenza dei sintomi in un caso primario e in un caso secondario da esso derivato. E' essenziale rilevare che i dati di sorveglianza riportano normalmente il secondo tipo di intervallo, quando presente, e non il primo.
E' infine definito periodo di incubazione l'intervallo temporale tra l'infezione e l'insorgenza dei sintomi, durante il quale l'individuo non è sintomatico.
La distribuzione degli intervalli seriali viene di solito considerata come indicatrice (proxi) della distribuzione del profilo di infettività per gli individui. Il profilo medio di infettività per gli infettori a seguito dell'infezione è descritto tramite una distribuzione di probabilitá w_S (quindi a somma 1) dipendente dal tempo dall'infezione s, ma non dal tempo di calendario t. Si può considerare che la trasmissione del contagio segua un processo di Poisson; in particolare, il tasso con cui un infetto al tempo t − s genera nuovi infetti al tempo t è dato da $R_t w_s$, per cui l'incidenza dei casi al tempo (time step) t è in media

$$E[I_t] = R_t \sum_{s=1}^{t} I_{t-s}\, w_s$$

dove E denota l'attesa e I_{t-s} l'incidenza al tempo t – s.

La stima di R_t risulta fortemente infuenzata dal profilo di infettività w_S, stimabile tramite la distribuzione dei tempi di generazione che risultano però raramente osservabili mentre lo sono più facilmente i tempi di insorgenza dei sintomi. Nel caso in cui la trasmissione del contagio sia monitorabile, i dati degli intervalli tra l'insorgenza dei sintomi negli infettori e negli infetti possono

essere utilizzati per stimare la distribuzione degli intervalli seriali. Tramite metodi di inferenza statistica Bayesiana risulta quindi possibile ottenere la distribuzione a posteriori di R_t assumendo una distribuzione a priori di tipo Gamma e utilizzando i dati di incidenza sintomatica. Per quanto riguarda la situazione in Italia, l'analisi di dati di cluster dell'epidemia in Lombardia ha consentito di ottenere per gli intervalli seriali un fit tramite una distribuzione Gamma (figura 1) con parametro di forma pari a 1.87 (sd 0.26) e parametro di scala pari a 0.28 (sd 0.04). Tali valori corrispondono a un valore medio di 6.6 giorni e una deviazione standard di 4.88 giorni.

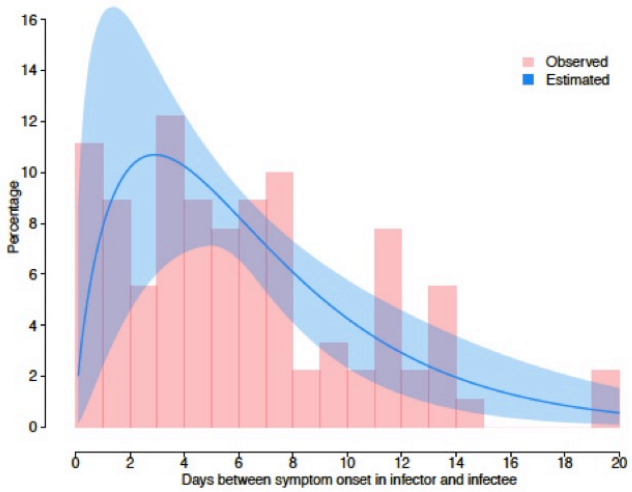


Figura 1

R_t risulta l'unico numero di riproduzione che si possa quindi stimare agevolmente in tempo reale, consentendo quindi di monitorare l'efficienza delle misure di contenimento adottate per l'epidemia e rilevarne eventuali modifiche nell'andamento del contagio.

Si assume che il numero di replicazione R_t rimanga costante per un intervallo di tempo pari a τ e si calcola quindi per ogni t la trasmissibilità media $R_{t,\tau}$ in una finestra temporale di ampiezza τ che termina in t. L'indice $R_{t,\tau}$ risulta altamente variabile e di difficile interpretazione se si considerano finestre temporali troppo piccole e vengono quindi impiegati di solito valori corrispondenti a $\tau = 4$ o $\tau = 7$ giorni. La stima di $R_{t,\tau}$ viene ottenuta a partire da un tempo non inferiore al valore massimo tra τ e l'intervallo seriale medio.

Uno degli strumenti che consentono di stimare i valori di $R_{t,\tau}$ è il software statistico R, tramite il quale si possono ottenere in output i grafici per la serie temporale di incidenza, per i valori di Rt in corrispondenza dell'ultimo giorno della finestra temporale corredati dai rispettivi intervalli di credibilità e per la distribuzione discretizzata degli intervalli seriali. È inoltre anche possibile tenere conto dell'incertezza nella stima della distribuzione Gamma per gli intervalli seriali usando come media μ_{SI} e deviazione standard σ_{SI} realizzazioni estratte da coppie di distribuzioni normali troncate aventi come medie i corrispondenti valori campionari per gli intervalli seriali e range opportuni.

Per ottenere una piena efficacia, il metodo di stima per il numero di riproduzione Rt presentato richiederebbe di poter disporre di un monitoraggio completo dell'incidenza di sintomatici; presuppone inoltre che la distribuzione degli intervalli seriali rimanga costante nel tempo e richiede l'utilizzo di un software dedicato.

Allegato 2
L'indice di Replicazione Diagnostica RD_t

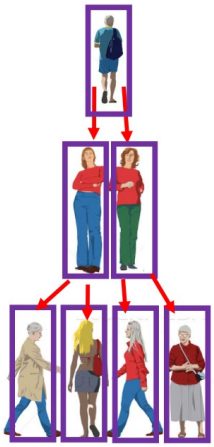
Razionale

Il razionale dell'indice RD_t deriva dalle tecniche della time series analysis che si usano per analizzare gli andamenti delle serie storiche, cioè per l'andamento di una misura o di una frequenza di un evento che si sviluppa in funzione del tempo. L'RD_t non è basato sulle incidenze di contagio, bensì di positività e per questo motivo viene detto "indice di replicazione diagnostica". Se il suo valore è >1 significa che gli eventi analizzati, contagi o positività, sono in crescita, se è =1 significa che sono stazionari, se è <1 significa invece che sono in decrescita.

Materiale e metodi per il calcolo

Lo sviluppo iniziale di un'epidemia, dove il contagio avviene tra persone, si può ritenere sia una progressione geometrica la cui ragione è determinata dalla capacità di un infetto di contagiare nell'unità di tempo *t* altre persone. Nell'esempio in Figura 1 l'indice RD_t misura l'intensità di replicazione (la ragione) in assenza di misure di contenimento e il suo valore è pari a 2.

Figura 1 - Esempio Intensità di Replicazione RD_0=2



L'esempio di calcolo proposto nella Tabella 1 ipotizza invece che il 1° marzo ci sia un solo infetto che in sei giorni, prima di essere isolato e impossibilitato a contagiare ulteriormente, contagi tre persone. Il 31 marzo con un andamento costante si arriverebbe a un totale di 243 infetti e un numero di nuovi contagi giornalieri pari a 40 casi.

Tabella 1

	Ragione	1,20093696
data	cumulativi	giornalieri
01-mar	1	0
02-mar	1,20	0,20
03-mar	1,44	0,24
04-mar	1,73	0,29
05-mar	2,08	0,35
06-mar	2,50	0,42
07-mar	3,00	0,50
08-mar	3,60	0,60
09-mar	4,33	0,72
10-mar	5,20	0,87
11-mar	6,24	1,04
12-mar	7,49	1,25
13-mar	9,00	1,51
14-mar	10,81	1,81
15-mar	12,98	2,17
16-mar	15,59	2,61
17-mar	18,72	3,13
18-mar	22,48	3,76
19-mar	27,00	4,52
20-mar	32,43	5,43
21-mar	38,94	6,52
22-mar	46,77	7,82
23-mar	56,16	9,40
24-mar	67,45	11,29
25-mar	81,00	13,55
26-mar	97,28	16,28
27-mar	116,82	19,55
28-mar	140,30	23,47
29-mar	168,49	28,19
30-mar	202,34	33,86
31-mar	243,00	40,66

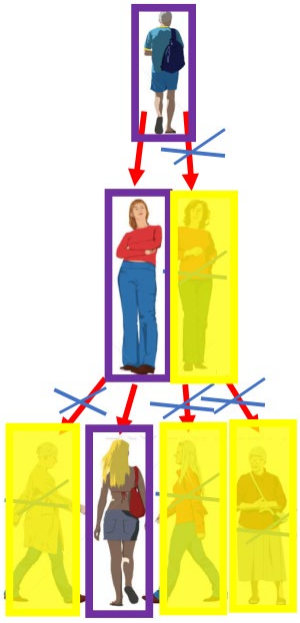
Se quindi la ragione fosse pari alla radice sesta di 3, cioè = 1,20093696, il rapporto tra la frequenza del sesto giorno (n=3) e quella del giorno zero (n=1) sarebbe esattamente 3. Come mostrato nella Tabella 2, l'indice di replicazione diagnostica RD_t a lag 6 risulterebbe esattamente costante con valore 3 e i casi triplicherebbero ogni 6 giorni: 1, 3, 9, 27, 81, 243.

Tabella 2

RDt lag 1	RDt lag 2	RDt lag 3	RDt lag 4	RDt lag 5	RDt lag 6
R**1	R**2	R**3	R**4	R**5	R**6
1,201					
1,201	1,442				
1,201	1,442	1,732			
1,201	1,442	1,732	2,080		
1,201	1,442	1,732	2,080	2,498	
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000
1,201	1,442	1,732	2,080	2,498	3,000

L'indice RD_t misura l'intensità di replicazione (la ragione) al tempo *t* in presenza di misure di contenimento e naturalmente diminuisce se in qualche modo si riesce a ridurre le capacità di contagiare, come mostrato in Figura 2. La progressione rimane sempre di tipo geometrico, ma la ragione diminuisce e se diventa inferiore ad uno l'epidemia regredisce.

Figura 2



Nelle Tabella 3 e Figura 3 sono mostrati i valori reali forniti dalla Protezione Civile dei contagi giornalieri (positivi) in Italia [https://github.com/pcm-dpc/COVID-19] nel periodo dall'1 marzo al 31 marzo. Come si può osservare dalla tabella, si passa da 566 positivi giornalieri al 1^0 marzo a 4053 positivi giornalieri al 31 marzo.

Calcolo delle medie mobili a 7 giorni centrate

Per evitare oscillazioni e fluttuazioni nei dati dovute alla differente esecuzione e rendicontazione dei test orofaringei (tamponi) eseguiti nei 7 giorni della settimana, come mostrato nella Figura 3, si procede calcolando le medie mobili non pesate di 7 giorni centrate sul valore intermedio, il quarto giorno, come mostrato dalla linea rossa in Figura 4. Questo produce una piccola perdita di informazione sui primi 3 giorni (1-3 marzo) e sugli ultimi 3 giorni (29-31 marzo), ma stabilizza notevolmente l'andamento eliminando la componente di 'stagionalità' giornaliera non associata allo sviluppo del contagio.

Figura 3

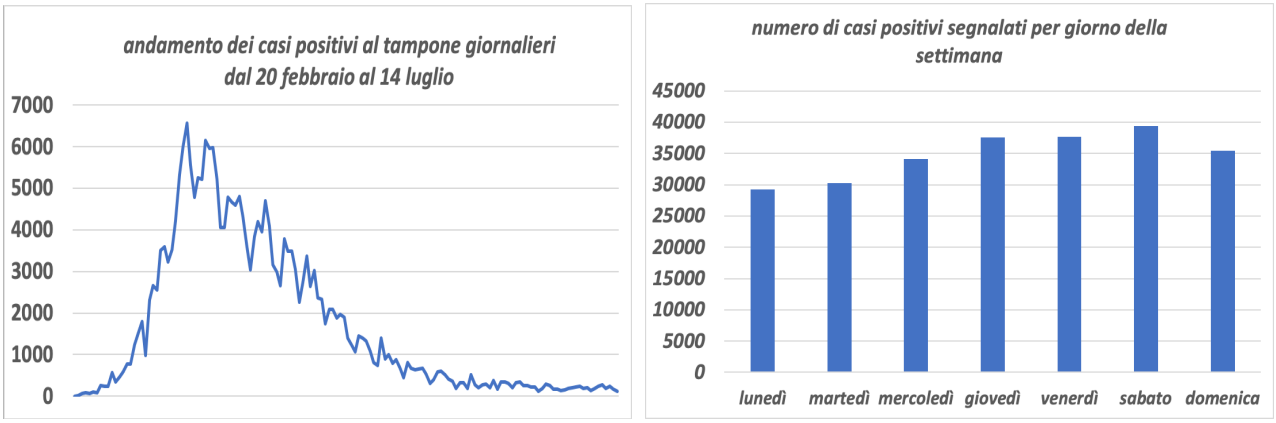


Figura 4

Calcolo dell'indice RD_t a vari lag

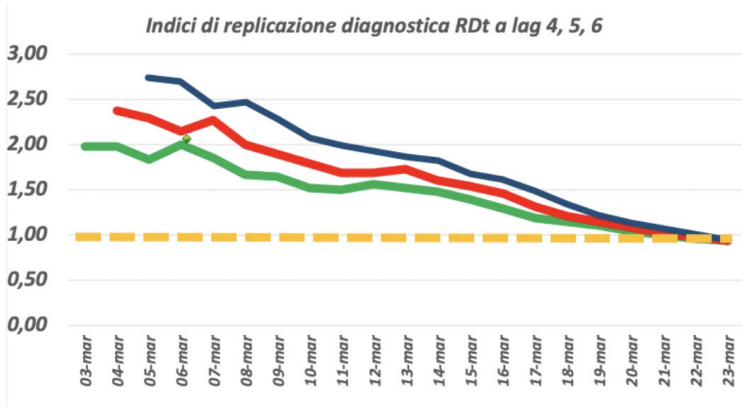
Come mostrato nellaTabella 3, eseguendo il rapporto tra la media mobile a 7 giorni riferita all'8 marzo (numeratore) e la media mobile a 7 giorni riferita al 4 marzo (denominatore) si ottiene un valore di RD_t a lag 4 pari a 1,97 (1339,0 / 679,3). Scorrendo ogni giorno per riga e variando i lag per colonna si ottengono i dati in Tabella 3 sull' RD_t giornaliero a lag 4, 5, 6. Ad esempio, l'RD_t a lag 6 del 15 marzo ha un valore di 2,07 ed è dato dal rapporto tra la media mobile del 15 marzo e la media mobile del 9 marzo, cioè 3321,6/1607,9.

Tabella 3

data	positivi	medie mob. 7		LAG 4		LAG 5		LAQG 6
01-mar	566							
02-mar	342							
03-mar	466		DENOMINATORE		DENOMINATOR		DENOMINATOR	
04-mar	587	679,3						
05-mar	769	811,6						
06-mar	778	1019,4						
07-mar	1247	1092,4						
08-mar	1492	1339,0		1,97				
09-mar	1797	1607,9	NUMERATORE	1,98		2,37		
10-mar	977	1860,6		1,83	NUMERATORE	2,29		2,74
11-mar	2313	2182,0		2,00		2,14	NUMERATORE	2,69
12-mar	2651	2481,7		1,85		2,27		2,43
13-mar	2547	2686,9		1,67		2,01		2,46
14-mar	3497	3051,0		1,64		1,90		2,28
15-mar	3590	3321,6		1,52		1,79		2,07
16-mar	3233	3703,1		1,49		1,70		1,99
17-mar	3526	4194,4		1,56		1,69		1,92
18-mar	4207	4631,6		1,52		1,72		1,87
19-mar	5322	4913,0		1,48		1,61		1,83
20-mar	5986	5135,3		1,39		1,55		1,68
21-mar	6557	5381,4		1,28		1,45		1,62
22-mar	5560	5524,7		1,19		1,32		1,49
23-mar	4789	5643,4		1,15		1,22		1,35
24-mar	5249	5639,6		1,10		1,15		1,22
25-mar	5210	5556,3		1,03		1,08		1,13
26-mar	6153	5507,3		1,00		1,02		1,07
27-mar	5959	5401,7		0,96		0,98		1,00
28-mar	5974	5230,9		0,93		0,93		0,95
29-mar	5217							
30-mar	4050							
31-mar	4053							

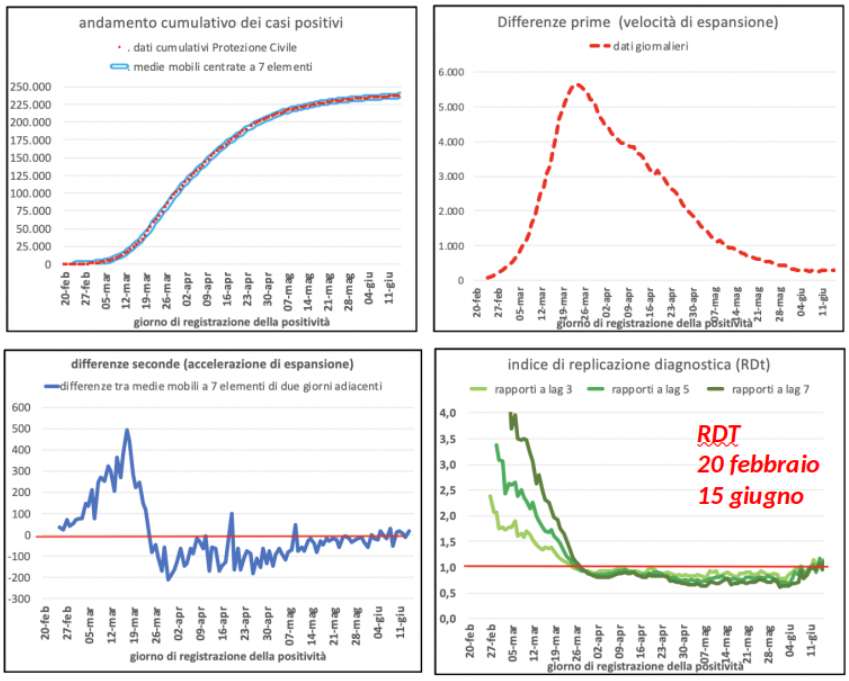
I predetti valori di RDd_t a lag 4, 5 e 6 sono visibili nella Figura 5.

Figura 5



L'indice RD_t è un indice da usare in un contesto descrittivo completo come quello ragionato e statico mostrato nella Figura 6, dove la sua lettura è associata all'andamento cumulativo dei casi positivi, alla velocità di espansione del contagio e all'accelerazione del medesimo.

Figura 6



Può altresì essere usato in un contesto descrittivo di tipo dinamico come quello presentato nella Figura 7, 8, 9 e visibile all'indirizzo prolea.shinyapps.io/covid19aie/, dove sia possibile effettuare la scelta dei giorni nella media mobile, dei lag e dell'area di analisi.

Figura 7

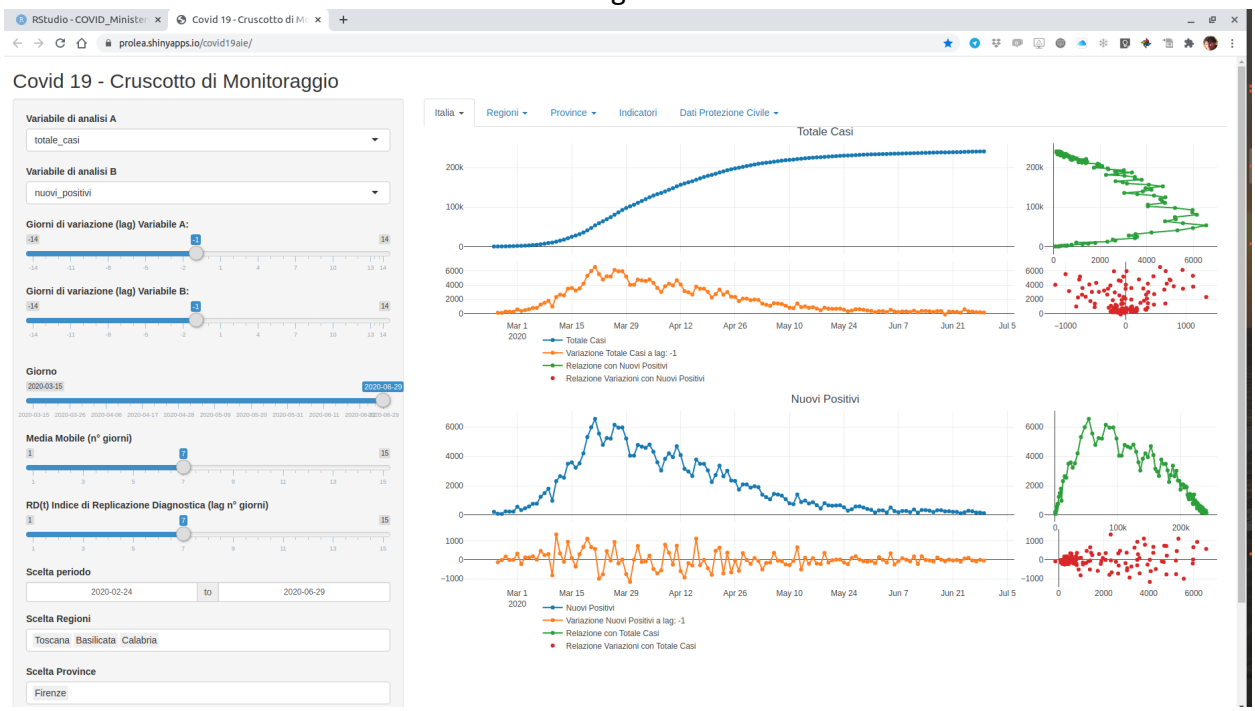


Figura 8

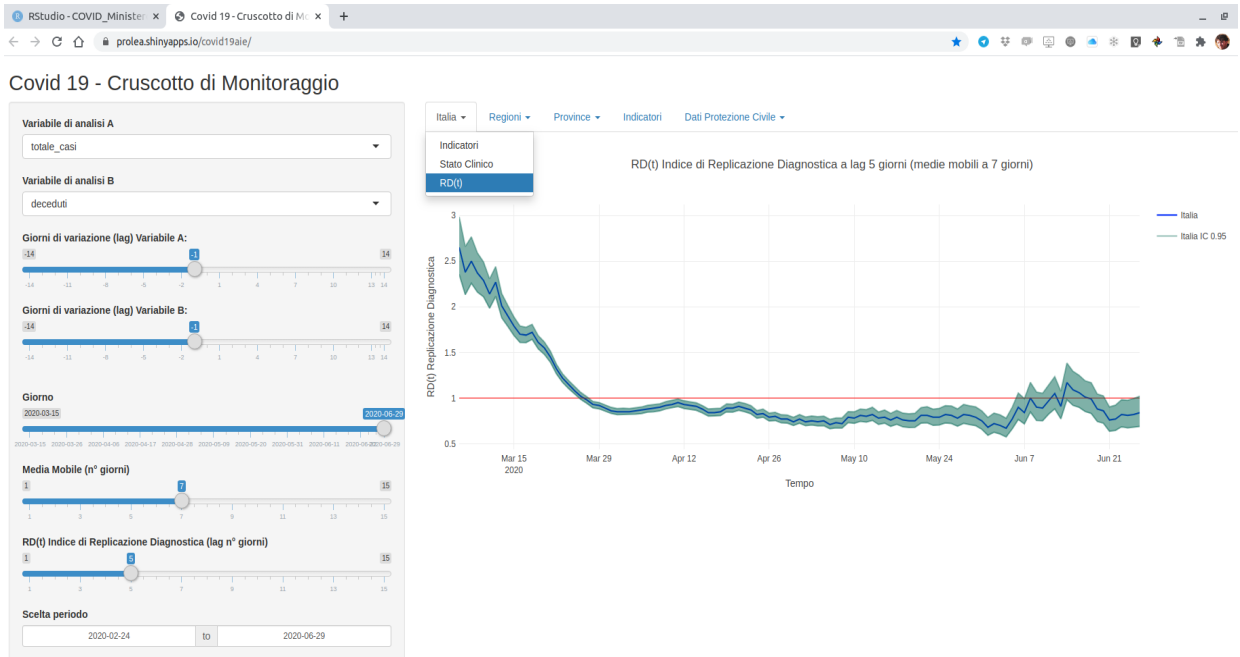
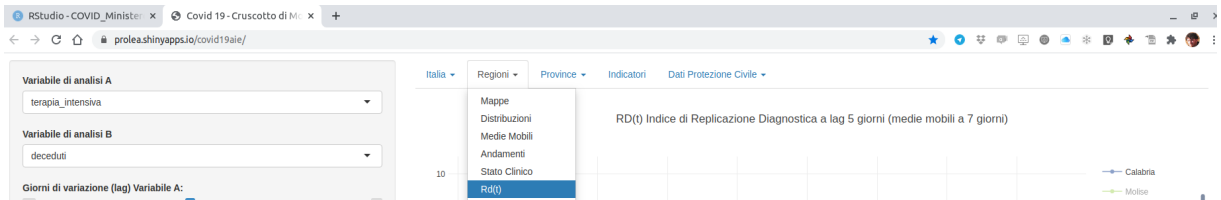
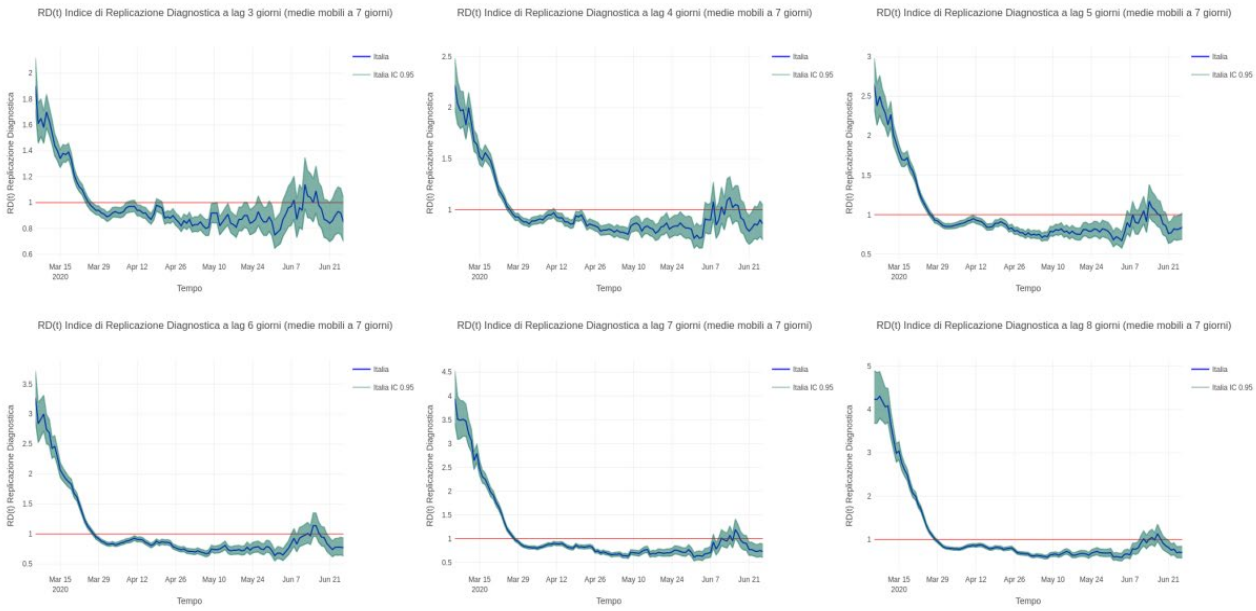


Figura 9



Conclusioni e scelta dei lag

Figura 10



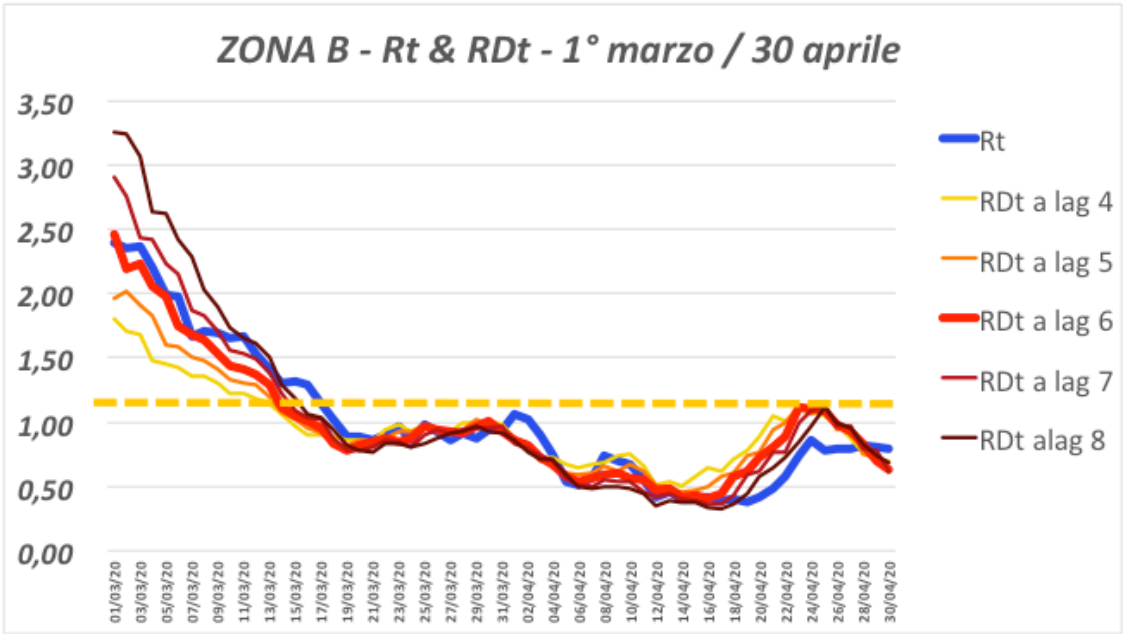
La scelta di lag differenti, mostrata in Figura 10, è un tentativo empirico di ovviare al problema della mancata costanza nei tempi di generazione. L'ipotesi che la distribuzione dei serial interval, cioè dei tempi di generazione tra un contagio ed il successivo, rimanga costante durante tutto il periodo appare debole e considerare l'andamento dell'RDt in corrispondenza di più lag può fornire un'indicazione a più ampio spettro sulla trasmissibilità dell'infezione.

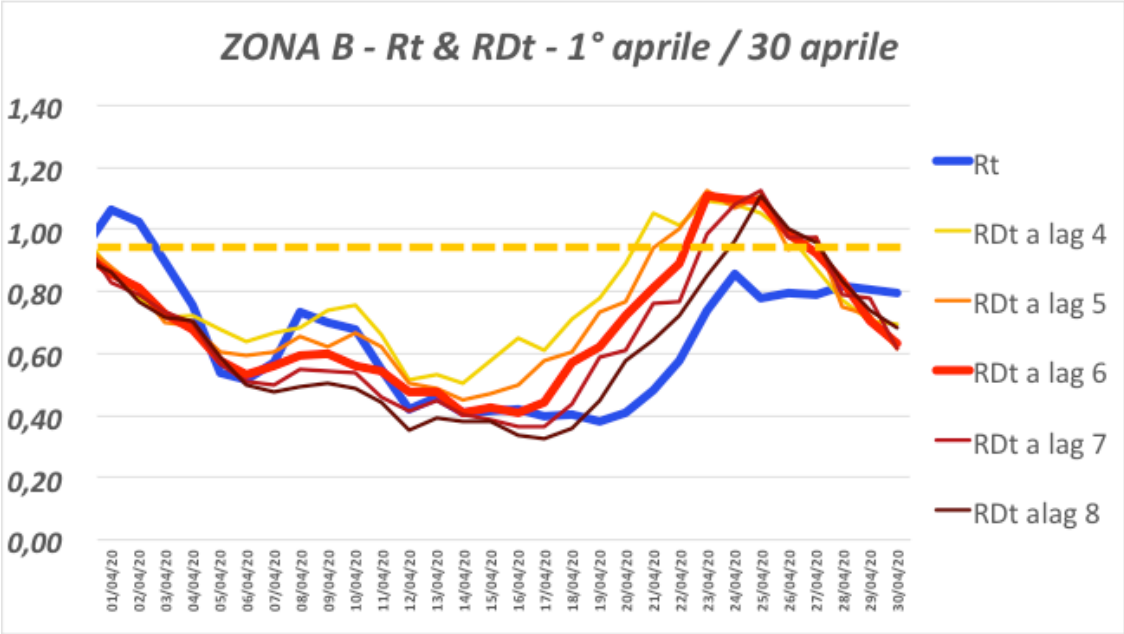
Allegato 3
Confronto tra R_t e RD_t calcolato in base a data sintomi

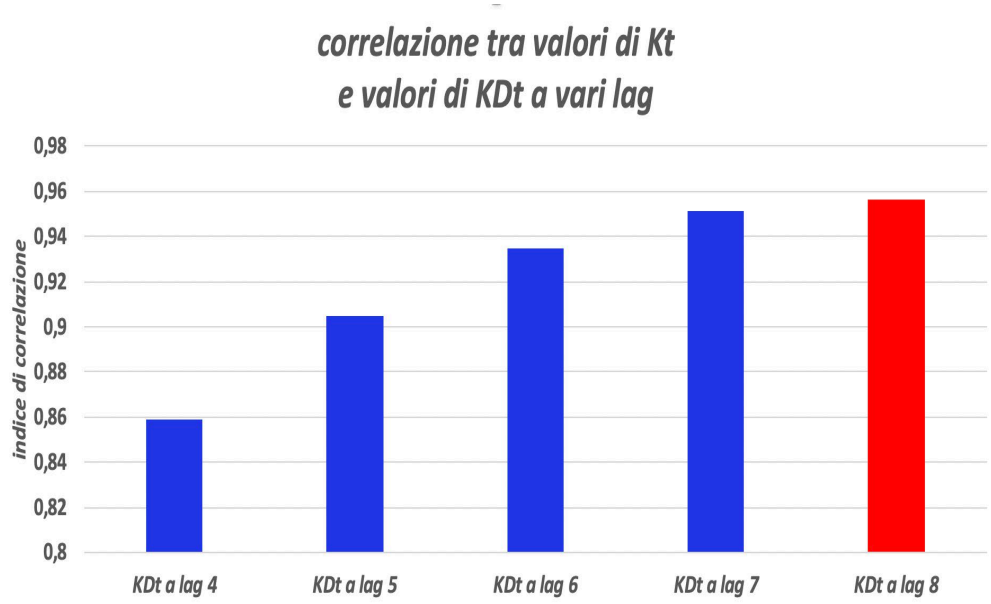
Utilizzando i dati individuali resi disponibili per quattro zone italiane (una Regione e una provincia del Nord, zone A e B rispettivamente, e due Regioni del Sud, zone C e D) e aggregando i soli record che contenevano la variabile "data inizio sintomi" si è calcolato l'indice R_t (secondo la metodica di Cori descritta) utilizzando il software R e con gli stessi dati si è calcolato l'indice RD_t di replicazione diagnostica illustrato per i lag 4, 5, 6, 7 e 8 giorni utilizzando il foglio di calcolo Excel.
Si è poi calcolato per ogni zona l'indice di correlazione di Pearson tra i valori di R_t e i valori di RD_t ai vari lag.
Nel seguito sono riportati per le zone B, C e D:

- i grafici dell'andamento di R_t e di RD_t ai vari lag con riferimento al periodo 1 marzo – 30 aprile
- i grafici analoghi per il solo mese di aprile
- i grafici dei valori dell'indice di correlazione tra R_t e RD_t ai diversi lag

Zona B

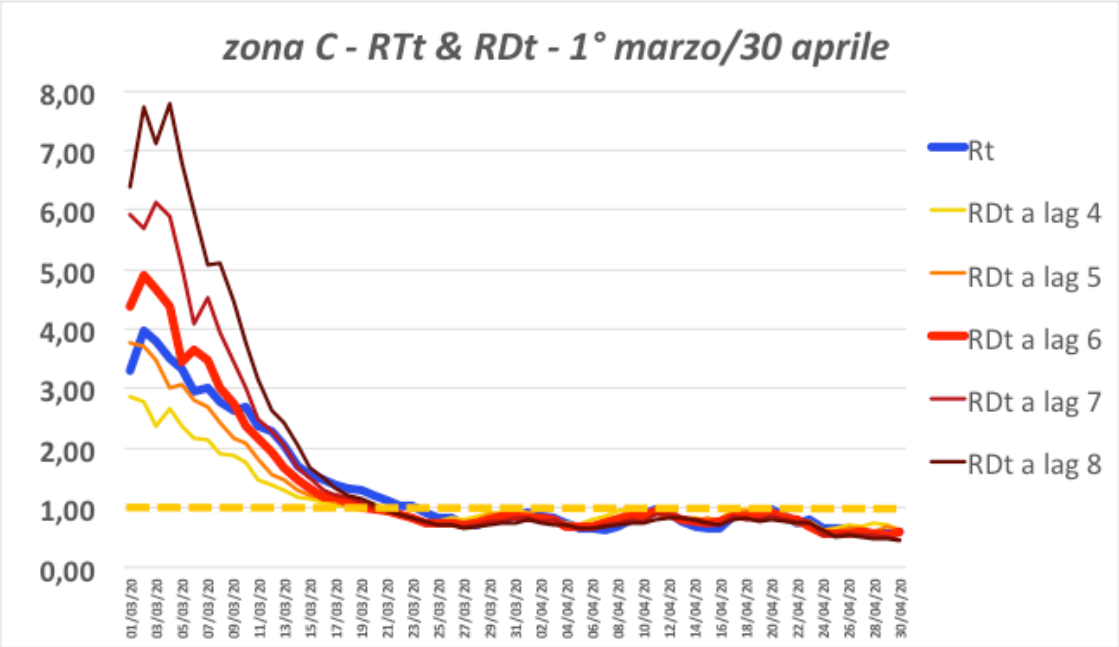


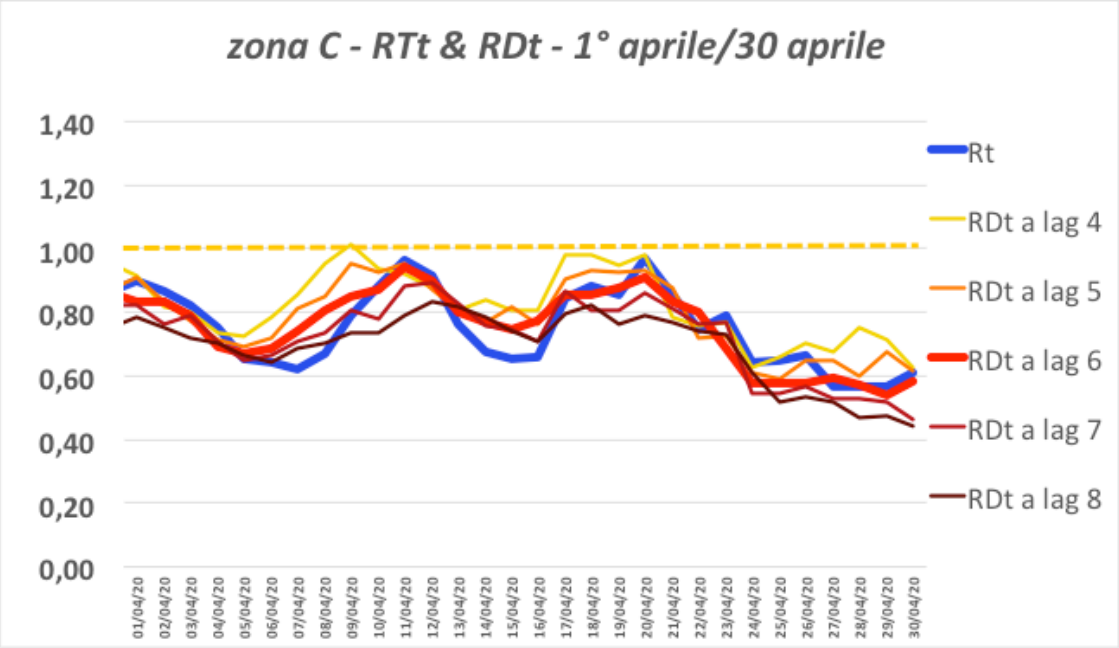


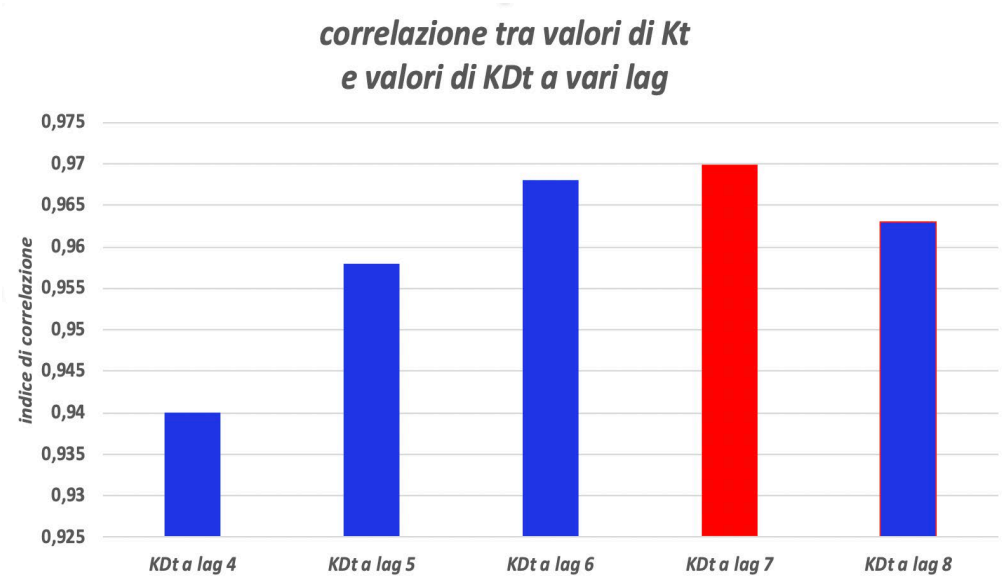


La corrispondenza tra gli indici R_t e RD_t ai vari lag risulta meno marcata soltanto nella seconda parte del mese di aprile. Va tenuto comunque presente che i due indici si basano su metodi di calcolo e su valori di medie mobili differenti e possono quindi risentire anche delle variazioni in frequenze non molto elevate come nel caso della Provincia a cui si riferiscono i risultati. La correlazione tra i due valori di R_t e RD_t risulta massima per lag pari a 8 giorni.

Zona C

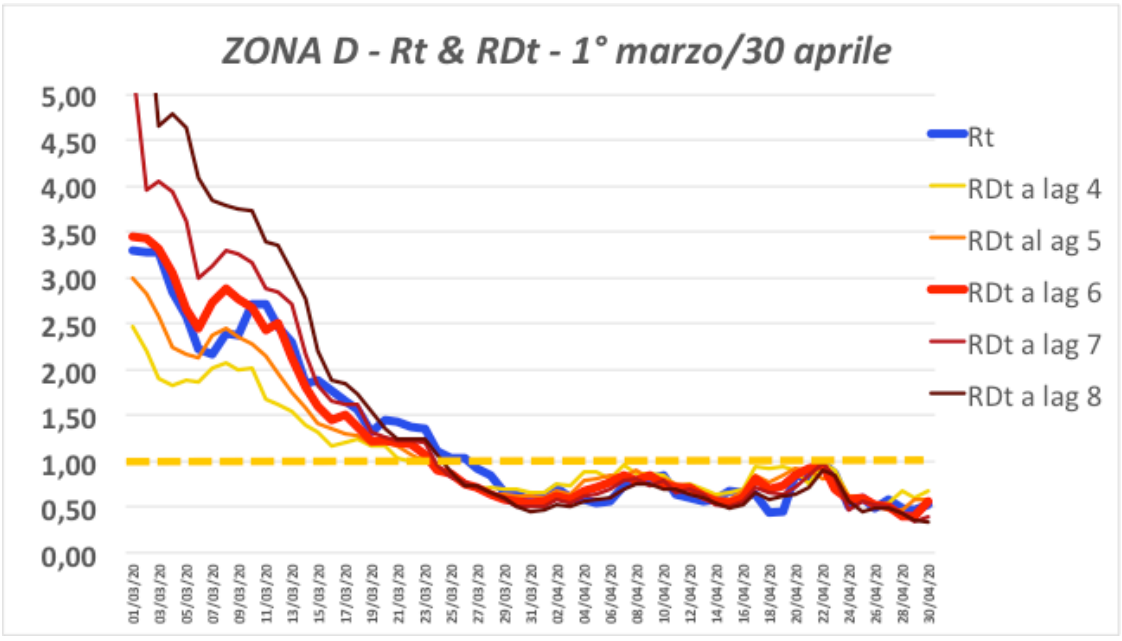

zona C - RTt & RDt - 1° marzo/30 aprile
8,00
7,00
6,00
5,00
4,00
3,00
2,00
1,00
0,00
Rt
RDt a lag 4
RDt a lag 5
RDt a lag 6
RDt a lag 7
RDt a lag 8

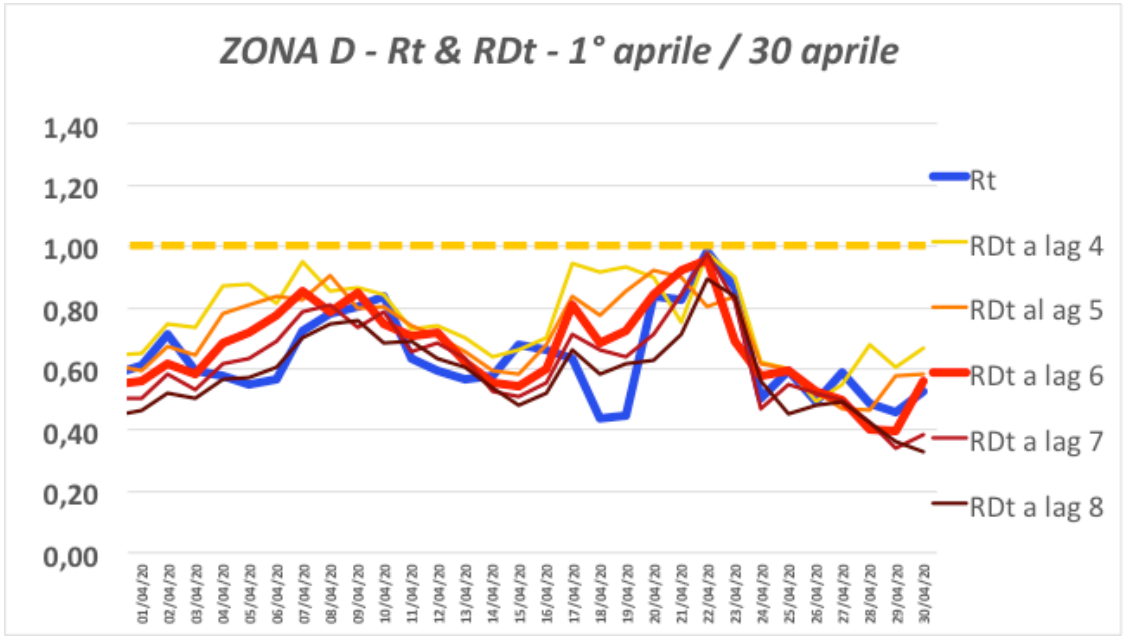

zona C - RTt & RDt - 1° aprile/30 aprile
1,40
1,20
1,00
0,80
0,60
0,40
0,20
0,00
Rt
RDt a lag 4
RDt a lag 5
RDt a lag 6
RDt a lag 7
RDt a lag 8

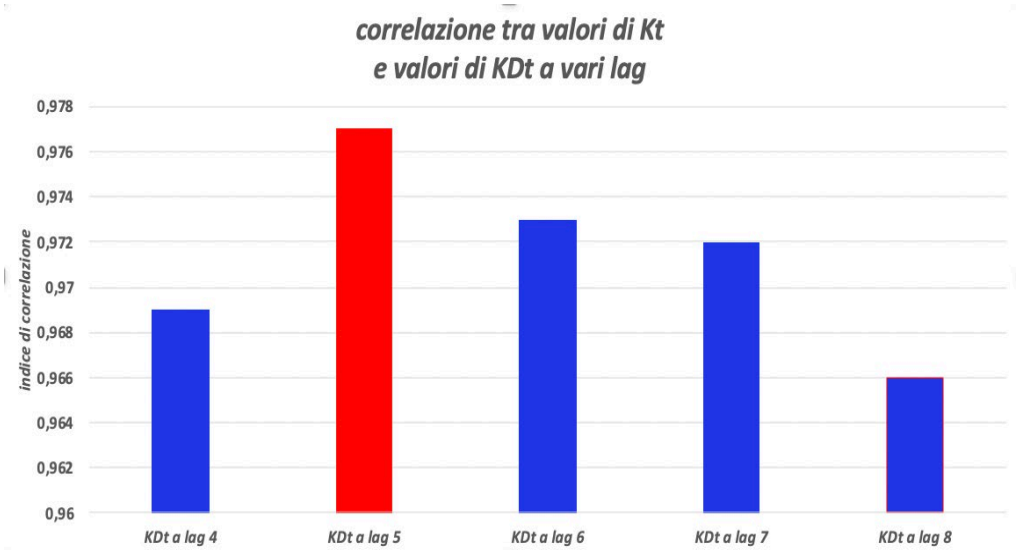


La corrispondenza tra gli indici R_t e RD_t ai vari lag risulta evidente a parte nella fase iniziale del monitoraggio per cui gli indici RD_t a lag maggiori tendono ad assumere valori più alti. La correlazione tra i due valori di R_t e RD_t risulta massima per lag pari a 7 giorni, ma ha comunque valori sempre superiori a 0.92.

Zona D







Il confronto tra gli indici rivela un comportamento analogo a quello della zona C, salvo che per alcuni giorni intorno alla metà del mese di aprile durante i quali probabilmente le oscillazioni nella registrazione delle incidenze sintomatiche sono smorzate maggiormente dall'indice RD_t che viene calcolato utilizzando medie mobili su un intervallo più esteso. La correlazione tra i due valori di R_t e RD_t risulta massima per lag pari a 5 giorni.

Risulta in conclusione evidente da tutte le zone analizzate che i due indici non sono coincidenti, ma forniscono informazioni sovrapponibili. Il fatto che la correlazione tra l'R_t e l'RD_t non sia massima in corrispondenza degli stessi lag potrebbe suggerire che i tempi di generazione non abbiano sempre la stessa distribuzione come si ipotizza invece per il calcolo dell'R_t.

Allegato 4
Il confronto tra gli indici di replicazione diagnostica calcolati sulla data dell'esito tampone e sulla data dell'inizio sintomi in due aree del Nord Italia e in due zone del Sud Italia

L'indice di Replicazione Diagnostica Rd_t è definito dal rapporto tra una stima degli eventi al tempo t e degli eventi al tempo t-s dove s sono i giorni che si stima siano necessari affinché un caso possa riprodurne un altro. L'intervallo s viene denominato in vario modo: tempo di generazione, serial interval, lag. In particolare il rapporto è tra due medie non pesate di due periodi distanti lag differenti, il rapporto è tra una media mobile centrata non pesata di n=7 elementi al tempo t e una media mobile centrata non pesata di n=7 elementi ai tempi t-s dove s può assumere diversi possibili valori, solitamente pari a 4, 5 , 6, 7 o 8 giorni. Generalmente RD_t utilizza la data di registrazione della positività, mentre altri indici come ad esempio l'indice R_t utilizzano la data dell'inizio sintomi. L'indice tuttavia può essere calcolato anche utilizzando le frequenze di positivi alla data di inizio sintomi.

Confronto tra RD_t calcolati su dati degli stessi soggetti positivi per data inizio sintomi o per data esito tampone

Il confronto tra l'uso delle diverse tipologie di dati di monitoraggio è stato effettuato limitatamente alle quattro aree per cui è stato possibile disporre di dati individuali completi relativi ai mesi di marzo e di aprile.
Nelle Figure 4.a, 4.b, 4.c e 4.d vengono riportati i confronti relativi rispettivamente a una Regione e una provincia nel Nord Italia (a e b) e a due Regioni del Sud Italia (c e d). Vengono riportati i valori dell'RD_t calcolato con le incidenze per data esito tampone e per data inizio sintomi relativi alle singole zone a, b, c e d rispettivamente, la media dei giorni di distanza tra la data di inizio sintomi e la data di esito tampone, e l'RD_t a lag 6 per data tampone e per data inizio sintomi anticipata o posticipata di un numero n variabile di giorni.

Figura 4.a

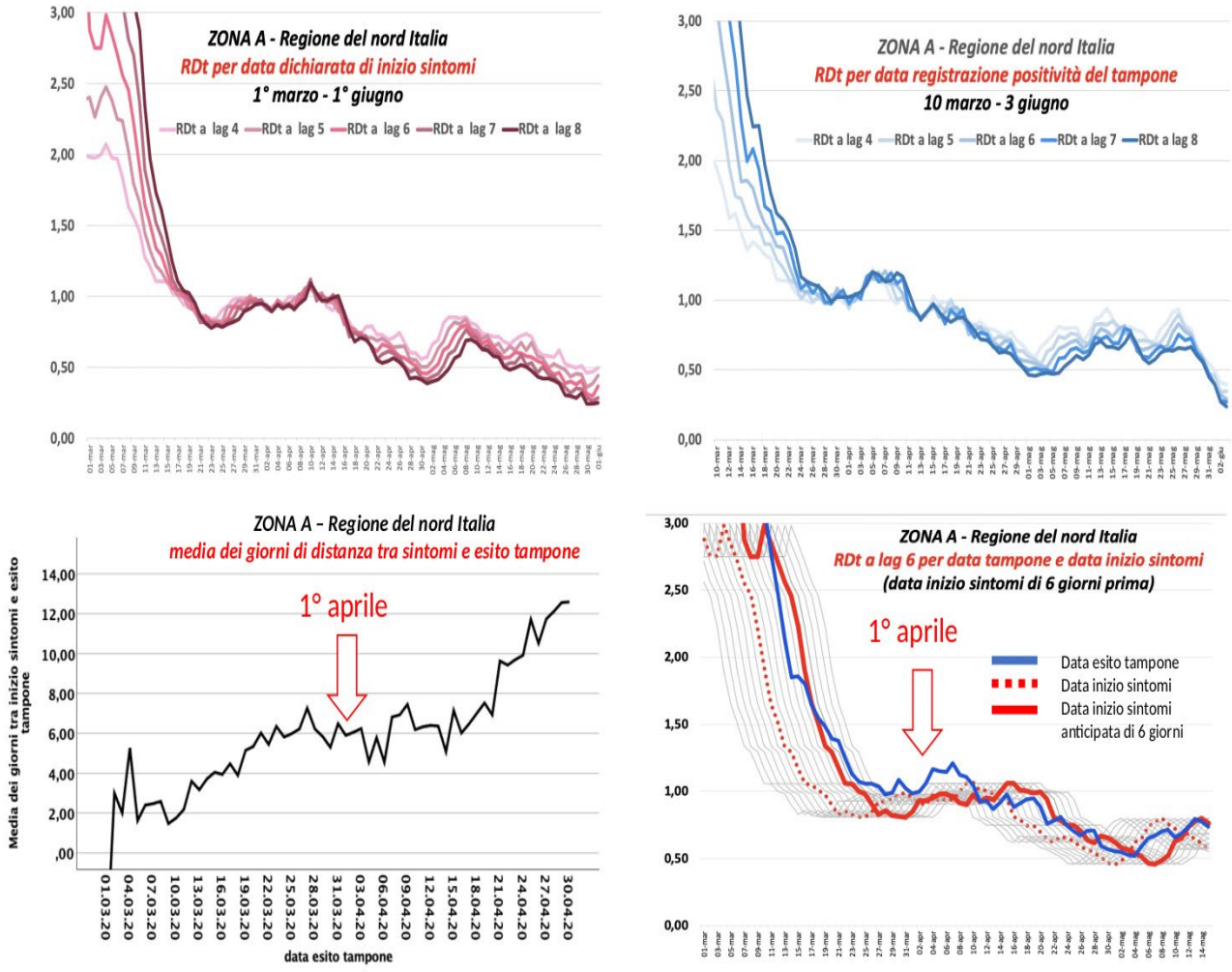
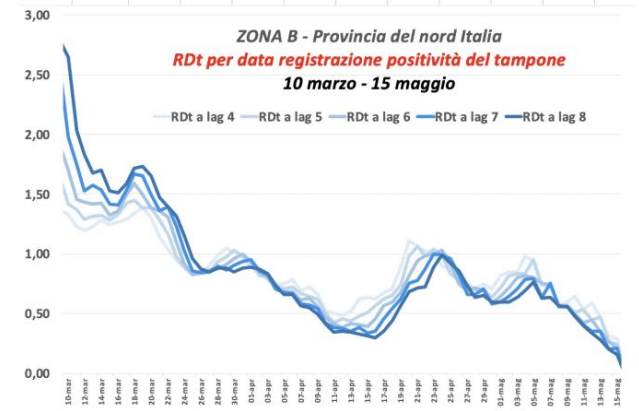
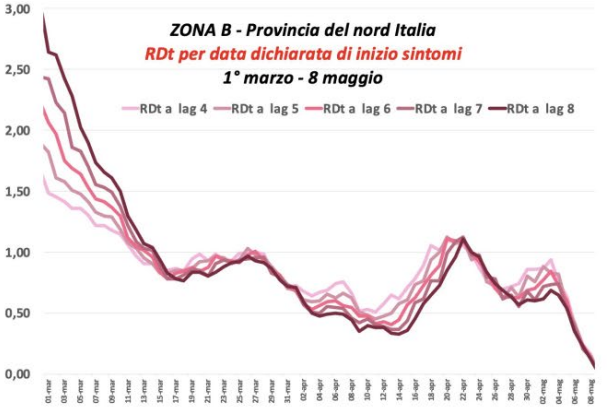
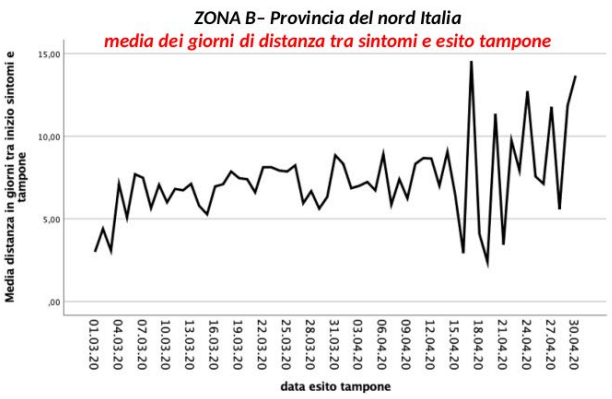

ZONA A - Regione del nord Italia
RDt per data dichiarata di inizio sintomi
1° marzo - 1° giugno
RDt a lag 4
RDt a lag 5
RDt a lag 6
RDt a lag 7
RDt a lag 8
ZONA A - Regione del nord Italia
RDt per data registrazione positività del tampone
10 marzo - 3 giugno
RDt a lag 4
RDt a lag 5
RDt a lag 6
RDt a lag 7
RDt a lag 8
ZONA A – Regione del nord Italia
media dei giorni di distanza tra sintomi e esito tampone
1° aprile
Media dei giorni tra inizio sintomi e esito tampone
data esito tampone
ZONA A - Regione del nord Italia
RDt a lag 6 per data tampone e data inizio sintomi
(data inizio sintomi di 6 giorni prima)
1° aprile
Data esito tampone
Data inizio sintomi
Data inizio sintomi anticipata di 6 giorni

Figura 4.b


ZONA B - Provincia del nord Italia
RDt per data registrazione positività del tampone
10 marzo - 15 maggio
RDt a lag 4
RDt a lag 5
RDt a lag 6
RDt a lag 7
RDt a lag 8


ZONA B - Provincia del nord Italia
RDt per data dichiarata di inizio sintomi
1° marzo - 8 maggio
RDt a lag 4
RDt a lag 5
RDt a lag 6
RDt a lag 7
RDt a lag 8


ZONA B– Provincia del nord Italia
media dei giorni di distanza tra sintomi e esito tampone
Media distanza in giorni tra inizio sintomi e tampone
data esito tampone

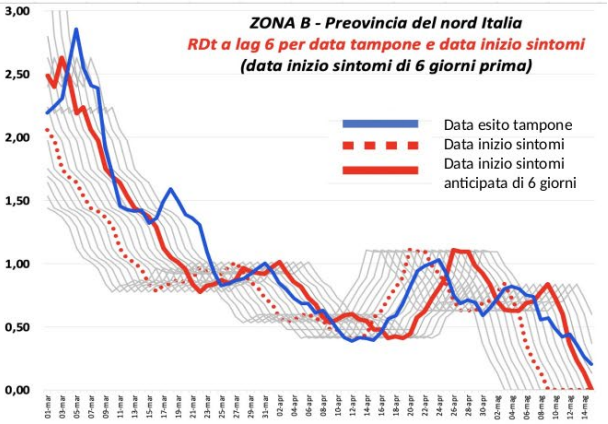
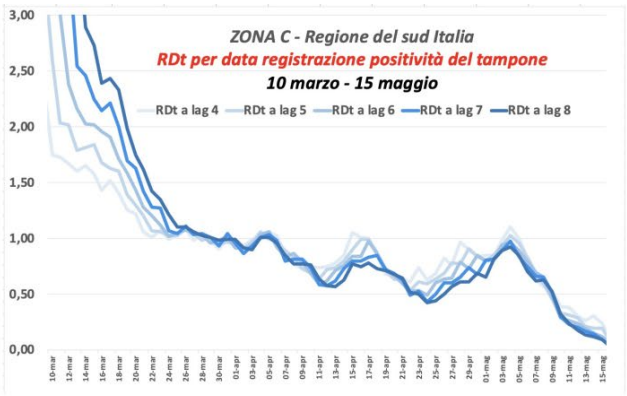
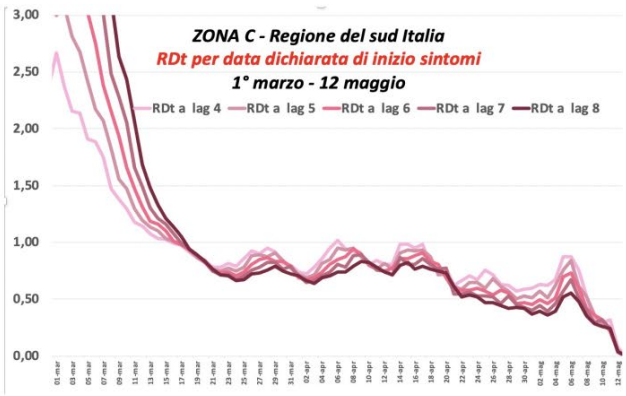
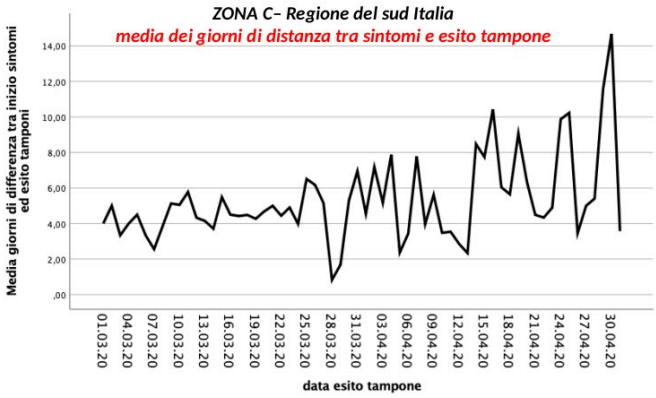

ZONA B - Preovincia del nord Italia
RDt a lag 6 per data tampone e data inizio sintomi
(data inizio sintomi di 6 giorni prima)
Data esito tampone
Data inizio sintomi
Data inizio sintomi anticipata di 6 giorni

Figura 4.c


ZONA C - Regione del sud Italia
RDt per data registrazione positività del tampone
10 marzo - 15 maggio
RDt a lag 4
RDt a lag 5
RDt a lag 6
RDt a lag 7
RDt a lag 8


ZONA C - Regione del sud Italia
RDt per data dichiarata di inizio sintomi
1° marzo - 12 maggio
RDt a lag 4
RDt a lag 5
RDt a lag 6
RDt a lag 7
RDt a lag 8


ZONA C– Regione del sud Italia
media dei giorni di distanza tra sintomi e esito tampone
Media giorni di differenza tra inizio sintomi ed esito tamponi
data esito tampone

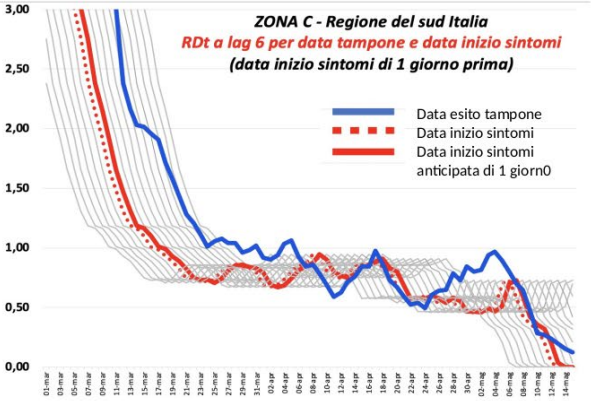
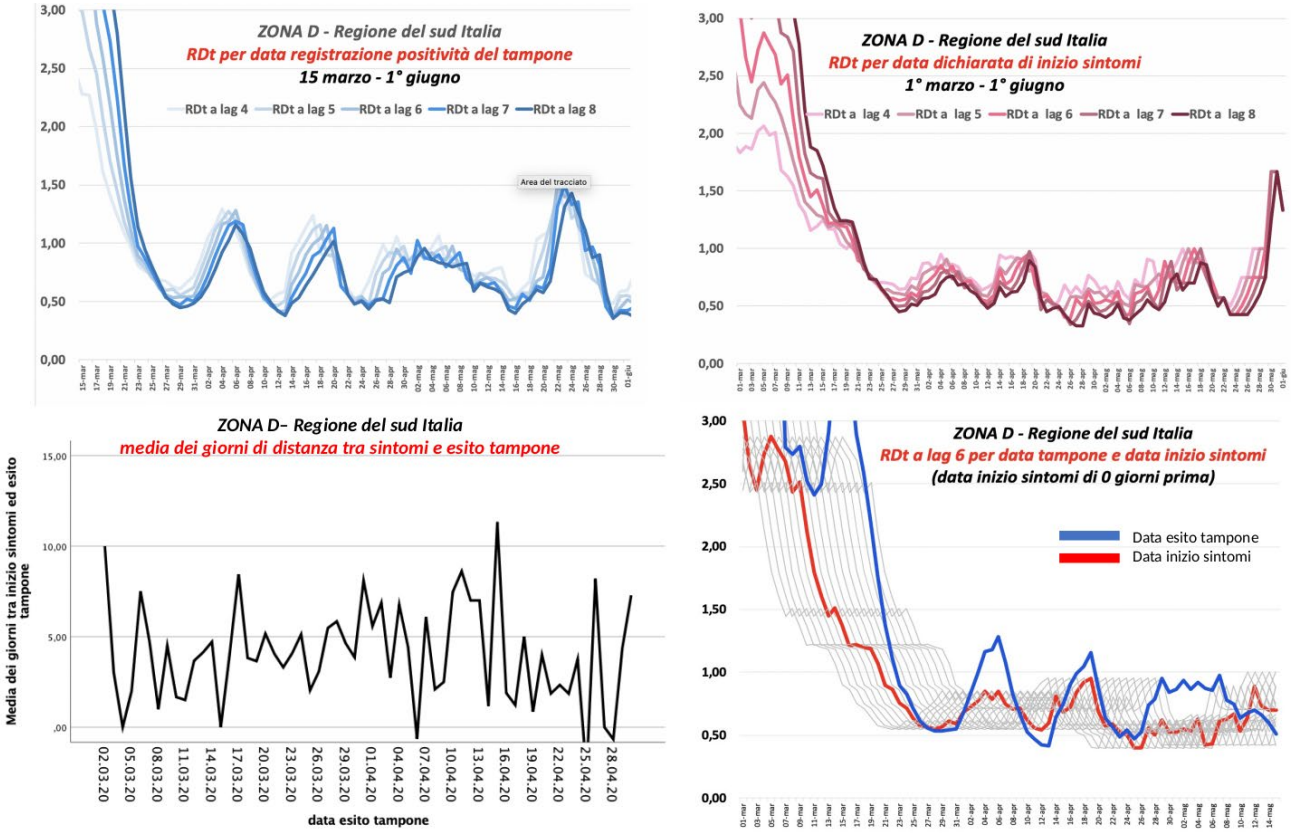

ZONA C - Regione del sud Italia
RDt a lag 6 per data tampone e data inizio sintomi
(data inizio sintomi di 1 giorno prima)
Data esito tampone
Data inizio sintomi
Data inizio sintomi anticipata di 1 giorn0

Figura 4.d



Nelle Figure 4.a, 4.b, 4.c e 4.d sono illustrati i differenti risultati che si ottengono calcolando l'RD_t ai vari lag per gli stessi casi positivi, ma utilizzando le due frequenze per data inizio sintomi oppure per data esito tamponi.

Si nota innanzitutto che, laddove l'andamento dell'epidemia cambia repentinamente, i valori dell'RD_t si differenziano maggiormente tra i vari lag in quanto, ad esempio, la modifica dell'andamento a quatto giorni risulta meno marcata di quella dell'andamento a otto giorni.

L'andamento dei due grafici per data inizio sintomi e data tampone è simile, ma non uguale perché innanzitutto il numero dei soggetti non è lo stesso in quanto molti casi non riportano la data di inizio sintomi e inoltre la distanza tra l'esordio dei sintomi e l'esito dei tamponi è variabile sia tra i soggetti che in media tra le diverse fasi dell'epidemia.

Come si può vedere nella figura 4.a in basso a sinistra, la media giornaliera della distanza tra inizio sintomi e esito tampone nell'area A qui descritta varia tra un valore di circa 2 giorni a inizio marzo per diventare di 6 giorni al primo di aprile e poi aumentare fino quasi a raddoppiare.

E' quindi evidente che questa distanza tra i due eventi monitorati produce degli andamenti differenti negli indici di replicazione diagnostica con discrepanze tra di essi a loro volta variabili nel tempo.

Nel grafico in alto a sinistra nelle rispettive aree è riportato l'andamento dell'indice RD_t calcolato in base alle frequenze per data esito tampone (linee blu). Nel grafico in alto a destra è riportato l'andamento dell'indice RD_t calcolato in base alle frequenze per data inizio sintomi (linee rosse). Nel grafico in basso a destra, per il solo RD_t a lag 6 vengono inoltre riportate, a fianco dei valori per data esito tampone (in blu), alcune copie dell' RD_t per data esordio sintomi traslate in verso positivo e negativo di vari giorni (linee sottili nere) e viene in particolare evidenziata in rosso quella spostata in avanti di 6 giorni.

Allegato 5

INTERVALLO DI CONFIDENZA DI RD_t

Vengono presentati nel seguito alcuni metodi che si possono utilizzare per ottenere un intervallo di confidenza per i valori stimati dell'indice RD_t inteso come rapporto tra due realizzazioni di conteggi.

METODO BINOMIALE ESATTO

RD_t è il rapporto fra due valori di variabili discrete, A/B, con A numero di casi "indotti" e B numero di casi "originari". A e B sono il numero di diagnosi fatte rispettivamente al tempo 1 e al tempo 0, lasciando che fra i due tempi intercorra un congruo periodo corrispondente al periodo medio di induzione di un nuovo caso (intervallo seriale: is).

Per dare questa interpretazione al rapporto, occorre procedere ad alcune assunzioni:

1) B è il numero di casi diagnosticati al giorno i-esimo, per i quali si assume una distribuzione di probabilità per i tempi di effettiva insorgenza fra i-k e i+h ad esempio di tipo Gamma;
2) A è il numero di casi diagnosticati al giorno (i + is) per cui vale la stessa distribuzione di probabilità dei reali tempi di insorgenza. Per tutti i casi (A e B) si considera fisso ed eguale il tempo medio di insorgenza;
3) A e B sono due realizzazioni di variabili casuali poissoniane, con massa-tempo di soggetti esposti a rischio costante ed eguale nei due tempi.

Sotto le condizioni 1 – 3 la distribuzione di campionamento dalla quale è possibile ricavare gli estremi dell'intervallo di confidenza al 95% può essere considerata assumendo "fissa" la somma dei valori A e B. Infatti l'informazione che interessa ricavare dai dati dipende solo dal modo di distribuirsi di questo totale fra i due tempi di osservazione. Da ciò deriva che la distribuzione di probabilità in gioco è la binomiale e RDt si può considerare un ODDS = P/(1 – P).

Da ciò il limite inferiore esatto dell'IC95% sarà ricavabile dal valore P(I) che assegna al valore osservato A o maggiore (fino a N = A + B) la probabilità del 2,5% o, detto in altro modo, il valore P(I) che assegna alla osservazione complementare, da 0 a (A – 1), la probabilità complementare di 97,5%.

Il limite superiore esatto dell'IC95% sarà ricavabile invece dal valore P(S) che assegna alla osservazione di un numero di eventi compreso fra 0 e A la probabilità del 2,5%.

La soluzione si può trovare, per prove ed errori, usando il foglio Excel, calcolando per diversi valori tentativi di P la probabilità cumulativa binomiale esatta per l'intervallo di valori di interesse (0,975 fra 0 e A – 1; 0,025 fra 0 e A) ricorrendo alla funzione "Distribuzione Binomiale".

Sempre con Excel si può usare la funzione "Ricerca obiettivo" in "Analisi di simulazione" ("Dati"), impostando la cella (nella quale compare la formula binomiale che calcola la probabilità cumulativa fra 0 e A) a 0,025 facendo variare la cella nella quale comparirà P(S) e la formula binomiale che calcola (1 – la probabilità cumulativa binomiale fra 0 e A-1), al valore 0,025 per P(I)).

Questo metodo esatto fu suggerito nel 1934 ed è noto come metodo esatto di Pearson – Clopper.

Ricavati gli estremi per P, quelli di RD_t saranno: $RD_t(I) = P(I)/(1-P(I))$; $RD_t(S) = P(S)/(1-P(S))$.

METODO BINOMIALE ESATTO CON DISTRIBUZIONE F DI FISHER

Il metodo esatto si può applicare anche sfruttando la relazione esistente fra la distribuzione binomiale e la distribuzione F di Fisher. Gli estremi dell'intervallo esatto di confidenza per P al 95% si ottengono così applicando le formule:

$$p(I) = 1+(N-A+1)/[A\ F_{[2A;2(N-A+1)]}$$

$$p(S) = 1+(N-A)/[(A+1)\ F_{2(A+1);\ 2(N-A);1-}$$

METODO GAUSSIANO APPROSSIMATO PER Ln(ODDS)

Considerato che RD_t, con totale fisso, è un odds, l'intervallo di confidenza si può calcolare ricorrendo all'approssimazione gaussiana della distribuzione di campionamento della sua trasformata logaritmica per cui: ES(Ln(A/B)) = Radq(1/A + 1/B) e gli estremi dell'IC95% sono dati da A/B Exp(±1,96 ES(Ln(A/B))).

ESEMPIO

In Piemonte il 4 marzo risultano A = 196 nuovi casi (media mobile su sette giorni: 1° - 7 marzo), B = 48 casi al 29 febbraio (media mobile: dal 26 febbraio a 3 marzo), RD_t = 196/48 = 4,08 (lag 5).

METODO BINOMIALE ESATTO:

LI(P) = 0,747780: Pr (196 o 197 o … o 244) = 0,02500 se p = 0,747780 $RD_t(I)$ = 2,96

LS(P) = 0,851246: Pr(0 o 1 o 2 o … o 195) = 0,97500 se p = 0,851246 $RD_t(S)$ = 5,72

METODO BINOMIALE ESATTO CON DISTRIBUZIONE F:

LI(P) = $[1+(248-196+1)/196\ 1,34916988]^{-1}$ = 0,74777958 $RD_t(I)$ = 2,96

LS(P) = (197/48) / (1+197/48 1,39431479) = 0,85124582 $RD_t(S)$ = 5,72

METODO APPROSSIMATO PER Ln(ODDS):

RD_t = ODDS = 196/48 = 4,0833333 $RD_t(I)$ = 2,98

ES(Ln(ODDS)) = Radq(1/196+1/48) = 0,16104463 $RD_t(S)$ = 5,60

Conclusioni

Come risulta dall'esempio, i due primi metodi, che fanno riferimento alla distribuzione binomiale, forniscono esattamente sempre gli stessi valori.

Il terzo metodo, che utilizza l'approssimazione gaussiana della distribuzione di campionamento della trasformata logaritmica di ODDS, fornisce risultati diversi, ma molto prossimi ai primi due, salvo nel caso di numeri molto piccoli, per cui l'approssimazione gaussiana risulta inadeguata.

Un possibile modo alternativo, semplice ma non preciso, per descrivere l'incertezza dell'indice RD_t dovuta a variazioni casuali

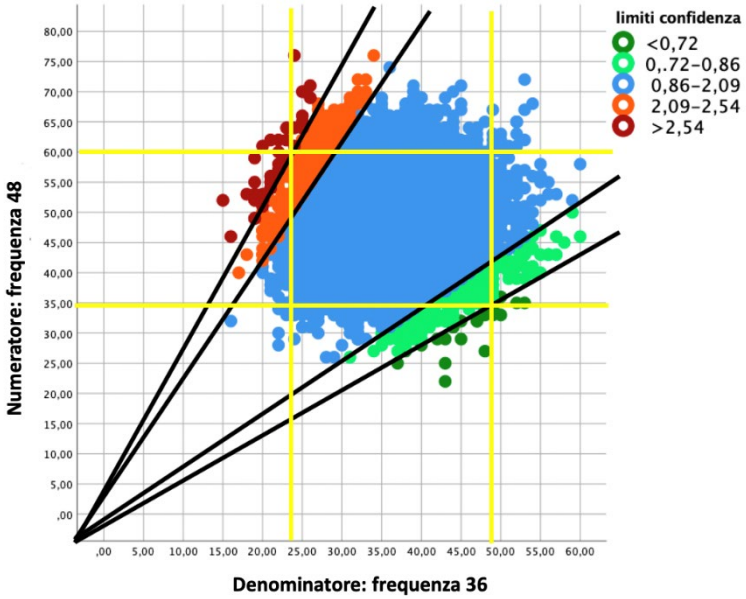
Per calcolare un intervallo di confidenza ci si deve rifare ad una ipotesi distributiva dell'indice di replicazione diagnostica come precedentemente indicato.

Chi avesse difficoltà a operare i calcoli necessari, peraltro non complessi, può seguire un'alternativa più semplice anche se non conduce a un intervallo di confidenza, bensì a quello che si può definire un intervallo di massima variabilità basato sui limiti di confidenza delle due variabili in questione.

Ipotizzando che l'RD_t sia il rapporto tra due frequenze, 48 e 36, si possono considerare le due frequenze come variabili distribuite secondo Poisson, per cui la varianza è uguale alla media e i limiti di confidenza al 95% sono dati dal prodotto di 1,96 per la loro radice quadrata, quindi 34,42 vs. 61,58 al numeratore e 24,24 vs. 47,76 al denominatore.

Il rapporto tra il valore massimo del numeratore ed il minimo del denominatore e il rapporto tra il valore minimo del numeratore ed il massimo del denominatore costituiscono l'intervallo massimo ottenibile con rapporti tra le due variabile se non eccedono dai loro limiti di confidenza singoli.

Questo “intervallo di massima incertezza” risulta quindi dato da
Limite inferiore: = 34,42 / 47,76 = 0,72
Limite superiore = 61,58 / 24,24 = 2,54
Il calcolo dell’intervallo di confidenza di RD_t calcolato con la formula indicata nella nota fornisce:
Limite inferiore: = 0,86
Limite superiore = 2,09



In figura si sono calcolati 10.000 rapporti tra coppie di valori casuali distribuiti come Poisson con medie pari a 36 e 48; i valori dei loro limiti di confidenza al 95% sono indicati con le linee gialle orizzontali e verticali. Tra i valori dei 10.000 rapporti simulati i 250 maggiori sono indicati in ocra (sia scuro che chiaro) e i 250 minori indicato in verde (sia scuro che chiaro). I livelli dell’intervallo di confidenza calcolato correttamente corrispondono esattamente ai valori simulati. L’intervallo massimo di incertezza invece corrisponde ai valori delle linee più esterne che intersecano quelle dei limiti delle due variabili; la percentuale di dati simulati che vi rientra è evidentemente più ampia rispetto alla precedente.
Questa possibile alternativa al calcolo degli intervalli di confidenza permette di approssimare l’area di incertezza dell’indice e può forse costituire per qualche utilizzatore una metodologia più semplice da implementare.
Nell’articolo il metodo presentato è indicato nella Figura 8 come *limiti di incertezza C*.

Bibliografia

C. J. Clopper and E. S. Pearson (1934): *The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial*. Biometrika 26: 404 – 413.