Luigi Biggeri e Luca Secondi

QUALE È STATA FINORA L'EVOLUZIONE DEI DATI SUL COVID-19

1. Premessa: i dati disponibili

Da un po' di tempo, come quasi tutti gli statistici abbiamo quotidianamente consultato ed elaborato i dati sul covid-19 che diffonde il Dipartimento della Protezione Civile (DPC) e le analisi che effettua l'Istituto Superiore di Sanità (ISS). Quasi tutti sostengono che i dati non sono adeguati per seguire l'espansione e l'evoluzione del virus e che la loro lettura è difficile e confusa o meglio non corretta. Eppure molti scienziati di tutte le discipline li usano per applicare modelli interpretativi dell'evoluzione dei fenomeni rilevati.

Da tempo insieme a Mauro Maltagliati abbiamo pensato di predisporre una rubrica, che dovrebbe essere inserita nella pagina sul corona virus del Dipartimento di Statistica, Informatica e Applicazioni (DiSIA) dell'Università di Firenze (www.disia.unifi.it) dedicata ad una diciamo "corretta" lettura dei dati divulgati e soprattutto alla presentazione di semplici analisi grafiche preliminari che sono la premessa indispensabile sia per cercare di individuare le cause che provocano i fenomeni, sia per la scelta di eventuali modelli di previsione della loro evoluzione futura.

Nei reports bisettimanali dell'ISS, che sono certamente ben fatti e facilmente leggibili, ci sono molte informazioni che aiutano a meglio comprendere, almeno in parte, il significato dei dati statistici pubblicati, ma purtroppo ne mancano molte altre. La critica che riguarda soprattutto i dati sugli individui attualmente positivi deriva dal fatto che questi chiaramente non rappresentano tutti i contagiati poiché dipendono dal numero di tamponi effettuati; ed è ovvio che se i positivi sono sottostimati, praticamente anche le altre statistiche a cascata sono inficiate. Tra l'altro, recentemente anche il numero dei morti per corona virus o con corona virus che sembrava essere il dato più attendibile è messo in dubbio, come risulta dalle polemiche riportate dai media.

Questa chiamiamola "carenza" è comprensibile considerando che anche coloro che raccolgono i dati sono in una situazione di grande emergenza e che i dati raccolti devono essere validati prima di essere pubblicati e sono in una continua fase di consolidamento.

Ad esempio, si sostiene giustamente che occorrerebbe conoscere il numero degli asintomatici e dei pauci-sintomatici. In realtà nel report ISS del 2 aprile 2020 c'è scritto che dallo stato clinico dei pazienti disponibile per 39.884 casi (su oltre 100.000 infettati) gli asintomatici risultano essere il 5,9% ed i pauci-sintomatici il 14 %, numeri non insignificanti! Ma dove poi sono classificati questi soggetti non lo sappiamo: sono coloro che si trovano in isolamento domiciliare? E' evidente che queste sono informazioni da avere per tutti gli individui per cui sono stati fatti i tamponi e per tutte le regioni.

Il problema più generale però la quantificazione dei positivi dipende proprio dai tamponi effettuati e dalla modalità di scelta degli individui ai quali fare i tamponi. Purtroppo all'inizio dell'epidemia lo svolgimento dei tamponi non è stato programmato, e forse non era possibile diversamente, come si dovrebbe fare in una rilevazione statistica (ricerca) ex-novo, con un disegno di campionamento di tipo probabilistico. L'emergenza ha spinto, come accade spesso quando si vuole seguire la diffusione di una epidemia, a organizzare la rilevazione con il metodo cosiddetto a valanga, che non consente stime probabilistiche per la variabile di interesse, alle quali si possa attribuire un certo grado di

affidabilità. E il problema principale è che non sono state date indicazioni stringenti alle regioni, ed ASL, su come dovevano procedere (o ad ogni modo ciascuna regione sembra che inizialmente abbia scelto una propria strada).

Come è noto, una survey a valanga (o snowball survey) parte con un certo numero di soggetti rilevati da inserire nella survey e di individuare (rilevare), i successivi soggetti, da studiare in base alla rete di relazioni che i nuovi soggetti hanno con quelli già contattati. Nel caso specifico dopo la prima rilevazione con i tamponi effettuati si sono inclusi altri soggetti cui fare tamponi considerando le persone con le quali quelle accertate positive erano state in contatto. Questa specie di rilevazione ha qualche vantaggio, soprattutto di costi, ma molti svantaggi e in particolare molti biases. Tuttavia è stato dimostrato come dopo vari steps di rilevazioni sia anche possibile effettuare stime asintoticamente non distorte.

Speriamo che, come è stato proposto, da ora in avanti i nuovi soggetti cui fare i tamponi vengano scelti in modo più consono, ad esempio dopo avere disegnato e attuato una indagine campionaria per individuare i potenziali "focolai" del virus (cioè dove nascono e si sviluppano i contagi). Indagine che a nostro avviso andrebbe fatta al livello di ASL (e Sistemi Locali del Lavoro interni alle ASL, dove come è noto ci sono i maggiori flussi di mobilità per motivi di studio e di lavoro). Il campione potrebbe essere basato sugli archivi Istat delle unità locali delle imprese industriali e di servizi, in particolare di farmacie e di imprese commerciali oltre una determinata dimensione, che sono aperti in base alle normative attualmente vigenti. Indagine che potrebbe essere integrata con una rilevazione nelle zone di traffico attraverso la tipologia della rilevazione "drive through" che alcune regioni hanno già iniziato ad implementare.

In attesa che ciò venga realizzato ci sentiamo comunque di affermare, sommessamente, che con più rilevazioni con i tamponi vengono effettuate, più i dati si avvicineranno ad una stima "sufficientemente" corretta almeno per alcune variabili, come in parte si rileva anche dai grafici che presenteremo. Ciò può quindi giustificare una loro analisi descrittiva specialmente se ci si riferisce all'ultima parte delle serie.

Comunque, per una interpretazione corretta delle serie storiche sono necessarie molte informazioni che non sono a nostra disposizione, come segnalato dal nostro presidente nella lettera inviata al presidente dell'ISS e anche nella recente intervista al Mattino di Napoli (quali dati elementari anonimi con informazioni sull'età del soggetto, sulla data dei sintomi e del tampone, delle patologie preesistenti, comune di residenza e di lavoro, ecc.). A noi sembrerebbe opportuno conoscere anche il numero delle persone sottoposte a tampone, per presidio o luogo dove è stato eseguito (in presidio ospedaliero, a casa, per strada, ecc.) e questo sia per coloro che sono ricoverati sia per coloro che sono in isolamento domiciliare. E inoltre sarebbe importante distinguere tra le morti quante sono avvenute mentre le sfortunate persone erano ricoverate con sintomi, in terapia intensiva o in isolamento domiciliare.

2. Analisi grafiche preliminari: time plots e log plots

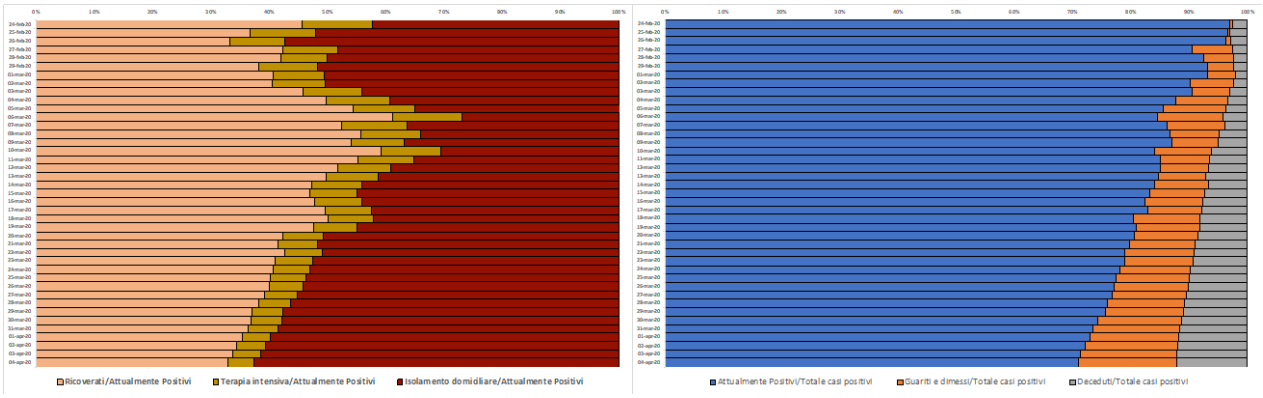
Accettata la carenza dei dati che comunque sono utili per aver dei "segnali" sulla evoluzione dei fenomeni, ecco le analisi grafiche preliminari che abbiamo fatto, come molti altri, sui dati pubblicati giornalmente dal DPC che presentiamo qui in sintesi dato il poco spazio a disposizione.

Una volta predisposte le serie storiche per l'Italia nel complesso e per le singole regioni, ci è sembrato opportuno calcolare, in primo luogo, le quote percentuali giornaliere di ricoverati con sintomi, di cui in terapia intensiva, e di coloro che sono in isolamento domiciliare sul totale degli attualmente positivi, come pure le quote percentuali degli attualmente positivi, totale guariti/dimessi e deceduti, sul totale dei casi.

I risultati sono interessanti perché queste quote sono diverse da regione a regione e ciò dipende anche dallo stato di avanzamento del virus e della rilevazione di tamponi.

A livello nazionale si vede dalla Figura 1, che la distribuzione degli attualmente positivi, negli ospedali o in isolamento domicilia mette in evidenza la sempre maggiore importanza dei soggetti attualmente positivi che si trovano in isolamento domiciliare e la quota dei dimessi/guariti rispetto al totale dei casi positivi accumulato nel tempo.

Fig.1. Distribuzioni percentuali degli attualmente positivi (a sinistra) del totale dei casi positivi (a destra)

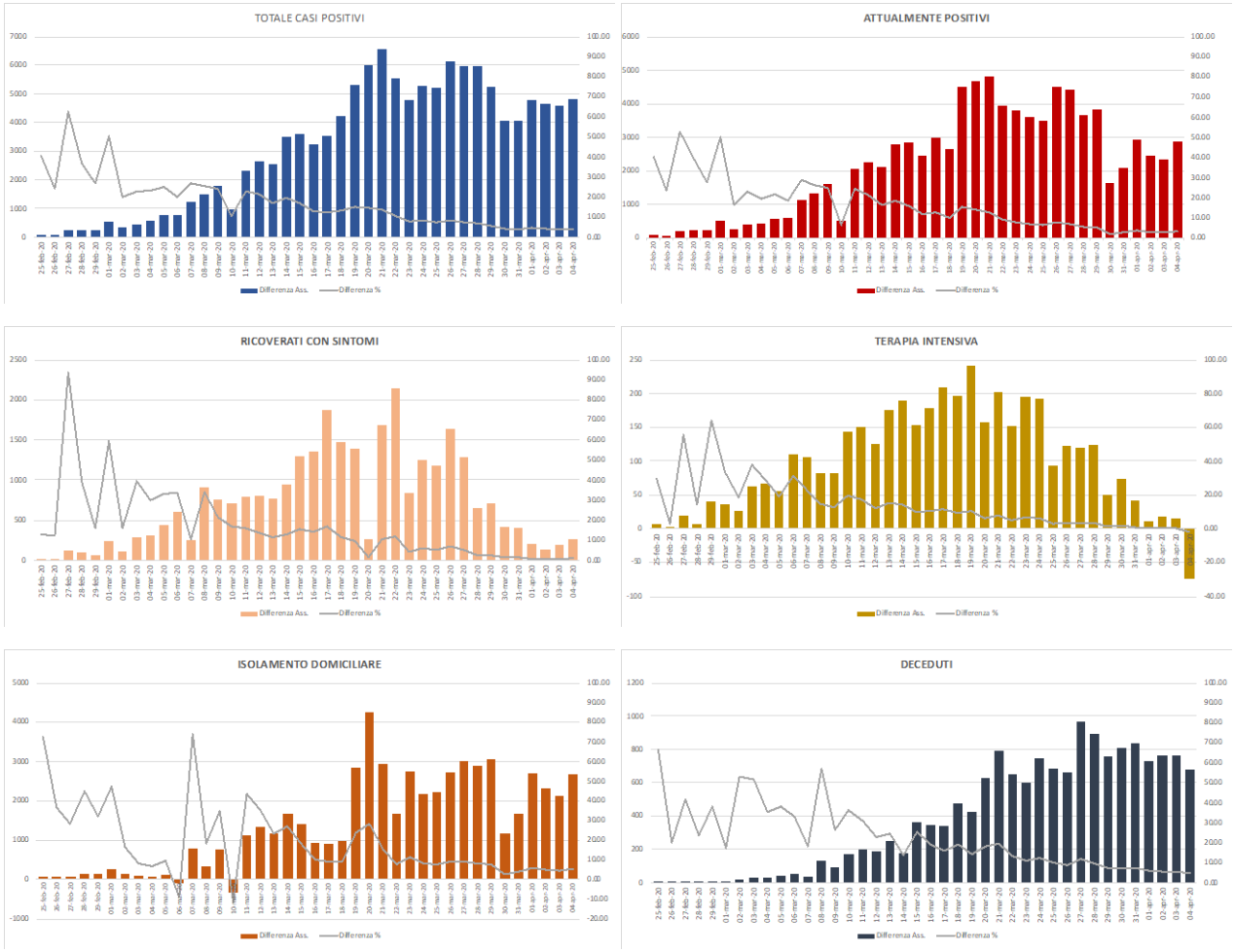


Per mettere in evidenza l'evoluzione dei fenomeni abbiamo calcolato le loro variazioni assolute e percentuali giornaliere che riportiamo nei grafici della Fig.2 per l'Italia, dove le variazioni assolute sono le colonnine e le variazioni relative sono indicate con la linea spezzata.

Qui presentiamo soltanto alcuni commenti di carattere generale, anche perché i grafici sono di immediata e facile lettura. Tuttavia per una adeguata interpretazione degli andamenti occorre tenere presente del lag temporale tra i vari fenomeni (le cui medie e mediane sono state stimate dall'ISS), nonché delle date di definizione dei provvedimenti governativi, e locali, per il contenimento del virus, e ovviamente del tempo necessario affinché si manifesti la loro efficacia.

Dall'esame dei singoli grafici, è chiaro che le variazioni assolute sono inizialmente aumentate e molto, e ora si stanno stabilizzando e in alcuni casi riducendosi. Le variazioni percentuali dopo molte oscillazioni iniziali sono quasi sempre nettamente diminuite e ora si stanno stabilizzando. E' ovvio che tutti speriamo che queste ultime diventino negative e le prime tendano rapidamente a zero.

Fig. 2 Variazioni assolute e variazioni relative: totale casi positivi, attualmente positivi, ricoverati con sintomi, terapia intensiva, isolamento domiciliare, deceduti



Come è noto, nella analisi descrittiva preliminare delle serie temporali, il grafico che mette meglio in evidenza l'evoluzione di fenomeni che crescono secondo una legge esponenziale modificata (cioè con un asintoto) è quello cosiddetto semilogaritmico, che riporta in ascissa il tempo e in ordinata il logaritmo del fenomeno. Inseriamo qui, nelle Figure 3-4-5, una esemplificazione riguardante il totale degli attualmente positivi (Serie 1 nei grafici sottostanti) e dei deceduti (Serie 2) con riguardo all'Italia e ad alcune regioni diciamo più significative per il numero di attualmente positivi e di deceduti.

Come si può rilevare i grafici sono molto esplicativi mettendo in evidenza il diverso inizio della diffusione del virus e la sua differente esplosione nei tassi di sviluppo e la sua direzione verso l'asintoto, quando il fenomeno non dovrebbe fortunatamente aumentare più o presentare aumenti insignificanti. Sono evidenti i differenti comportamenti sia delle serie del numero di soggetti attualmente positivi che di quelle dei deceduti nelle varie regioni italiane. Cosa che ha ben messo in evidenza Mauro Maltagliati nel suo ultimo articolo pubblicato su Neodemos con riferimento ai morti.

Come si vede, in alcune regioni il numero degli attualmente positivi tende all'asintoto, mentre il numero deceduti è un po' lontano dall'asintoto. In altre regioni accade il viceversa.

Naturalmente la descrizione dell'evoluzione di ciascuna serie deve essere fatta esaminandola molto attentamente per individuare i cambi di pendenza delle curve e gli eventuali punti di svolta. Necessariamente, in questa sede, i grafici sono di piccola dimensione, e non è nostra intenzione descrivere l'evoluzione delle singole serie; lasciamo al lettore di fare tale analisi. Tuttavia, in alcuni casi, anche allargando o restringendo i grafici non è facile individuare bene gli eventuali punti di svolta che sono molto importanti sia ai fini interpretativi che previsivi.

Si noti che in alcuni grafici regionali vi sono oscillazioni non giustificate se non per il fatto (come scrive l'ISS nei suoi reports) che gli stessi vengono aggiornati giornalmente da ciascuna regione anche se alcune informazioni possono richiedere qualche giorno per il loro inserimento. A volte la diminuzione dei casi può essere interpretata come un ritardo di notifica, successivamente recuperato. Piuttosto evidente è il caso dei dati relativi alla regione Marche per il 4 aprile 2020 (l'ultimo giorno della serie), dove gli attualmente positivi sono addirittura inferiori a quelli rilevati il 2 aprile 2020.

Fig.3. Evoluzione temporale Totale attualmente positivi e Deceduti in scala logaritmica: Italia, Lombardia, Emilia-Romagna e Piemonte

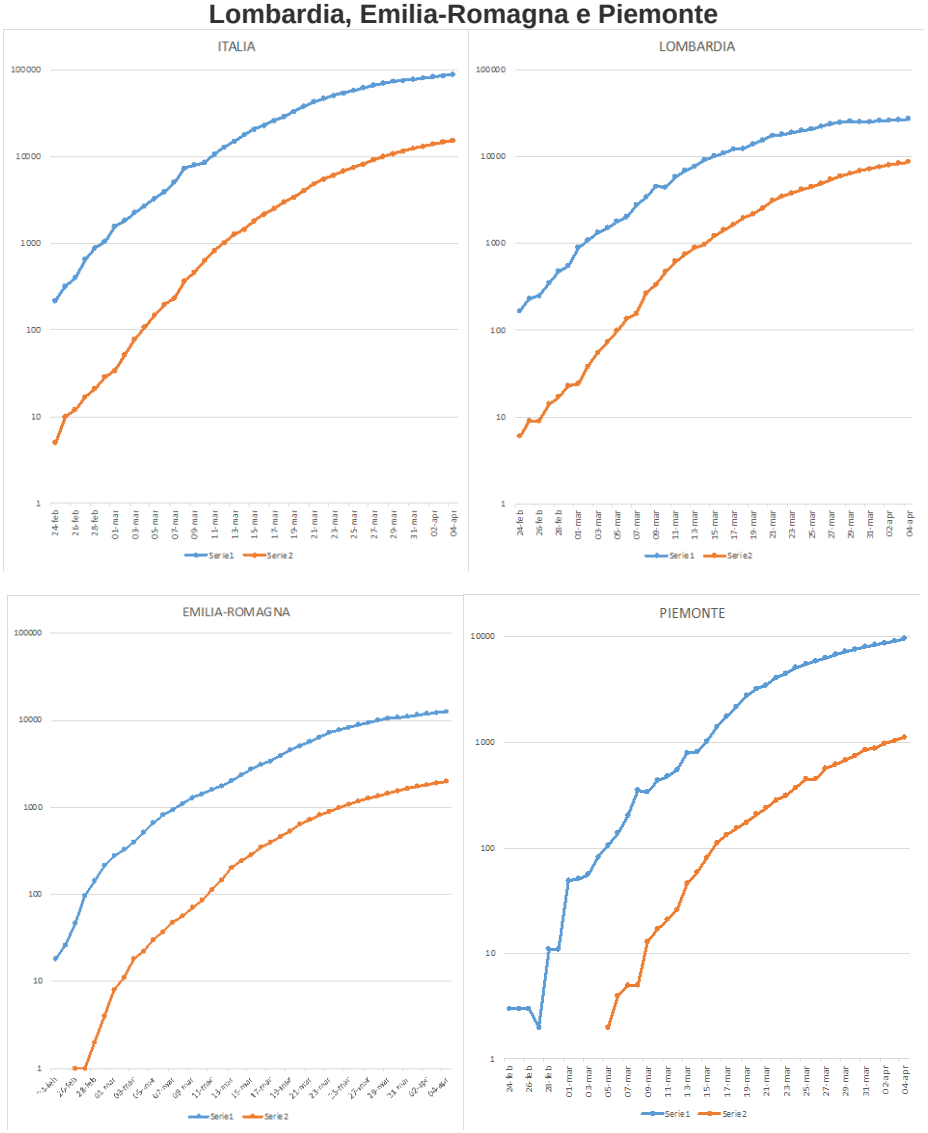


Fig.4. Evoluzione temporale Totale attualmente positivi e Deceduti in scala logaritmica: Veneto, Liguria, Toscana e Marche

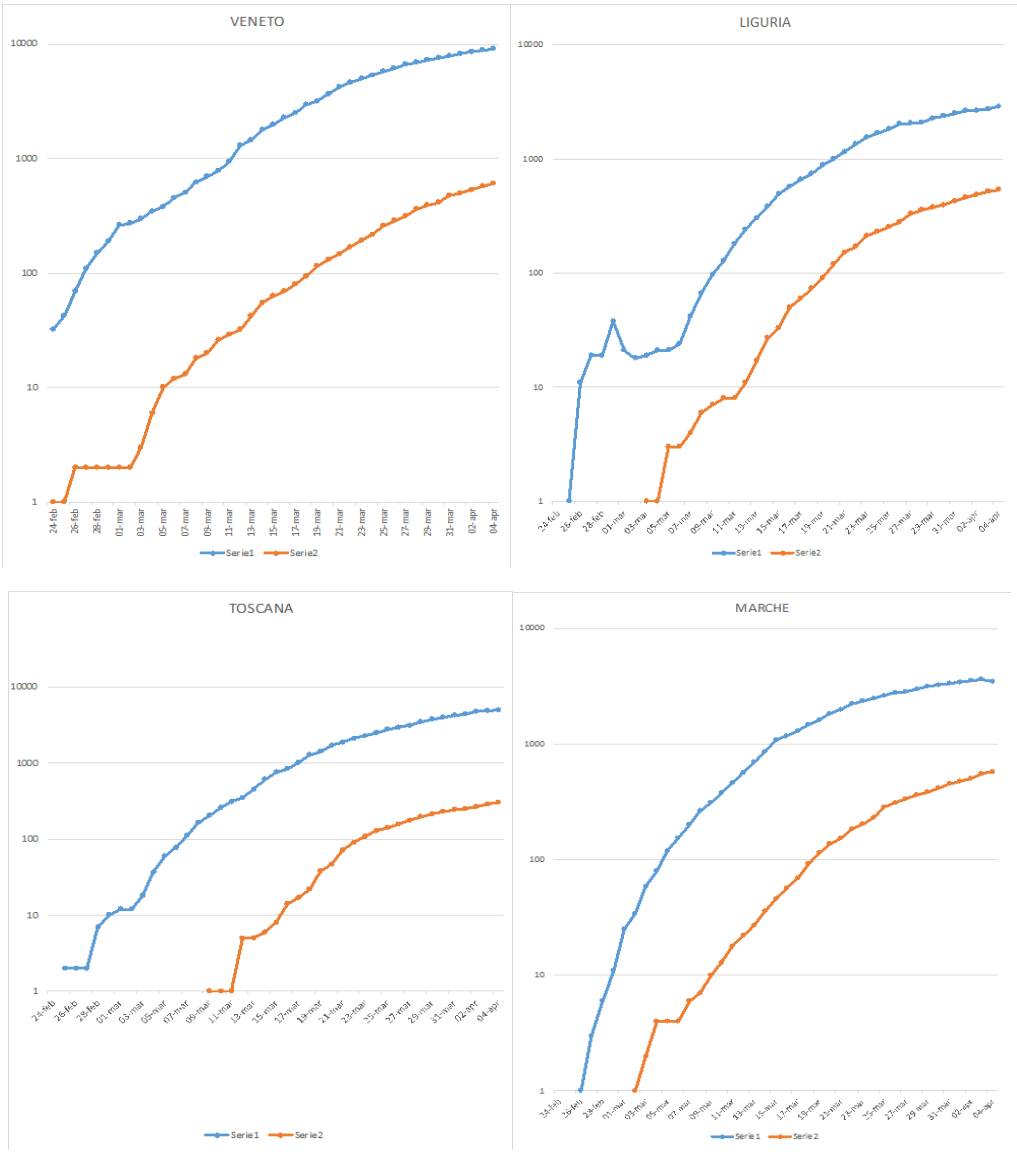
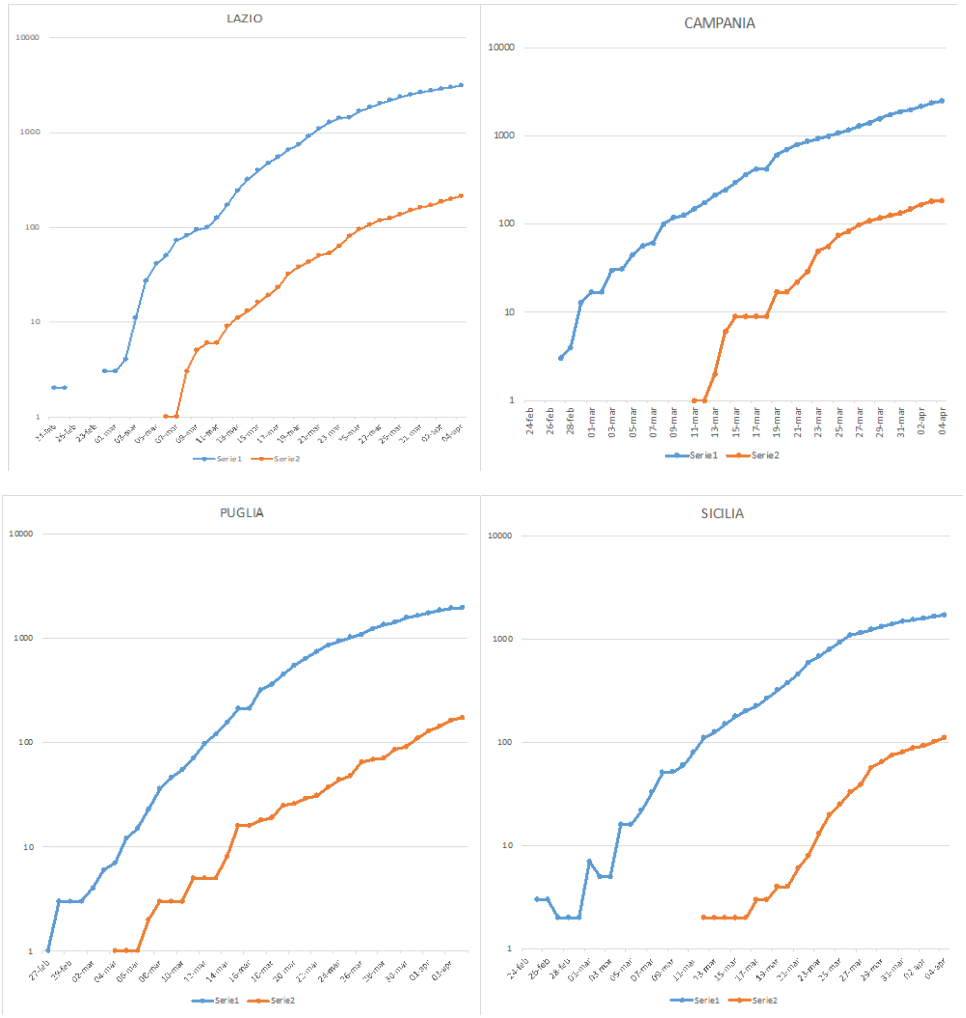


Fig.5. Evoluzione temporale Totale attualmente positivi e Deceduti in scala logaritmica: Lazio, Campania, Puglia e Sicilia



Molto interessante è anche il confronto tra le serie relative al numero degli attualmente positivi con quelle dei deceduti all'interno di ciascuna regione e tra regioni che mettono in evidenza se e dopo quanto tempo la curva dei deceduti ha un andamento simile a quella degli attualmente positivi, e quale è l'ampiezza della differenza numerica tra le due serie.

3. Una nuova rappresentazione grafica: log plots con adattamento di rette mobili

Per cercare di illustrare meglio i cambiamenti di pendenza e gli eventuali punti di svolta abbiamo perciò pensato di costruire dei grafici basati sui "Moving Least Squares Methods for scattered data". Sostanzialmente, nel nostro caso, si tratta di stimare, sulle serie dei logaritmi, i parametri delle rette basate su un certo numero di termini e poi di volta in volta ristimarle togliendo e aggiungendo un termine.

Indicando con t_i i tempi (nel nostro caso i giorni) e con log y_i i valori dei fenomeni rappresentati in ordinata (totale casi positivi, attualmente positivi, deceduti, ecc,) è necessario effettuare i seguenti passi:

1. iniziare stimando una retta su un certo numero <u>n</u> di termini (di giorni) e quindi per t_i con i= 1,…,n); $\log y_i^* = a + bt_i + \varepsilon_i$

2. stimare successivamente delle nuove rette, cosiddette, mobili togliendo il valore del primo termine della serie precedente stimata e includendo il valore del termine successivo ad n. La seconda stima della retta sarà quindi effettuata sempre per *n* termini per t_i con i= 2,…,n+1), e così via per gli altri termini della serie.
3. Effettuare la rappresentazione grafica delle rette (spezzate) mobili. Nel far questo si può decidere di rappresentare la serie stimata proiettandola per un certo numero *k* di termini, con *k*=1, 2…) quindi la proiezione sarà a partire dall'ultimo termine inserito nella stima della retta. Nella prima stima sarà a partire dal tempo t_{n+1}. Ovviamente è bene rimarcare che si tratta di "proiezioni" e non di previsioni, cioè di proiezioni che ci dicono soltanto quale sarebbe il valore di log y_i nei *k* tempi (giorni) successivi se la pendenza della curva rimanesse la stessa di quella stimata sugli *n* tempi precedenti.

I log plots con rette (segmenti) mobili saranno diversi a seconda della combinazione dei valori scelti per *n* e *k* , nonché delle caratteristiche di evoluzione di ciascun fenomeno. Con *n* grande le stime delle rette risentiranno meno di eventuali oscillazioni, ma ovviamente troppi termini limitano la capacità visiva del grafico, cioè si vede in ritardo quando c'è un vero cambiamento di pendenza o un punto di svolta; *k* deve essere necessariamente non grande e comunque inferiore ad *n*.

Come si vedrà nei grafici che presentiamo, esemplificativamente, quando la pendenza delle rette cambia spesso (anche perché la serie presenta oscillazioni) le spezzate rappresentate nel grafico sono distanti l'una dall'altra e disposte quasi a ventaglio; quando invece una pendenza rimane stabile per un certo periodo, numero di giorni, le spezzate sono molto più ravvicinate. Quando la pendenza cambia si evidenzia quasi un leggero o più grande angolo. Viene fuori un grafico che presenta più o meno facilità di lettura a seconda delle scelte precedenti.

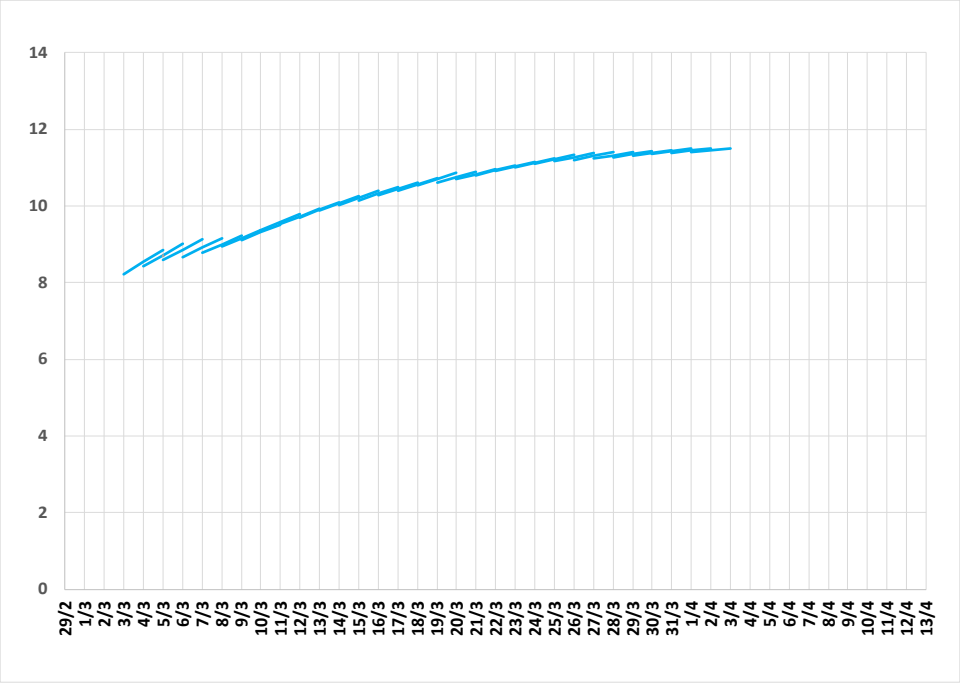
Non abbiamo trovato nessun programma che ci consentisse di fare le stime dei log plots con rette mobili. Mauro, essendo un espertissimo programmatore in particolare con Excel, ha prontamente risolto il problema delle stime e della preparazione dei grafici, creando la possibilità di farli scegliendo *n* e *k* (forse si tratta di una novità da copyright informatico?).

Gli esempi che riportiamo di seguito servono per far vedere l'utilità di questi grafici anche per verificare gli effetti dei provvedimenti di contenimento del virus, e come base per decisioni riguardanti l'evoluzione futura dei fenomeni oggetto di studio. Non diciamo che "parlano da soli", anzi a prima vista sembrano un po' complessi, ma tenendo conto di quanto sopra specificato e analizzandoli attentamente crediamo possano essere utilizzati con profitto. Per evidenziare i cambiamenti di pendenza e gli eventuali punti di svolta.

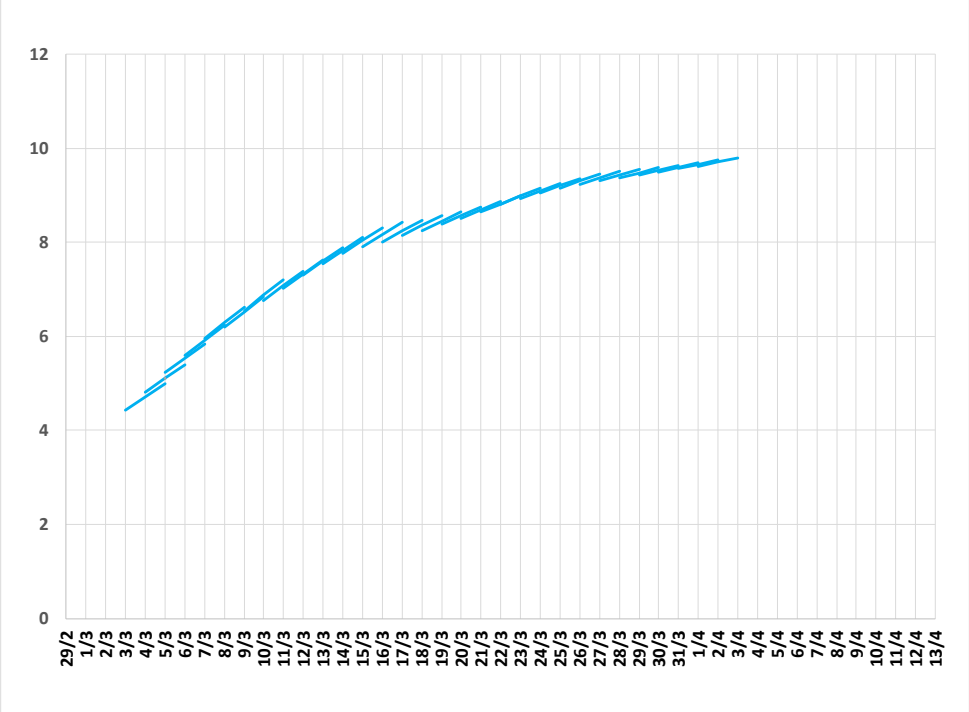
A titolo esemplificativo, riportiamo nella Fig. 6, i **Log Plots con rette mobili**, **per *n* = 7 e *k* = 3**, per gli attualmente positivi e per i deceduti, per l'Italia e la Toscana. Abbiamo scelto *n* = 7 considerando che per verificare se i fenomeni stanno mantenendo o modificando la loro evoluzione occorre almeno una settimana.

FIG. 6

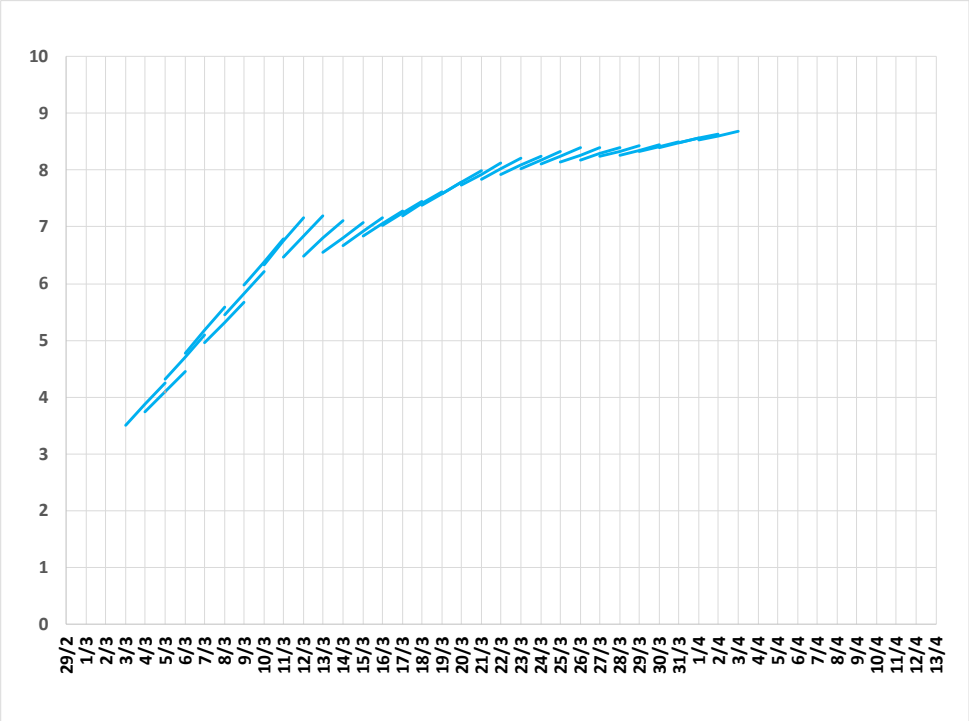
ITALIA ATTUALMENTE POSITIVI



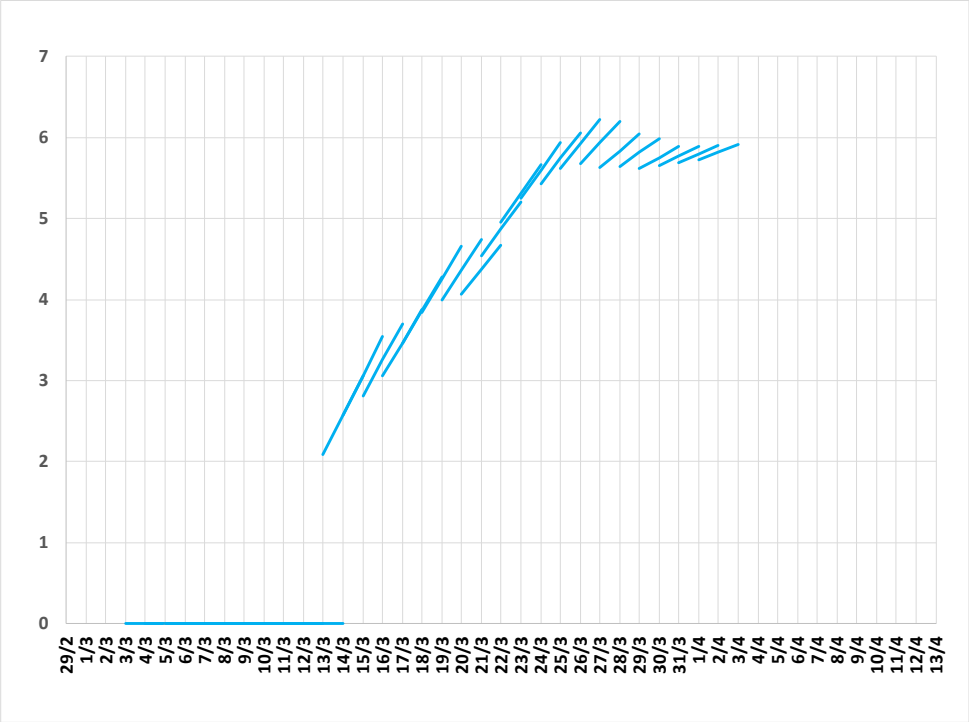
ITALIA-DECEDUTI



TOSCANA- ATTUALMENTE-POSITIVI



TOSCANA-DECEDUTI



I grafici all'inizio e in occasioni di oscillazioni della serie, come avviene per le serie della Toscana ed in particolare per quella di deceduti, non presentano una pendenza ben definita e hanno una rappresentazione, per così dire a ventaglio, ma successivamente gli spazi tra i segmenti di retta si restringono e la "lettura" diviene più facile, comunque sempre tenendo presente che la retta più in basso di ogni fascio è l'ultima rappresentata di ciascun periodo. Il cambiamento di pendenza nella evoluzione esponenziale del fenomeno

è chiaro quando vi è un periodo abbastanza lungo di spezzate (rette) che confermano il cambiamento (che quindi non è conseguenza di un solo dato che può essere anche anomalo).

4. Alcune informazioni da utilizzare e analizzare per fare corretti confronti tra regioni

Infine, come hanno già chiarito Mauro ed altri ricercatori, è evidente che nel fare i confronti tra regioni e per valutare la diffusione del virus, i tassi di mortalità e di letalità devono essere standardizzati, in primo luogo rispetto all'ammontare della popolazione, ma possibilmente anche per tener conto della struttura per età della popolazione, della struttura del numero di morti per causa, dello stato di salute della popolazione (che è indicativo delle possibili patologie pregresse di coloro che purtroppo poi sono più facilmente contagiati dal virus). Elementi tutti che sono molto diversi da regione a regione. Come forse, sempre nei confronti si dovrebbe tenere conto anche della densità della popolazione per km^2, del numero di centri abitati (o comuni) con più di 15.000-20.000 abitanti, della presenza di imprese e di esercizi sanitari e commerciali e così via.

Per fortuna, come ben sanno, gli epidemiologi, i demografi e coloro che si occupano di questi problemi l'Istat pubblica regolarmente moltissimi dati di qualità a livello territoriale. Ad esempio, nell'Annuario Statistico Italiano del 2019, nel capitolo 3, che riguarda la Popolazione e gli indicatori demografici, risulta, come è stato rilevato da altri, che la popolazione con 65 anni e più rappresenta una quota della popolazione maggiore nelle regioni del Nord rispetto a quelle del Sud (agli estremi ci sono la Liguria, con il 28,5% e la Campania con il 18,8%). E nel capitolo 4, che riguarda la Sanità e salute, si trovano informazioni sul ricorso alla ospedalizzazione, sulle malattie del sistema circolatorio, la graduatoria delle principali cause di morte e i quozienti di mortalità per età, lo stato di salute e la presenza di malattie croniche. Praticamente tutte queste informazioni mettono in evidenza la peggiore situazione di salute dei residenti nelle regioni del Nord rispetto a quelli residenti nelle regioni del Sud. Il che potrebbe anche far sperare che nel Sud il corona virus abbia una espansione minore, sempre che tutti rispettino le misure di contenimento adottate.