Luigi Biggeri*, Mauro Maltagliati* e Luca Secondi°

Rubrica su "I dati sul Covid-19 in Italia a livello regionale: interpretazione e validità dei dati, analisi grafiche preliminari ed esplorative"

*Università di Firenze; ° Università della Tuscia

Struttura

1. **Premessa: obiettivi e contenuti della Rubrica**
2. **Significato e validità dei dati disponibili**
3. **Analisi grafiche preliminari delle serie storiche**
   - **3.1 La composizione dei casi positivi totali e degli attualmente positivi**
   - **3.2 Grafici a barre**
   - **3.3 Time plots in scala naturale e in scala logaritmica**
   - **3.4 I confronti tra le regioni dell'intensità del covid-19**
4. **Una nuova rappresentazione grafica: log plots con adattamento di rette mobili**
5. **Il modello Holt-Winters con trend, per descrivere l'andamento delle serie e fare proiezioni a brevissimo termine**

Appendice

Preparazione del Data Base e di programmi di elaborazione flessibili per effettuare tutte le analisi: Manuale per l'uso del file autoDB.xlsm

1. Premessa: obbiettivi e contenuti della Rubrica

Da tempo, come stanno facendo molti altri ricercatori di varie discipline e in particolare di quelle statistiche, anche noi analizziamo i dati disponibili sulla evoluzione del Covid-19 per cercare di interpretarne l'evoluzione e ottenere indicazioni sul loro possibile sviluppo.

Una importante e completa informazione sulle analisi e discussioni sul tema del Covid-19 si trova nella Rassegna-Raccolta predisposta dalla Società Italiana di Statistica (SIS) dove i lavori sono classificati in tre grandi gruppi: "Web, App, Interviste e Video", "Numeri, Grafici e Articoli Divulgativi" e "Modelli Statistici e Previsioni"(http://www.sis-statistica.it/ita/20/index.php?p=9996). Utile è anche il riferimento al Forum SIS al quale molti soci e non soci hanno inviato links a lavori propri o di altri, aprendo una vivace e interessante discussione degli stessi (https://www.sis-statistica.it/ita/3207/Forum%20di%20posta%20elettronica).

Da tutti questi lavori, emerge che i principali obbiettivi delle analisi possono essere:

(i)Analisi e valutazione delle cause e/o fattori che determinano, o quanto meno influenzano, il fenomeno (con modelli ad equazioni strutturali? Cfr. Bruno Chiandotto, in Rivista SIS, Statistica & Società, 4.4.2020);

(ii) Analisi degli "impatti" del covid-19 dal punto di vista sanitario, sociale ed economico (con modelli di micro e macro simulazione);

(iii) Analisi dell'evoluzione attuale e della previsione futura del fenomeno nei suoi vari aspetti sanitari, per individuare i punti di aumento, di attenuazione e di svolta del suo sviluppo, e l'avvicinarsi all'asintoto (cioè quando gli incrementi tendono ad essere insignificanti), nonché per valutare gli effetti delle misure di contenimento della diffusione del virus.

In questo ambito la maggior parte delle analisi hanno riguardato l'impiego di adeguati modelli statistici per effettuare la previsione dello sviluppo del covid-19. I modelli impiegabili e impiegati sono molteplici come è facile rilevare dalle proposte e dalle analisi svolte riportate nel forum SIS e, in parte, nella rivista Neodemos (www.neodemos.info/articoli/covid-19), nel sito del Florence Centre for data Science, dove sono ci sono anche i collegamenti ai gruppi italiani di ricerca sul Covid19 (https://datascience.unifi.it/index.php/projects/covid-19-fds/) e in una pagina dedicata al virus da parte del nostro Dipartimento, Disia (https://www.disia.unifi.it/art-423-coronavirus-disia.html) . Poiché la diffusione dei fenomeni epidemici è, in genere, caratterizzata da una fase iniziale di sviluppo modesto e da una successiva fase di accelerazione che poi si attenua con il passare del tempo con incrementi assoluti che si riducono fino praticamente ad annullarsi quando il fenomeno è vicino all'asintoto (o livello di saturazione), i modelli statistici più frequentemente utilizzati sono l'esponenziale modificata, il logistico e, con maggior frequenza, varie versioni dei modelli SIR.

Tutte le analisi menzionate hanno interesse ed utilità a livello nazionale, ma soprattutto a livello locale di regione e provincia, viste le diverse date di inizio della diffusione del virus nelle varie aree territoriali del Paese, le diverse modalità (intensità e velocità di diffusione del virus) nelle differenti aree, nonché il loro diverso ammontare della popolazione e le diverse caratteristiche (struttura per

età, ecc.) della stessa. A nostro avviso sarebbe, anzi, importante effettuare le analisi a livello dei 610 Sistemi locali del Lavoro (SLL), che sono aree territoriali identificate da un insieme di comuni legati tra loro dai flussi degli spostamenti quotidiani per motivi di lavoro e di studio (Cfr. Istat, 30.10.2019).

Tutti sappiamo che per poter condurre adeguatamente le suddette analisi occorre:

1. conoscere le caratteristiche e la qualità dei dati a disposizione;

2. disporre di un adeguato data base con tutte i dati rilasciati dal Dipartimento della Protezione Civile (DPC);

3. effettuare le indispensabili analisi preliminari, soprattutto grafiche (grafici di composizione, time plots e log plots) per evidenziare l'evoluzione delle differenti serie temporali fino al momento attuale (o di interesse);

4. effettuare qualche semplice analisi esplorativa sulla possibile evoluzione futura delle serie esaminate;

5. disporre di programmi di elaborazione flessibili per effettuare con rapidità tutte le analisi che interessano.

Questo è ciò che cerchiamo di fare in questa Rubrica, che sarà aggiornata due volte a settimana, sperando che sia di utilità per potenziali i lettori.

In questo momento, il virus sembra che sia indebolito e che abbia perso la sua forza sia di diffusione, sia di intensità su coloro che sono di nuovo contagiati. Infatti, in diverse regioni i nuovi casi non si traducono in aumento del numero degli attualmente contagiati e la maggior parte di loro sono nella situazione di isolamento domiciliare. Questa Rubrica potrebbe perciò apparire tardiva o addirittura inutile. Ma così non è, perché la situazione non è già tutta rosea e il processo di uscita dal covid-19 è e sarà diverso da regione a regione. Il nostro obiettivo è perciò di monitorare, anche nel periodo della cosiddetta Fase 2, gli specifici andamenti delle varie casistiche del virus nelle diverse regioni, cercando di individuare "campanelli di allarme" riguardanti la eventuale interruzione del miglioramento della situazione o la ripresa di forza del virus o, infine, la nascita di nuovi focolai.

2. Significato e validità dei dati disponibili

Quasi tutti i ricercatori sostengono che i dati disponibili sul Covid-19 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. Molto utile al riguardo è un recente interessante contributo di F. De Luca dal titolo "I dati sul Covid19: maneggiare con prudenza" pubblicato in Neodemos il 22 aprile 2020.

In questa sede il nostro contributo è diverso perché cerchiamo di valutare il significato e la validità dei dati disponibili in base a quanto è scritto nei documenti ufficiali degli enti che li diffondono.

Come è noto, la interpretazione dei dati e delle informazioni statistiche dipende dalle definizioni operative del fenomeno, delle unità e della popolazione, delle variabili e modalità oggetto di rilevazione, nonché dalle modalità di rilevazione (che possono comportare errori di campionamento e/ di osservazione).

Il Dipartimento della Protezione Civile (DPC) diffonde tutti i giorni le seguenti informazioni sul Covid-19 a livello regionale e nazionale sul numero di:

- Tamponi,
- Totale di Casi Positivi al nCoV di cui:
  - Attualmente Positivi, Dimessi/Guariti, Deceduti
- Totale degli Attualmente Positivi distinto, in relazione alla gravità della situazione clinica del contagiato e al luogo dove si trova, in:
  - Ricoverati con Sintomi, Ricoverati in terapia intensiva, in Isolamento Domiciliare

Le definizioni dei casi e la gestione-trasmissione delle informazioni

Le definizioni delle varie casistiche non vengono fornite in occasione della diffusione delle informazioni forse dando per scontato il loro significato, ma così non è come vedremo tra poco, e neppure sono forniti i cosiddetti Metadati (cioè le informazioni che permettono ai dati i "parlare" di farci comprendere appieno il loro significato (glossari, classificazioni, definizioni, metodologie di rilevazione, ecc.).

L'Istituto Superiore di Sanità (ISS) nelle sue analisi e pubblicazioni sul Covid-19 in particolare nella Infografica e nei Reports bisettimanali - che sono certamente ben fatti e facilmente leggibili - fornisce molte informazioni che aiutano a meglio comprendere, almeno in parte, il significato delle informazioni statistiche pubblicate, ma purtroppo ne mancano molte altre e soprattutto anche l'ISS non fornisce i Metadati.

La critica principale da parte degli utilizzatori di queste statistiche riguarda soprattutto i dati sul totale dei Casi Positivi accertati e il totale degli Attualmente Positivi e deriva dal fatto che i primi chiaramente non rappresentano tutti i contagiati poiché dipendono dal numero di tamponi effettuati e non includono gli asintomatici (che non presentano sintomi specifici pur essendo infettati) e i pauci-sintomatici (cioè quelli che presentano pochi sintomi); 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 o con corona virus che sembrava essere il dato più attendibile è messo in dubbio, come risulta dalle polemiche riportate dai media nelle ultime settimane.

L'ISS illustrando il Sistema di sorveglianza nazionale (https://www.epicentro.iss.it/coronavirus/sars-cov-2-sorveglianza), che gli è stato affidato dal 22 gennaio 2020 e che raccoglie le segnalazioni delle Regioni attraverso una piattaforma web dedicata, fornisce la definizione di caso di infezione da Coronavirus e le modalità per la diagnosi e la raccolta dei campioni clinici. L'ISS scrive testualmente: "la definizione di caso (sospetto, probabile e confermato) si basa sulle informazioni attualmente disponibili e può essere rivista in base alla evoluzione della situazione epidemiologica e delle conoscenze scientifiche disponibili", concludendo che "Per maggiori informazioni sulla definizione di caso si invita a fare riferimento all'ultima Circolare ministeriale disponibile". In effetti tale definizione deve necessariamente tener conto ed essere modificata in base alle indicazioni del gruppo di esperti (Gruppo permanente di lavoro costituito presso il Consiglio Superiore di Sanità) e delle organizzazioni internazionali, in particolare dell'Organizzazione Mondiale della Sanita (OMS).

Esaminando i tanti decreti e circolari si rileva che le varie definizioni sono state precisate più volte a partire dal 22 gennaio 2020.

Un primo richiamo è stato fatto, in tale data, dalla Direzione Generale della Prevenzione Sanitaria (Ministero della Salute) con riguardo alle segnalazioni che provenivano dalla Cina su "Polmonite da

nuovo coronavirus", fornendo una prima definizione provvisoria di caso e dei sintomi che si potevano manifestare nei viaggiatori provenienti dalla Cina e sulla modalità di gestione dell'eventuale paziente.

Una successiva circolare del 31 gennaio 2020, ha fatto riferimento ai Potenziali casi di coronavirus e alla loro relativa gestione, ma anche alla gestione delle persone che erano state in contatto con i nuovi casi individuati (con contatti definiti a rischio), nonché alle raccomandazioni in caso di isolamento domiciliare fiduciario.

Il 22 febbraio, una Circolare del Ministero della salute ha fornito "Nuove indicazioni e chiarimenti". Considerando l'evoluzione della situazione epidemiologica e le nuove evidenze scientifiche, sono state modificate le definizioni di caso, di "contatto stretto" e le modalità di notifica dei casi.

I casi sono distinti in: (i) caso sospetto (una persona con infezione respiratoria acuta che ha richiesto o meno il ricovero in ospedale); (ii) caso probabile (un caso sospetto il cui risultato ai test è dubbio o inconcludente); (iii) caso confermato (con una conferma di laboratorio per infezione da SARS-COV-2, indipendente dai segni e dai sintomi clinici).

Per quanto riguarda le notifiche i medici di medicina generale che vengono a conoscenza di un caso sospetto devono attuare varie misure precauzionali: (i) per il paziente sintomatico (devono segnalare il caso al 112/118 e alla unità di malattie infettive); (ii) per il paziente paucisintomatico/contatto stretto negativo al test (devono predisporre assistenza domiciliare e/o segnalare il caso al dipartimento di prevenzione della ASL per la sorveglianza attiva); (iii) per il soggetto riscontato positivo al tampone e al momento asintomatico (devono prescrivere la quarantena domiciliare con sorveglianza attiva per 14 giorni). E' evidente che la decisione è lasciata al medico, che può comportarsi diversamente nelle varie aree del Paese, anche in relazione ai posti disponibili negli ospedali, sia in generale che in terapia intensiva, dell'area.

Un aggiornamento della suddetta circolare è stato diffuso il 27 febbraio, anche al fine di garantire la rapida ed efficace rintracciabilità dei contatti. In particolare il caso sospetto è definito Caso sospetto di Covid 19 che richiede esecuzione di test diagnostico. Questa circolare specifica meglio i casi sospetti che richiedono l'esecuzione di test diagnostico (il cui risultato dipende però dal tipo di test e dalle caratteristiche delle persone sottoposte test, ndr) e la definizione di "contatto stretto".

Sempre il 27 febbraio una circolare del Ministero della salute, fa sua la seguente conclusione del Gruppo permanente di lavoro sui criteri per sottoporre soggetti clinicamente asintomatici alla ricerca d'infezione da coronavirus: "…il rischio di trasmissione in fase asintomatica/prodomica sembra essere basso o molto basso….(riferendosi a quanto avvenuto in Cina, ndr) ..il gruppo di lavoro ritiene che trasferire un numero elevato di campioni che risulterebbero poi essere, nella larghissima maggioranza dei casi, negativi a laboratori di virologia non sia scientificamente giustificabile.. e condivide le indicazioni emanate dal Ministero della salute raccomandando che l'esecuzione dei tamponi sia riservata ai soli casi sintomatici di ILI ((influenza-like Illness) non attribuibili ad atra causa …oltre che ai casi sospetti di Covid 19.

Il 28 febbraio un documento su la definizione di paziente guarito da Covid 19, precisa che è colui il quale risolve i sintomi dell'infezione e che risulta negativo in due test consecutivi effettuati a distanza di 24 ore; precisando però che non si può escludere che, in una percentuale dei casi, ad alcuni campioni venga attribuito un risultato non idoneo.

Infine, in una circolare del 9 marzo vengono forniti ulteriori aggiornamenti (precisazioni) della definizione di caso e per la certificazione di decesso a causa di Covid-19 che dovrà essere accompagnata da parere dell'ISS. A tale scopo le cartelle cliniche dei pazienti deceduti, positivi da Covid-19 (anche per coloro che si trovano a casa? ndr) e le schede di morte Istat recanti le cause di decesso dovranno essere inviate all'ISS attraverso il sito di sorveglianza.

Tutte queste definizioni dipendono certamente (e in parte sono giustificate) dal fatto che inizialmente si trattava di rilevare un fenomeno "sconosciuto", negli esatti sintomi e modalità di diffusione. Tuttavia limitano o rendono incerta la comparabilità di vari casi rilevati all'inizio con quelli rilevati quando il sistema di sorveglianza si è stabilizzato. Destano in particolare preoccupazione i dati sui casi positivi, sulla data di inizio dei sintomi, sulla gravità clinica dei pazienti, sulle caratteristiche di quelli in isolamento domiciliare e sui guariti, nonché la mancanza dei dati sugli asintomatici.

Questa chiamiamola "carenza" dei dati diffusi dalla DPC è, in parte 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. Non sono disponibili per il motivo a segnalato. In realtà, l'ISS effettua le analisi sulle cartelle cliniche e nel report del 23 aprile 2020 ha scritto che dallo stato clinico dei pazienti disponibile per 57.048 casi (su oltre 177.143 infettati) gli asintomatici risultano essere il 13,1% ed i pauci-sintomatici il 17,2 %, 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.

La individuazione di soggetti a cui fare i tamponi

Il problema più generale però è la quantificazione dei positivi che dipende 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 costo, 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 vari ricercatori (si veda il Forum SIS) e in parte è stato fatto, 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.

<u>Insufficienza delle informazioni disponibili per la ricerca scientifica</u>

Comunque, per una interpretazione corretta delle serie storiche e per fare adeguate analisi sono necessarie molte informazioni che non sono a nostra disposizione, come segnalato dal presidente della SIS nella lettera inviata al presidente dell'ISS e anche nella sua intervista al Mattino di Napoli Occorre certamente che la comunità scientifica possa disporre di dati elementari anonimi con informazioni sull'età del soggetto, sulla data dei sintomi e del tampone, sulle patologie preesistenti, sulla gravità clinica dei pazienti, sulle caratteristiche di quelli in isolamento domiciliare e sui guariti, sul 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: di queste ultime non sappiamo praticamente niente!

Date le modalità di rilevazione i dati attuali non sono certamente adeguati per svolgere ricerche epidemiologiche sulla stima statistiche della proporzione di persone che sono state infettate, sulle modalità con la quale la "trasmissione è avvenuta e su quante persone hanno sviluppato una risposta anticorpale specifica utilizzabile come tracciante della circolazione virale. Proprio per questo motivo, il Ministero della Salute e l'Istat stanno ora organizzando una indagine campionaria nazionale su circa 150.000 persone, che dovrebbe partire tra pochi giorni, alla quale la SIS collabora per la parte metodologica (si veda il sito web della SIS dedicato al Covid-19).

3. Analisi grafiche preliminari delle serie storiche

Accettata la carenza dei dati che comunque sono utili per aver dei "segnali" sulla evoluzione dei vari aspetti del covid-19, presentiamo le analisi grafiche preliminari che abbiamo fatto, come molti altri, e che riteniamo più adeguate per illustrare i dati pubblicati giornalmente dal DPC.

Siamo ben consci che le analisi preliminari ed esplorative delle serie storiche non si limitano soltanto alla costruzione dei grafici e che quelli che presentiamo non sono gli unici che si possono costruire. Anche perché l'utilità di un grafico dipende dagli obiettivi per cui lo costruiamo (da ciò che vogliamo mettere in evidenza), dalle conoscenze a priori che abbiamo sui "meccanismi" che generano il fenomeno che stiamo osservando e dal suo attuale andamento. Senza contare che abbiamo a che fare con un Virus molto complesso mai visto prima, di cui al momento non sappiamo niente o molto poco.

Riteniamo però che una rappresentazione grafica se esaminata attentamente (e non superficialmente) costringe a vedere e a cercare di capire i movimenti e le oscillazioni delle serie fornendo informazioni utili anche per l'applicazione di modelli descrittivi e previsivi che non possono essere astratti o automatici. D'altra parte gli statistici - la cui scienza è indispensabile per prendere decisioni in condizioni di incertezza - possono e devono cercare di rilevare e analizzare tali oscillazioni, senza però dimenticare che queste analisi vanno condotte con adeguati scambi culturali con gli scienziati che hanno competenze specifiche e consolidate nell'ambito dello studio delle epidemie.

Noi ci limitiamo a suggerire le rappresentazioni grafiche che ci sembrano più utili, a livello regionale per: (i) analizzare lo stadio di diffusione e di intensità del virus, anche in relazione ad eventuali strategie terapeutiche diverse da parte dei servizi sanitari locali; (ii) individuare le diverse fasi di sviluppo e gli eventuali punti di svolta delle varie serie storiche sul Covid-19, evidenziando, se possibile, eventuali cambiamenti nell'ultimo periodo di tempo (ultimi giorni). Come sarà illustrato nel paragrafo 6, forniamo anche la possibilità, in automatico, di usare anche i valori per la macro area Italia senza Lombardia e di effettuare confronti tra coppie di regioni nello stesso grafico.

Alcune avvertenze per una loro corretta lettura e interpretazione.

In primo luogo, occorre tenere presente che le date di inizio di diffusione del coronavirus non sono state le stesse nelle varie regioni e questo si rileva anche dall'inizio con valori positivi delle serie storiche dei vari aspetti del virus.

In secondo luogo, può essere interessante ricordare le date dei più importanti provvedimenti adottati con DPCM e Ordinanze per il contenimento del covid-19:

- 23 febbraio, decreto legge del Consiglio dei Ministri, con misure per il divieto di accesso e allontanamento nei comuni dove sono presenti focolai del coronavirus e la sospensione di manifestazioni ed eventi;
- 25 febbraio, Dpcm attuativo del precedente decreto;
- 1 e 4 marzo, due Dpcm, con estensione dei divieti a tutto il territorio nazionale, anche di congressi riunioni, manifestazioni ecc.;
- 8 e 9 marzo, due Dpcm, per limitare tutti i movimenti della popolazione non essenziali (chiamati "io resto a casa");
- 11 marzo, Dpcm che chiude le attività commerciali non di prima necessità;
- 22 marzo, Ordinanza 2 per limitare gli spostamenti delle persone in comuni diversi da quello in cui si trovano, salvo che per comprovate esigenze lavorative, di assoluta urgenza ovvero per motivi di salute;

- 22 marzo, Dpcm che ha previsto per tutto il territorio nazionale la chiusura delle attività produttive non essenziali o strategiche fino al 3 aprile (con proroga alla stessa data anche dei precedenti divieti);
- 1 aprile, Dpcm, con proroga delle misure fino al 13 aprile;
- 10 aprile, Dpcm che ha prorogato tutte le misure fino al 3 Maggio.

Ricordiamo che anche alcune regioni tra le quali ad esempio la Basilicata e la Campania hanno iniziato fin dal 23 febbraio 2020 a pubblicare Ordinanze per estendere e precisare le limitazioni nell'ambito regionale.

Infine, si avvertono i lettori che nel testo i grafici non saranno riportati per tutte le regioni, ma soltanto per alcune a titolo esemplificativo. Chi lo desidera può costruire i dati che gli interessa esaminare usando l'applicativo "autoDB.xlsm" (su base Microsoft Excel) in cui è integrato un database che si aggiorna automaticamente attivando una macro (nell'Appendice si descrive il suo funzionamento). I grafici sono riportati come immagine e quindi possono essere allargati o allungati a piacimento.

3.1 La composizione dei casi positivi totali e degli attualmente positivi

Ci è sembrato opportuno calcolare, in primo luogo, le quote percentuali giornaliere del totale dei casi positivi (scomposti in attualmente positivi, dimessi/guariti e deceduti), e quelle del totale degli attualmente positivi (scomposti tra i ricoverati con sintomi, in terapia intensiva, e in isolamento domiciliare).

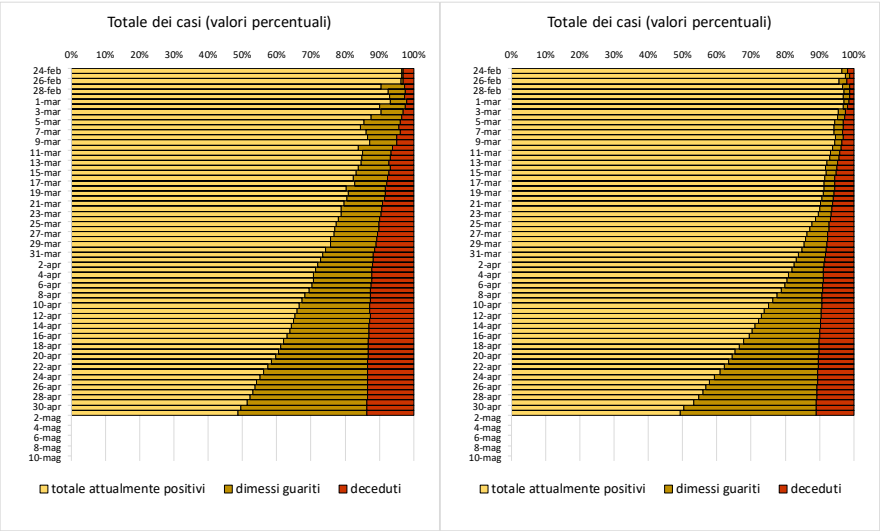
I risultati sono interessanti perché come si vede dai grafici della Fig. 1, queste quote sono diverse da regione a regione e ciò dipende anche dallo stato di avanzamento e di intensità del virus, ma anche dalla numerosità della rilevazione di tamponi.

Dalla figura si vede che a livello nazionale i casi positivi accumulati nel tempo si sono sempre più risolti positivamente (speriamo!) e purtroppo in aumento del numero di deceduti. Al tempo stesso gli attualmente positivi, sono nell'ultimo periodo sempre meno negli ospedali (ricoverati per sintomi o in terapia intensiva) e sempre più in isolamento domiciliare, mettendo evidenza, almeno sembra(?), che ultimamente il virus agisce facendo minori danni.

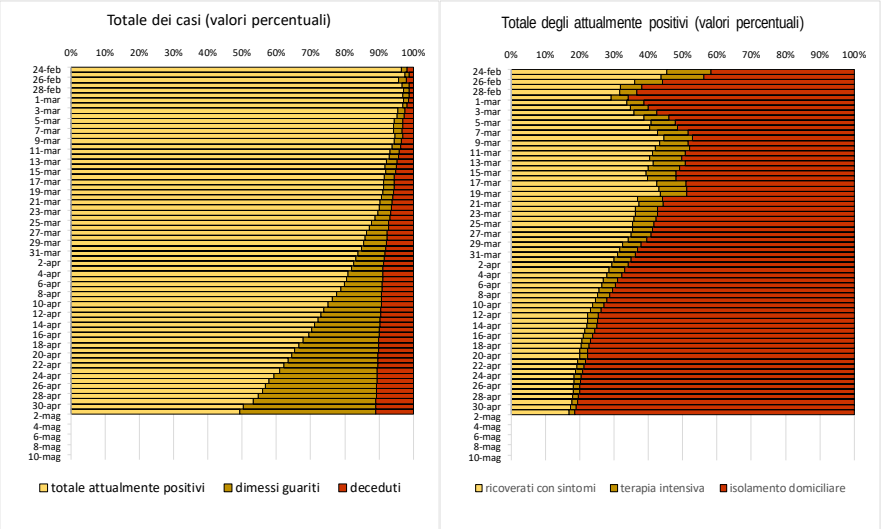
Le differenze regionali sono molto chiare e non necessitano di illustrazione. Il virus ha finora avuto effetti minori nelle regioni del centro sud. Ma come abbiamo detto prima ci possono esse state, nelle varie aree territoriali, valutazioni diverse della insorgenza (presenza) di casi positivi da parte dei medici, od anche strategie terapeutiche diverse da parte dei servizi sanitari locali.

Fig.1. Distribuzioni percentuali del totale dei Casi Positivi (a sinistra) e del totale degli Attualmente positivi (a destra)

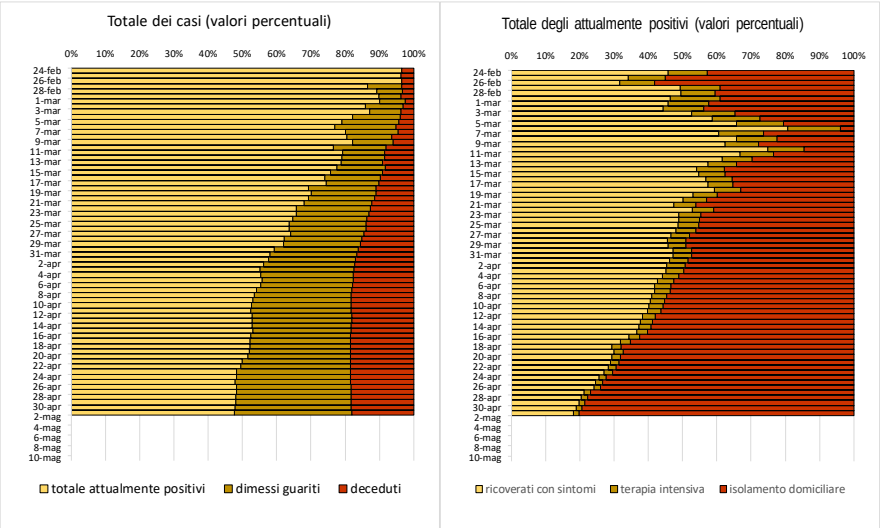
Italia



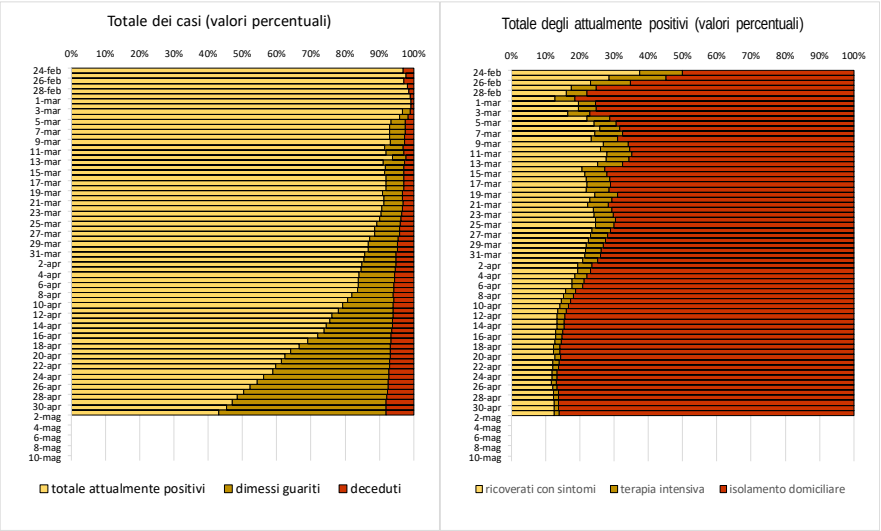
Italia-Lombardia



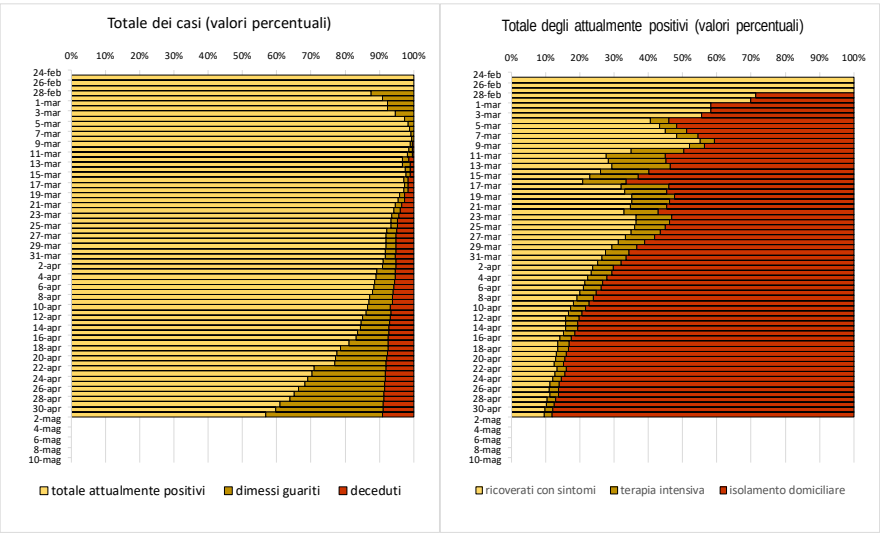
Lombardia



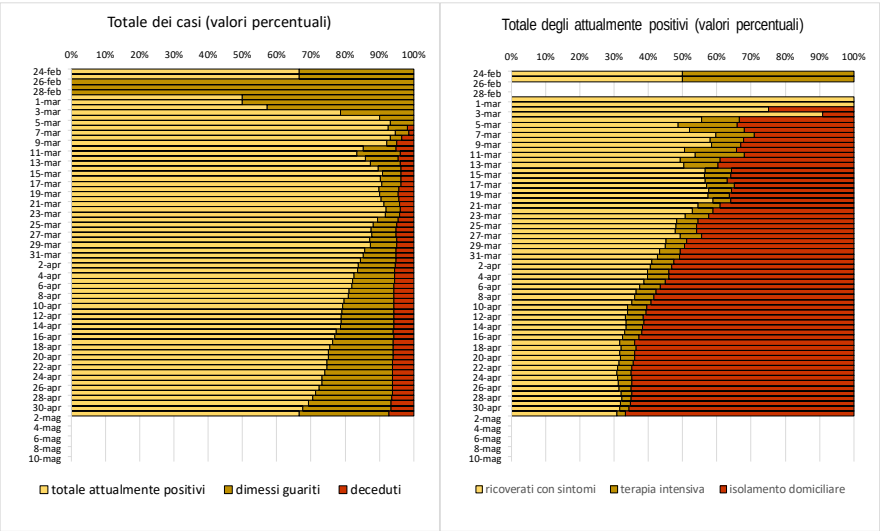
Veneto



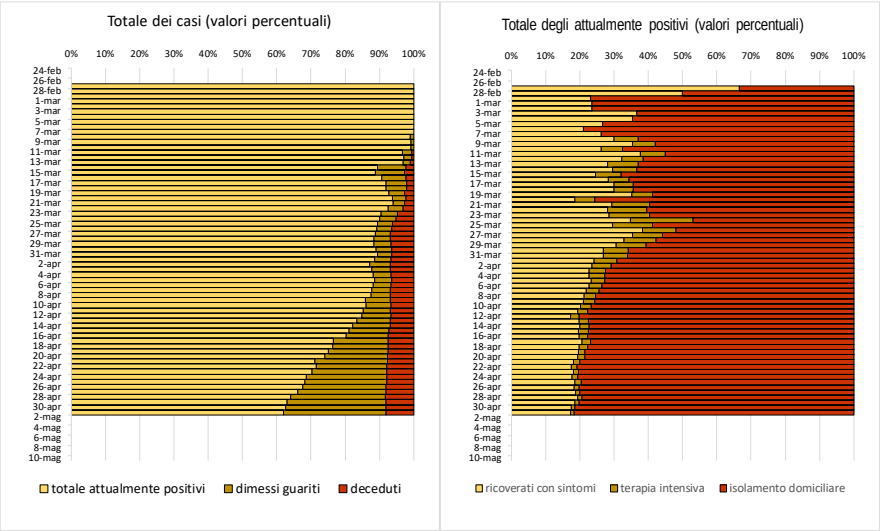
Toscana



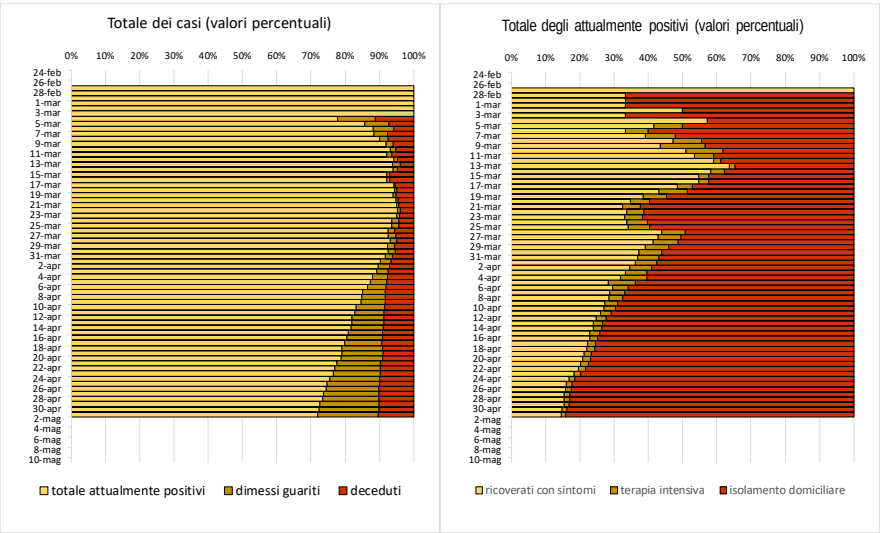
Lazio



Campania



Puglia



3.2 Grafici a barre

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

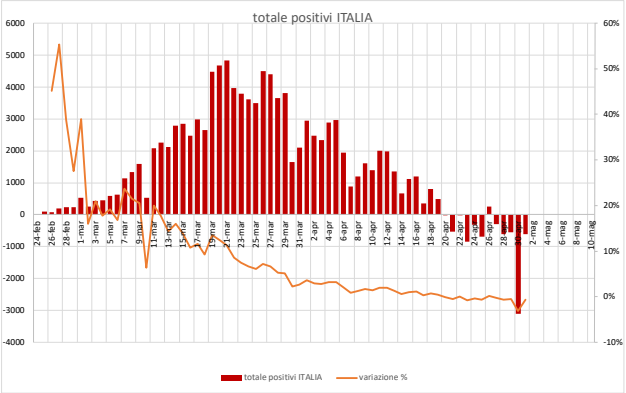
Qui presentiamo soltanto alcuni commenti di carattere generale, anche perché i grafici sono di immediata lettura e di facile approfondimento per analisi approfondite. Tuttavia, per una adeguata interpretazione degli andamenti occorre tenere presente del lag temporale tra i vari fenomeni (le cui medie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

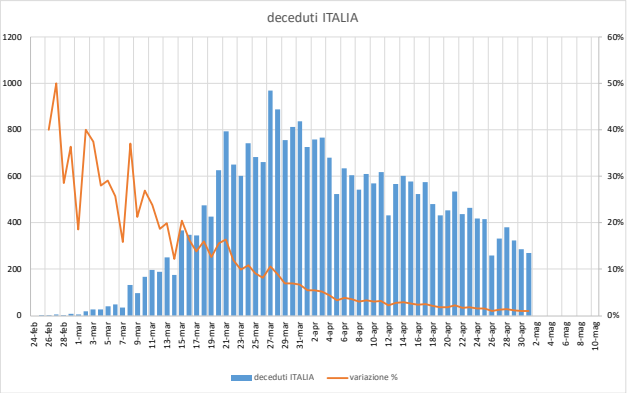
Dall'esame dei singoli grafici, è chiaro che le variazioni assolute sono inizialmente aumentate e molto, e ora si stanno stabilizzando e in alcuni casi riducendo (anche abbastanza velocemente). Le

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.

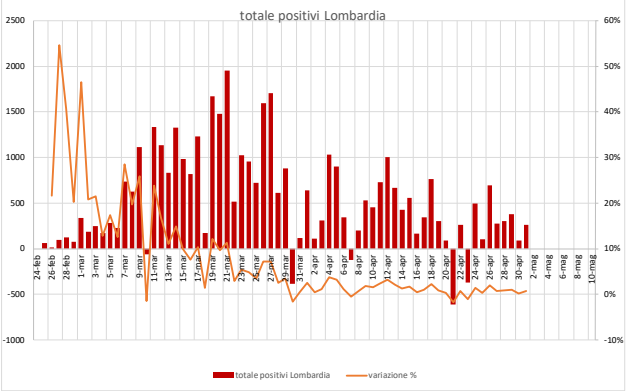
Tav.2. Differenze assolute e percentuali per ITALIA e regioni- totale Casi Positivi (1° grafico); Deceduti (2° grafico). 1° Maggio 2020

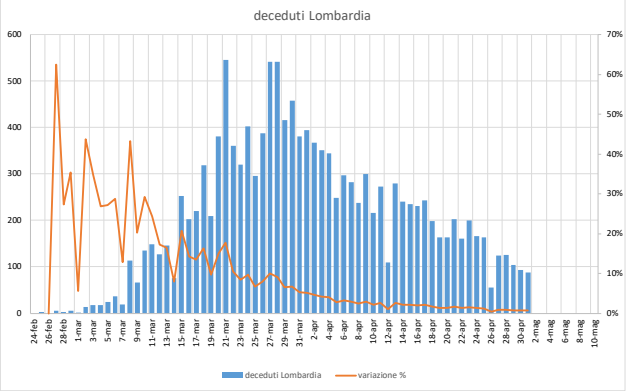
Italia



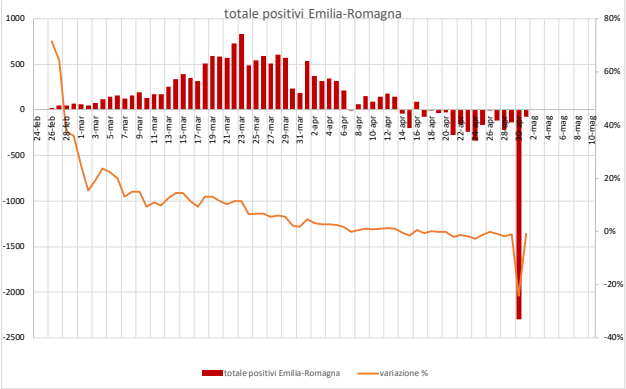


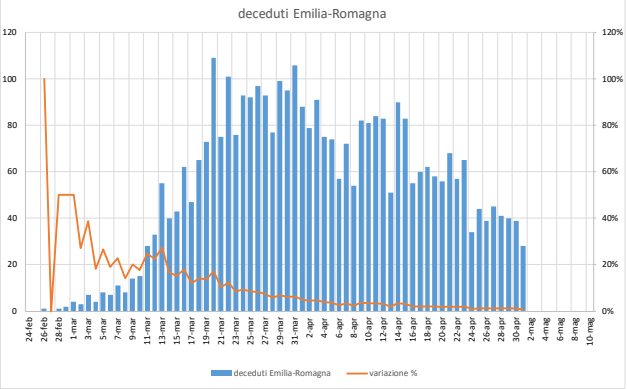
Lombardia



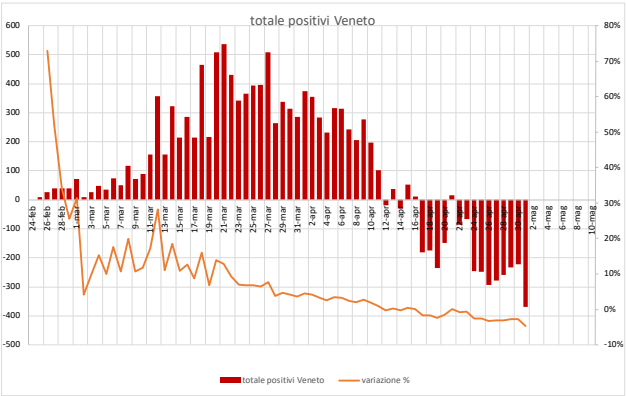


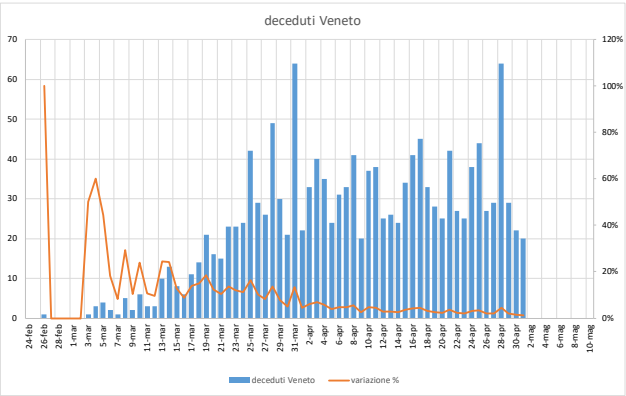
Emilia-Romagna



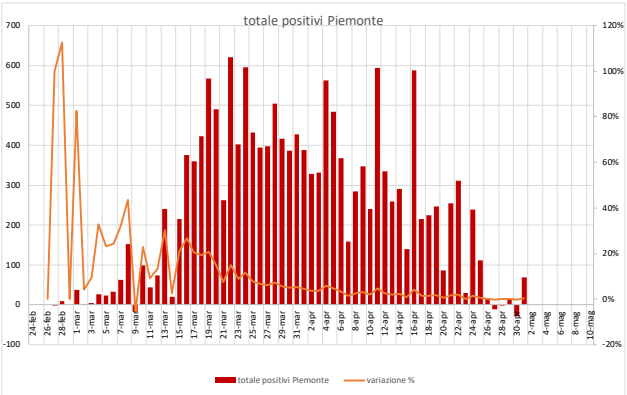


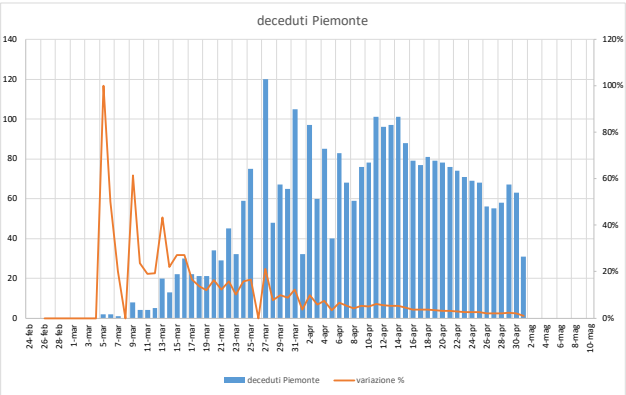
Veneto



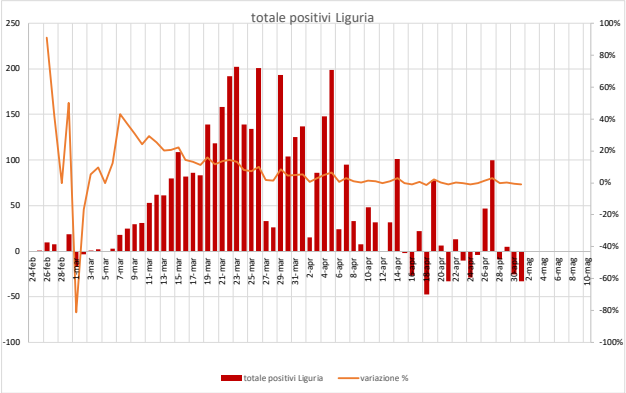


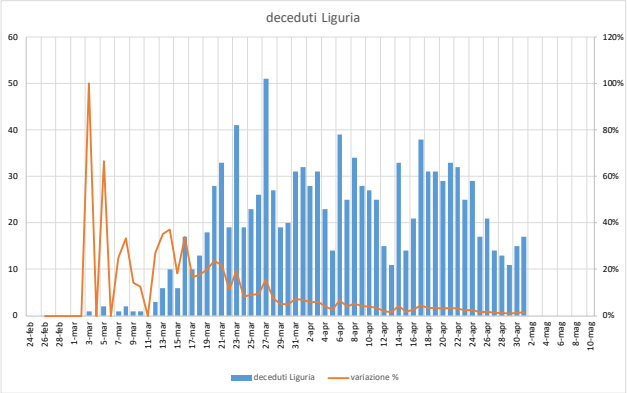
Piemonte



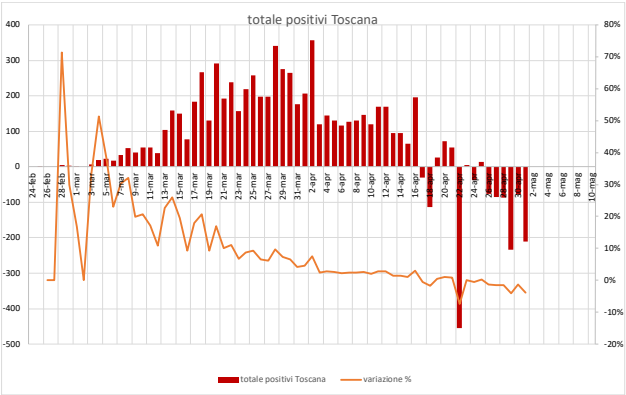


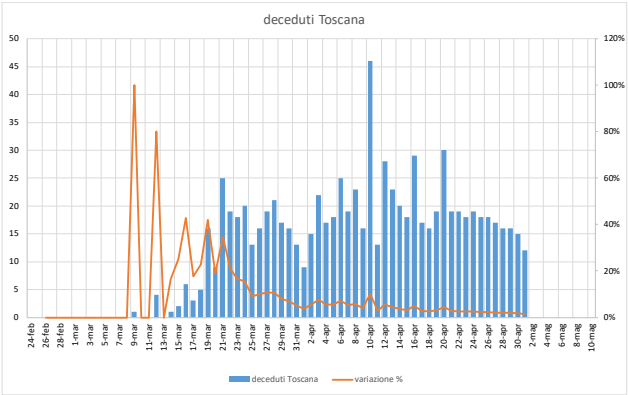
Liguria



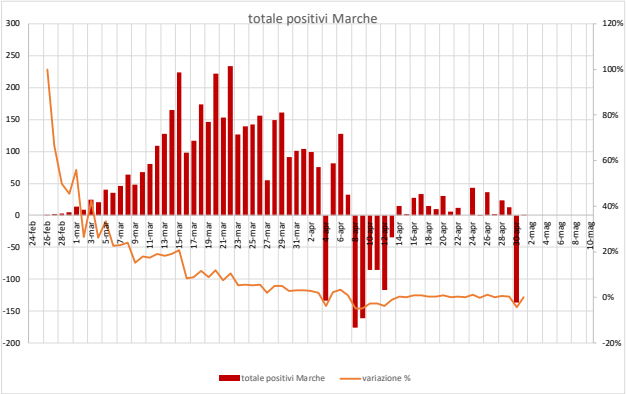


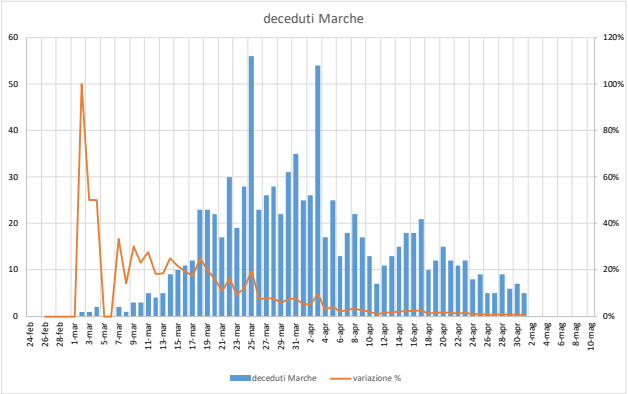
Toscana



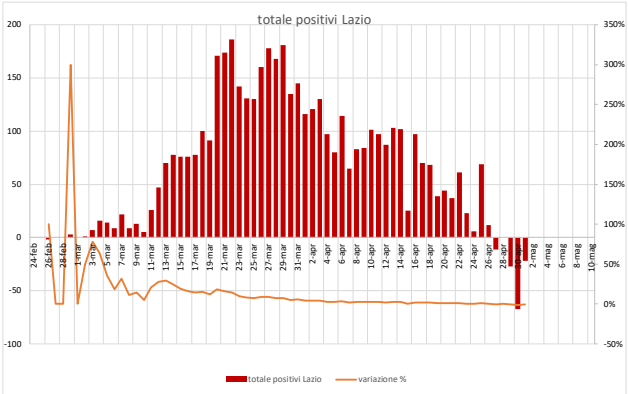


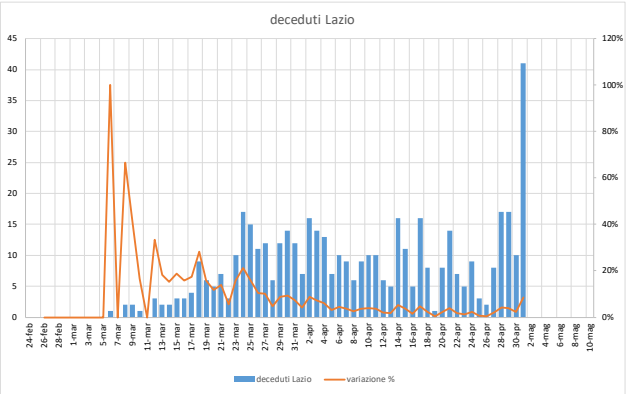
Marche



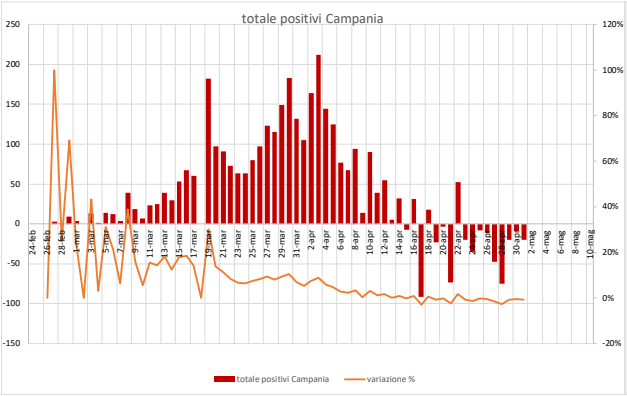


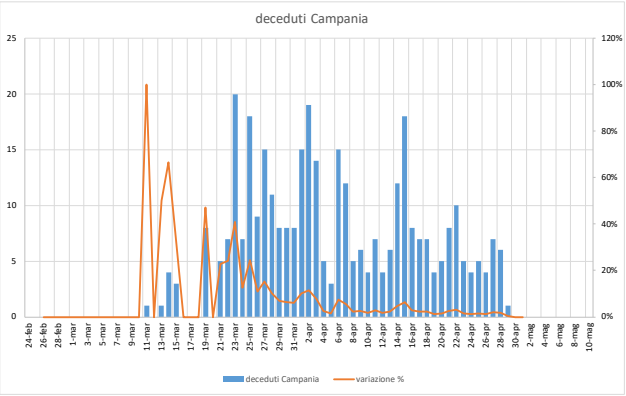
Lazio



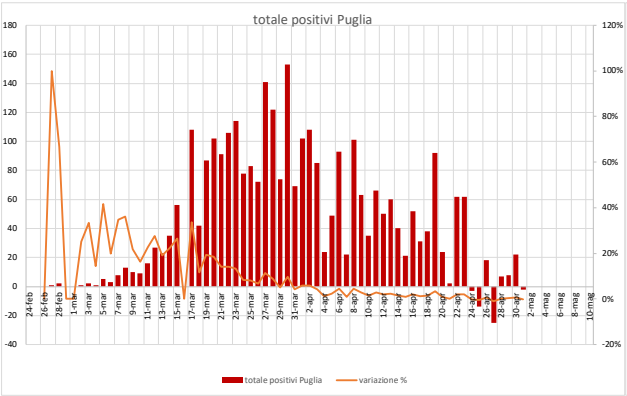


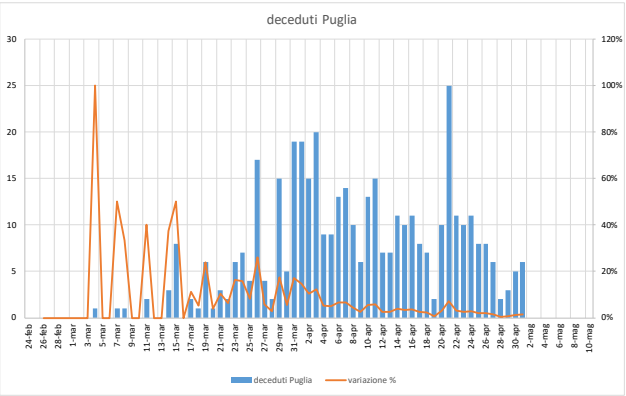
Campania



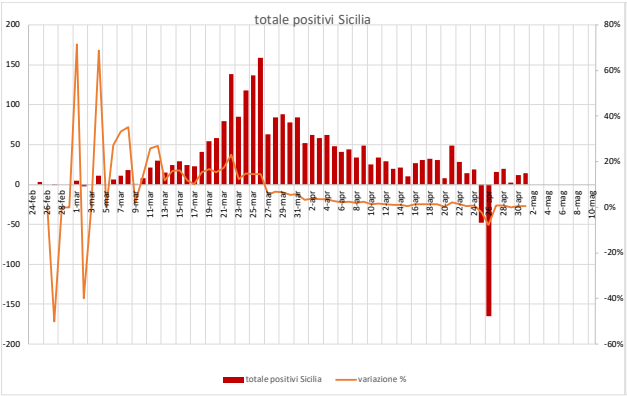


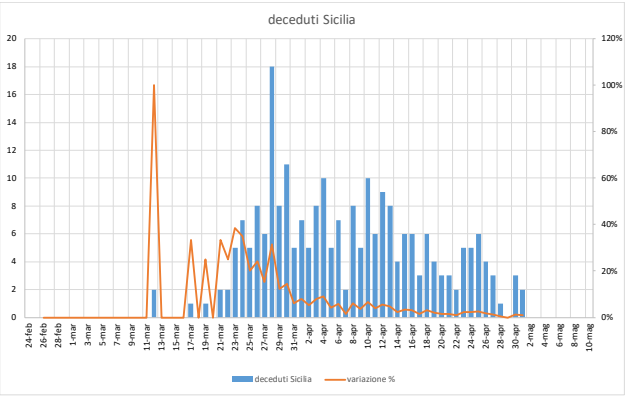
Puglia





Sicilia





3.3 Time plots in scala naturale e in scala logaritmica

Come è noto, nella analisi descrittiva preliminare delle serie storiche, i grafici che mettono meglio in evidenza l'evoluzione di fenomeni sono quelli cosiddetti time plots in scala naturale e in scala logaritmica. Questi ultimi, che riportano in ascissa il tempo e in ordinata il logaritmo, sono i più adeguati per fenomeni che crescono secondo una legge esponenziale modificata o logistica (cioè con un asintoto).

Inseriamo qui, nella Fig. 3-, una esemplificazione riguardante il Totale dei Casi Positivi (serie 1 nei grafici sottostanti), degli Attualmente Positivi (Serie 2) e dei Deceduti (serie 3) con riguardo all'Italia e ad alcune regioni diciamo più significative per l'estensione e gli effetti del virus.

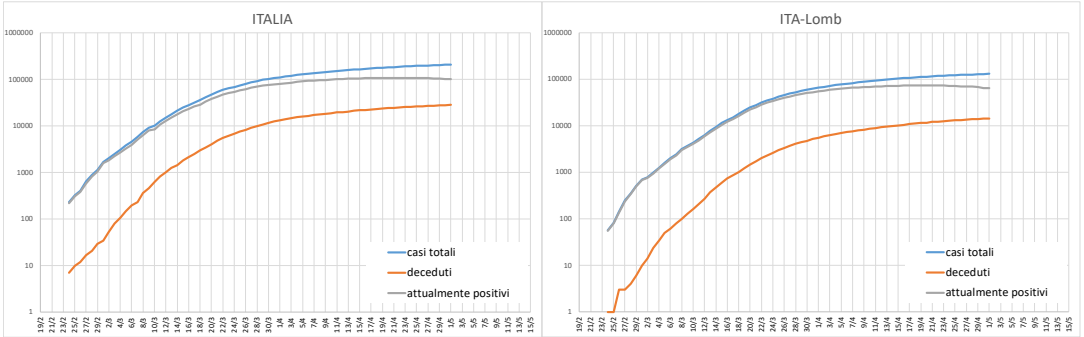
Come si può rilevare i grafici sono molto esplicativi mettendo in evidenza il diverso inizio della diffusione del virus e la sua differente esplosione nei tassi di sviluppo e la sua direzione verso l'asintoto, quando il fenomeno non dovrebbe fortunatamente aumentare più o presentare aumenti insignificanti. Sono evidenti i differenti comportamenti sia delle serie del numero di soggetti attualmente positivi che di quelle dei deceduti nelle varie regioni italiane.

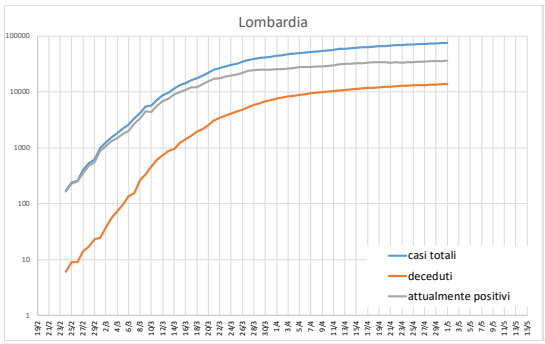
Come si vede, in alcune regioni il numero degli attualmente positivi tende all'asintoto, mentre il numero deceduti è un po' lontano dall'asintoto. In altre regioni accade il viceversa.

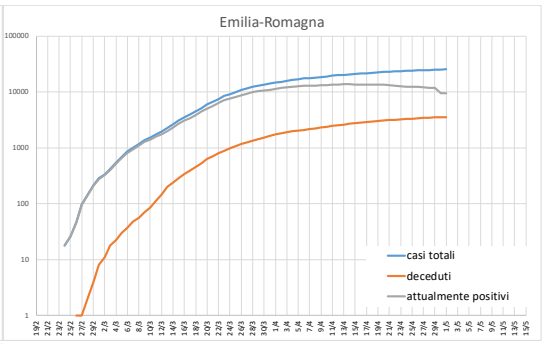
Naturalmente la descrizione dell'evoluzione della serie di ciascuna area territoriale deve essere fatta esaminandola molto attentamente per individuare i cambi di pendenza delle curve e gli eventuali punti di svolta, in particolare negli ultimi giorni. Necessariamente, in questa sede, i grafici sono di piccola dimensione, e questo fa sì che non mettono in evidenza tutte le oscillazioni e possono dare l'impressione che un unico modello di sviluppo si adatti a tutte le serie e per tutte le regioni (magari non considerando la diversa data di inizio del virus nelle diverse regioni). Ma così non è, e non essendo nostra intenzione descrivere l'evoluzione delle singole serie lasciamo al lettore tale analisi. Tuttavia, in alcuni casi, anche allargando o restringendo i grafici non è facile individuare bene gli eventuali punti di svolta che sono avvenuti e sono molto importanti sia ai fini interpretativi che previsivi.

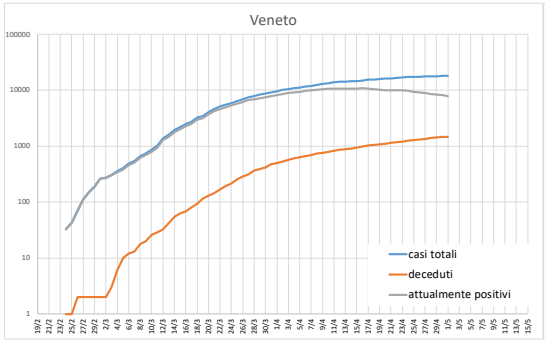
Si noti che in alcuni grafici regionali vi sono oscillazioni non giustificate se non per il fatto (come scrive l'ISS nei suoi reports) che gli stessi vengono aggiornati giornalmente da ciascuna regione anche se alcune informazioni possono richiedere qualche giorno per il loro inserimento. A volte la 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 dove gli attualmente positivi sono addirittura inferiori a quelli rilevati il 2 aprile 2020; ma come il lettore può rilevare non è il solo caso!

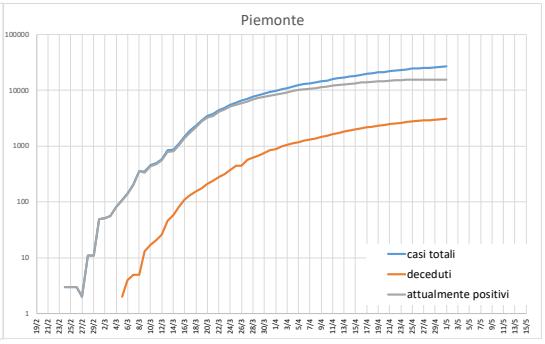
FIG. 3. LOG PLOTS SU COVID-19 PER ITALIA E 11 REGIONI- 1 Maggio-2020 -Casi Totali (serie 1) ; Attualmente positivi (serie 2) e Deceduti (serie 3). In ordinata scala logaritmica, in base 10

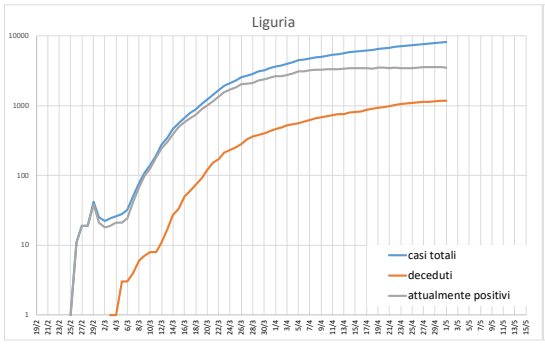


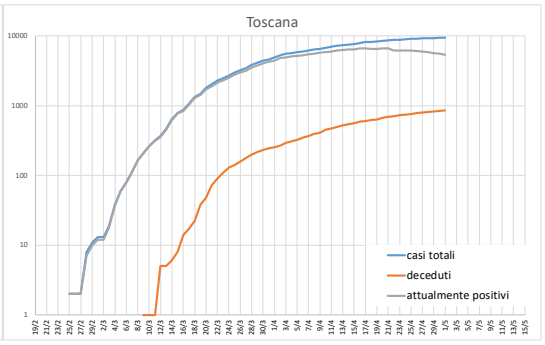

Lombardia
casi totali
deceduti
attualmente positivi

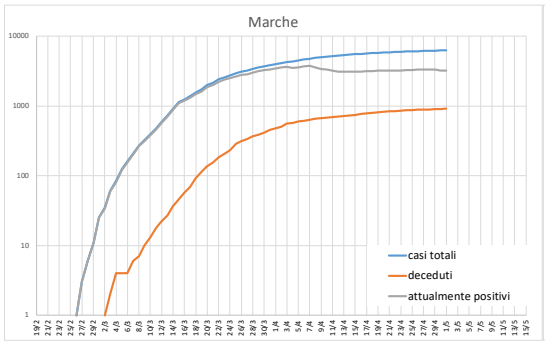

Emilia-Romagna
casi totali
deceduti
attualmente positivi

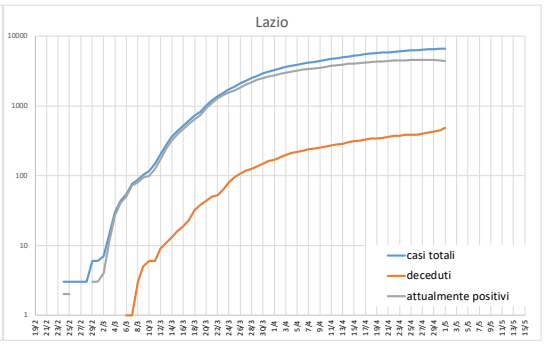

Veneto
casi totali
deceduti
attualmente positivi

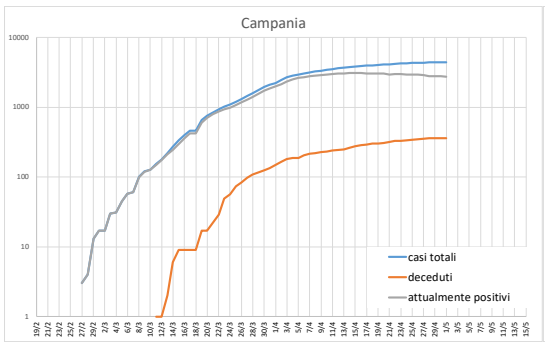

Piemonte
casi totali
deceduti
attualmente positivi

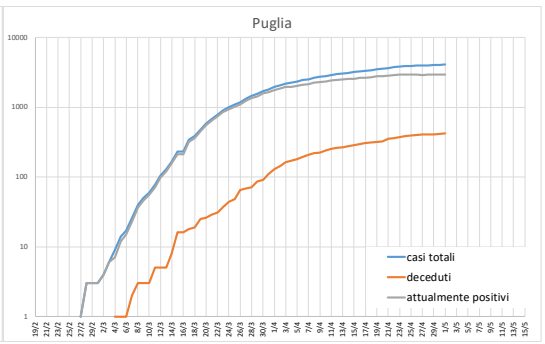

Liguria
casi totali
deceduti
attualmente positivi

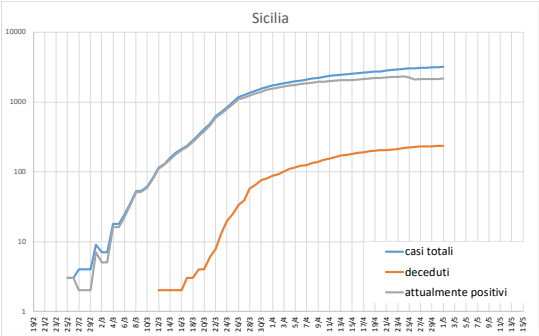

Toscana
casi totali
deceduti
attualmente positivi


Marche
casi totali
deceduti
attualmente positivi


Lazio
casi totali
deceduti
attualmente positivi


Campania
casi totali
deceduti
attualmente positivi


Puglia
casi totali
deceduti
attualmente positivi



3.4 I confronti tra le regioni dell'intensità del covid-19

I grafici finora illustrati consentono di confrontare le diverse evoluzioni temporali dei vari aspetti del coronavirus nelle regioni, ma non consentono di fare i confronti tra regioni per valutare comparativamente la diffusione del virus, i tassi di mortalità e di letalità.

A questo fine i dati e i tassi devono essere standardizzati, in primo luogo rispetto all'ammontare della popolazione di ciascuna regione, ma possibilmente anche per tener conto della struttura per età della popolazione, della struttura del numero di morti per causa e 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. Per fare 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.

Al momento, abbiamo preso in considerazione soltanto il diverso ammontare della popolazione e come si può vedere dal prospetto qui riportato

Tabella con valori per milione di abitanti aggiornati al 1° maggio

	ricoverati con sintomi	terapia intensiva	totale ospedalizzati	isolamento domiciliare	totale positivi	dimessi guariti
Abruzzo	226,4	12,2	238,6	1218,4	1457,0	543,6
Basilicata	87,1	7,1	94,2	248,7	342,9	284,3
P.A. Bolzano	215,0	21,1	236,1	1217,1	1453,3	2866,2
Calabria	53,9	1,5	55,5	317,9	373,4	153,6
Campania	82,4	5,0	87,4	387,1	474,5	229,6
Emilia-Romagna	473,6	44,2	517,8	1608,9	2126,7	2821,2
Friuli Venezia Giulia	107,8	4,9	112,7	804,8	917,5	1343,0
Lazio	233,4	17,9	251,2	505,0	756,2	296,6
Liguria	429,5	43,9	473,4	1795,4	2268,7	2208,1
Lombardia	658,8	56,0	714,8	2910,6	3625,3	2597,9
Marche	270,8	28,8	299,6	1805,6	2105,2	1411,6
Molise	55,6	3,3	58,9	562,8	621,7	291,2
Piemonte	574,6	41,5	616,1	2956,1	3572,2	1842,1
Puglia	106,2	9,7	115,9	615,5	731,4	181,4

Sardegna	51,2	7,3	58,6	395,2	453,8	275,7
Sicilia	79,8	6,0	85,8	348,4	434,2	157,2
Toscana	139,4	33,2	172,7	1268,0	1440,6	862,8
P.A. Trento	271,3	35,3	306,6	2095,8	2402,3	4488,8
Umbria	68,0	14,7	82,8	148,5	231,3	1271,0
Valle d'Aosta	572,9	23,9	596,8	135,3	732,1	7193,7
Veneto	199,1	22,4	221,6	1364,1	1585,7	1801,9
ITALIA	291,1	26,1	317,3	1355,4	1672,7	1296,7
Italia, esclusa la Lombardia	217,6	20,2	237,8	1044,3	1282,1	1036,3

	deceduti in percentuale sulla popolazione (mortalità)	percentuale di deceduti su totale casi (letalità)
Abruzzo	0,025	11,0
Basilicata	0,004	6,6
P.A. Bolzano	0,053	11,0
Calabria	0,004	7,7
Campania	0,006	8,1
Emilia-Romagna	0,080	14,0
Friuli Venezia Giulia	0,024	9,7
Lazio	0,008	7,2
Liguria	0,076	14,6
Lombardia	0,138	18,1
Marche	0,060	14,5
Molise	0,007	7,0
Piemonte	0,071	11,6
Puglia	0,010	10,3
Sardegna	0,007	8,9
Sicilia	0,005	7,4
Toscana	0,023	9,0
P.A. Trento	0,079	10,2
Umbria	0,008	4,9
Valle d'Aosta	0,109	12,1
Veneto	0,030	8,2
ITALIA	0,047	13,6
Italia, esclusa la Lombardia	0,029	11,0

Prossimamente effettueremo anche una standardizzazione rispetto alla diversa struttura per età della popolazione.

4 **Una nuova rappresentazione grafica: log plots con adattamento di rette mobili**

Già nell'articoletto apparso il 7 Aprile in questo sito e in quello della SIS, abbiamo ritenuto opportuno cercare evidenziare meglio, di quanto non si possa fare con i Time Plots e Log Plots, i cambiamenti di pendenza e gli eventuali punti di svolta in particolare nella parte finale delle serie. Abbiamo pensato che forse poteva essere utile costruire dei grafici basati sui "Moving Least Squares Methods for scattered data" (che abbiamo visto applicato nel paper "PPP-based Stratification of CIS-EU/OECD Economies" presentato da A. Kosarev alla IARIW–HSE Conference, Moscow, September 17-18, 2019).

Sostanzialmente, nel nostro caso, si tratta di stimare, sulle serie dei logaritmi (ma si può fare per qualunque serie storica!), 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 $\underline{n}$ 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 $\underline{n}$ 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 $\underline{k}$ di termini, con $\underline{k}$=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 $\underline{k}$ tempi (giorni) successivi se la pendenza della curva rimanesse la stessa di quella stimata sugli $\underline{n}$ tempi precedenti.

I log plots con rette (segmenti) mobili saranno diversi a seconda della combinazione dei valori scelti per $\underline{n}$ e $\underline{k}$, nonché delle caratteristiche di evoluzione di ciascun fenomeno. Con $\underline{n}$ 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; $\underline{k}$ deve essere necessariamente non grande e comunque inferiore ad $\underline{n}$.

Come si vedrà nei grafici che presentiamo a titolo esemplificativo, quando la pendenza della curva rimane stabile per un certo periodo, numero di giorni, le spezzate sono molto più ravvicinate mettendo in evidenza sia i cambi di pendenza che i punti di svolta; quando invece la pendenza delle rette cambia spesso (anche perché la serie presenta frequenti oscillazioni) le spezzate rappresentate nel grafico sono distanti l'una dall'altra e disposte quasi a ventaglio; Quando la pendenza cambia poi cambia in modo più stabile, 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 di $\underline{n}$ e $\underline{k}$. Non abbiamo trovato nessun programma che ci consentisse di fare le stime dei

log plots (o in generale dei time plots) con rette mobili. Il programma è stato predisposto da Mauro con Excel.

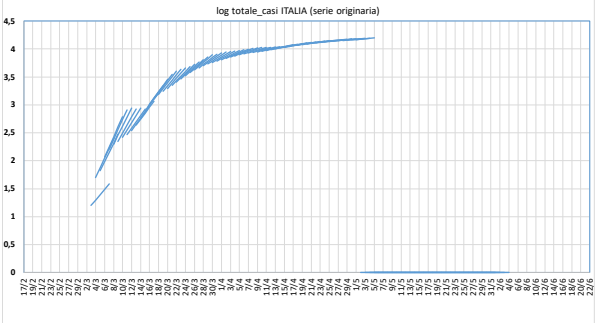
Gli esempi che riportiamo di seguito servono per far vedere l'utilità di questi grafici anche per verificare gli effetti dei provvedimenti di contenimento del virus, e come base per ipotizzare l'evoluzione futura, a breve termine, dei fenomeni oggetto di studio. Non diciamo che "parlano da soli", anzi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

A titolo esemplificativo, riportiamo nella Fig. 4, i **Log Plots con rette mobili**, per *n* = da 7 a 10 a e *k* = 3 o 4. Abbiamo scelto *n* = uguale come minimo a 7 considerando che per verificare se i fenomeni stanno mantenendo o modificando la loro evoluzione occorre almeno una settimana; mentre *k* deve essere necessariamente piccolo trattandosi di una proiezione e non di una previsione.

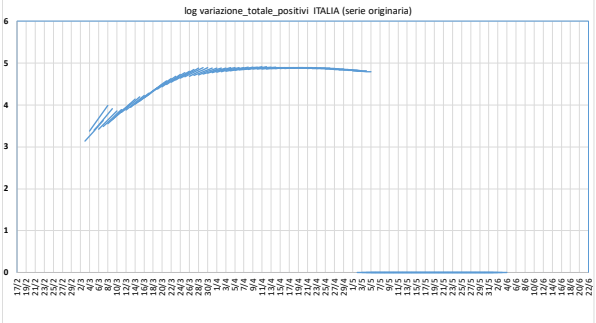
FIG. 4

ITALIA

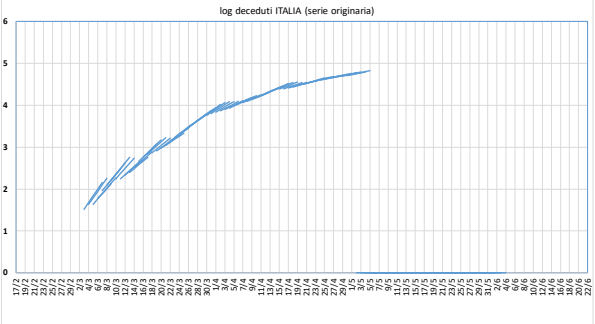
Totale Casi Positivi, scala log: n=7 , *k* =4



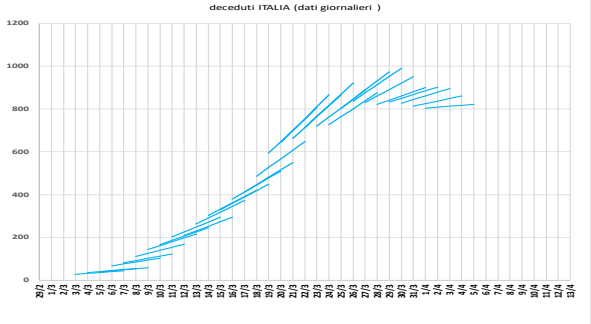
Totale Attualmente Positivi, scala log: n=7 , *k* =4



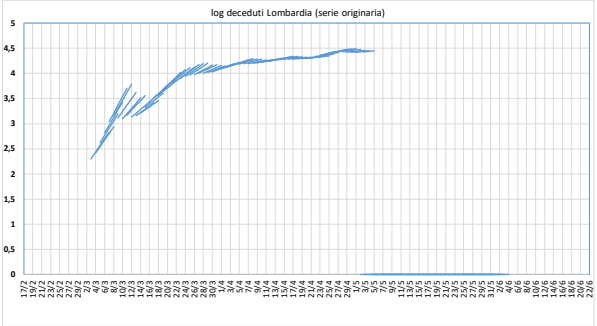
Deceduti, scala log: n=7 , *k* =4



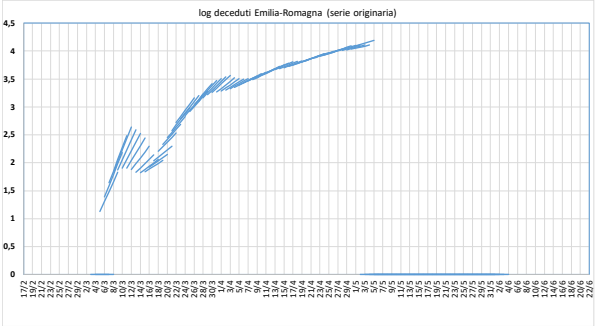
Deceduti Giornalieri, scala naturale: n=10 , *k* =5



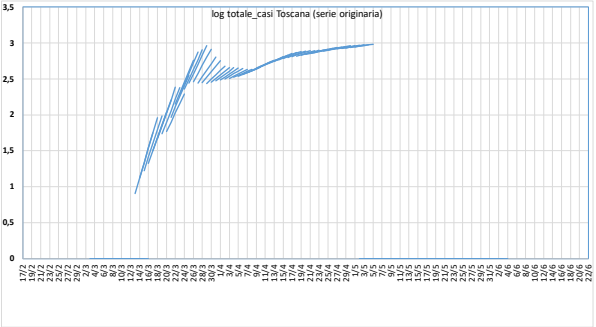
LOMBARDIA Deceduti, scala log: n=7 , *k* =4



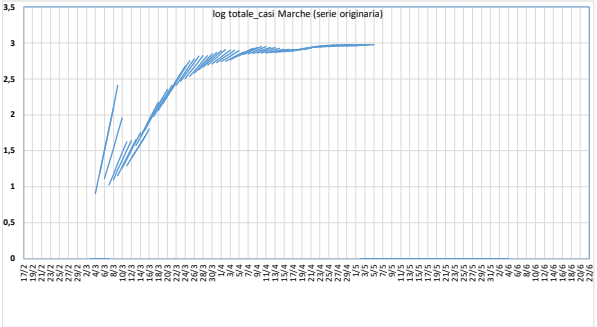
EMILIA-ROMAGNA Deceduti, scala log: n=7 , *k* =4



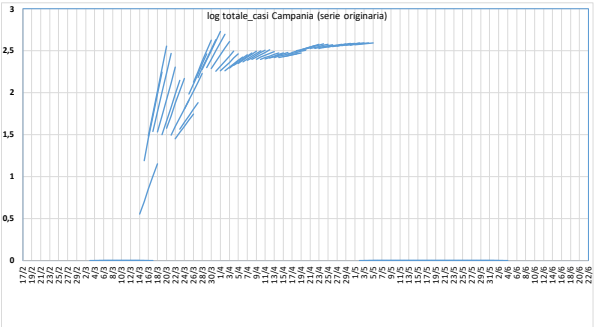
TOSCANA Totale casi, scala log: n=7 , *k* =4



MARCHE Totale casi, scala log: n=7 , *k* =4



CAMPANIA Totale casi, scala log: n=7 , *k* =4



I grafici all'inizio e in occasioni di oscillazioni della serie (come avviene per le serie della Toscana e altre regioni ed in particolare per la serie dei 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 l'esame dei grafici va sempre fatto tenendo presente la retta più in basso di ogni fascio e 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).

5 Il modello Holt-Winters con trend, per descrivere l'andamento delle serie e fare proiezioni a brevissimo termine

Nell'ottica di descrivere meglio gli eventuali punti di svolta delle serie e ottenere utili informazioni sugli andamenti dell'ultimo periodo abbiamo provato ad applicare i modelli di livellamento Esponenziale (Exponential Smoothing) di tipo Holt-Winters.

Come è noto, il livellamento esponenziale è stato inizialmente proposto soprattutto come metodo di previsione a breve e brevissimo termine nell'ambito delle analisi delle serie storiche tramite le medie mobili. Il lisciamento della serie storica con medie mobili evidenzia l'andamento della stessa in modo più o meno preciso in relazione al numero dei termini della media mobile utilizzata e l'ultima stima non è riferita al tempo corrente t_i, ma è centrata su un tempo precedente, in relazione al numero dei termini usati per il calcolo della media mobile. Quindi anche nella loro rappresentazione grafica si perdono gli ultimi termini della serie che sono quelli che più interessano. Inoltre, le medie mobili danno la stessa importanza a tutti i termini usati, mentre appare logico attribuire più importanza ai dati più recenti e questo è quanto viene fatto con il livellamento esponenziale, che attribuisce maggiore peso ai dati più recenti e minore peso, decrescente in modo esponenziale, ai termini precedenti della serie.

Nell'ipotesi che la serie storica presenti un livello (M) approssimativamente costante nel breve periodo, in genere si usa il livellamento esponenziale costante che possiamo scrivere in questo modo:

$$\widehat{M}_t = \hat{y}_{t+1} = \delta_1 y_t + (1 - \delta_1)\hat{y}_t = \hat{y}_t + \delta_1(y_t - \hat{y}_t)$$

dove il parametro di smussamento δ_1 è il peso attribuito all'ultimo termine della serie. Come risulta dalla formula finale sostanzialmente si modifica la previsione (i valori stimati) in modo iterativo e nell'effettuare la previsione al tempo t+1 si modifica la previsione precedente del tempo t tenendo conto dell'errore di previsione $(y_t - \hat{y}_t)$ commesso nel prevedere y_t ponderato secondo il valore del parametro di smussamento δ_1.

Poiché le serie storiche delle diverse casistiche rilevate sul Covid-19 presentano trend, in questo caso è necessario impiegare il Modello di Holt-Winters (H-W) per serie storica con trend (ovviamente se ci fossero anche le oscillazioni stagionali dovremmo prendere in considerazione anche la stima della stagionalità).

Questo modello è composto di due parti: (i) il livello medio della serie stimato per il tempo t ($\widehat{M}_t$), più (ii) la stima del trend al tempo t ($\hat{T}_t$), cioè tra il tempo t-1 e il tempo t. Questa seconda componente è interpretata e stimata come media ponderata tra la variazione di livello tra i due tempi e il trend al tempo t sempre utilizzando il livellamento esponenziale.

La formula che esprime il modello H-W con trend può così essere scritta:

$$\widehat{M}_t = \delta_1 y_t + (1 - \delta_1)\left(\widehat{M}_{t-1} + \hat{T}_{t-1}\right)$$

dove $\hat{T}_{t-1} = \delta_2\left(\widehat{M}_t - \widehat{M}_{t-1}\right) + (1 - \delta_2)\hat{T}_{t-1}$

I due parametri di smussamento δ_1 e δ_2 più adeguati in relazione alla serie storica impiegata, sono individuati minimizzando la somma dei quadrati degli scarti degli errori di previsione, cioè la SSQ tra valori osservati e valori *fitted.*

Per fare la previsione a un tempo futuro inizialmente si prevede y al tempo t+1 e poi passo dopo passo si prevede y ai tempi successivi.

Poiché si tratta di una previsione, che sarebbe opportuno chiamare proiezione, condizionata molto dagli ultimi termini della serie è necessario verificarne il comportamento effettuando le cosiddette previsioni ex-post (che comunque seguono l'usuale logica del test di *goodness of forecast* condotto a posteriori). Questo calcolo fatto sui periodi precedenti a t consentirà di valutare la sua utilizzabilità in relazione alla serie storica presa in considerazione ed anche il più opportuno intervallo temporale di previsione.

Di seguito riportiamo alcuni esempi di applicazione del modello per le serie già esaminate in precedenza con il Log Plots del totale di casi positivi per l'Italia e la Toscana. I grafici sono talmente esplicativi che non ci sembra necessitino di commenti.

Modello H-W per Totale Casi Positivi (scala log) - ITALIA

Previsione a 5 giorni

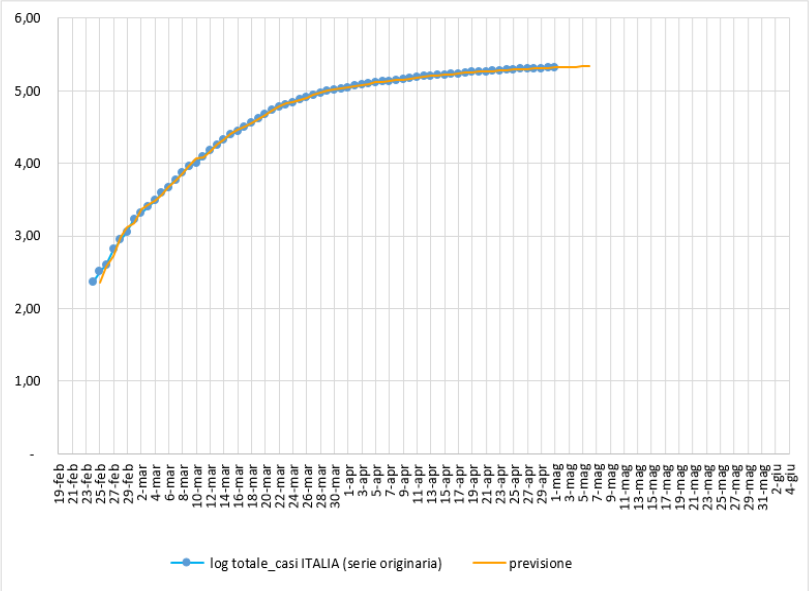
gg. Previsione

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

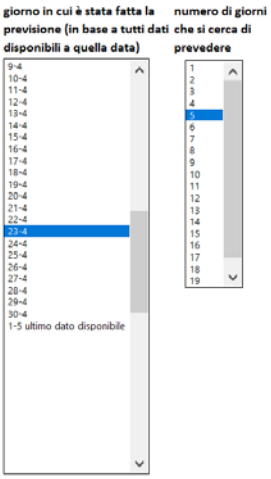
δ1	0,7579079
δ2	1

SSQ 0,0486228

data	giorno	previsione	osservato
1/5	0	5,3169829	5,3168674
2/5	1	5,3209683	
3/5	2	5,3248507	
4/5	3	5,328836	
5/5	4	5,3327185	
6/5	5	5,3367038	

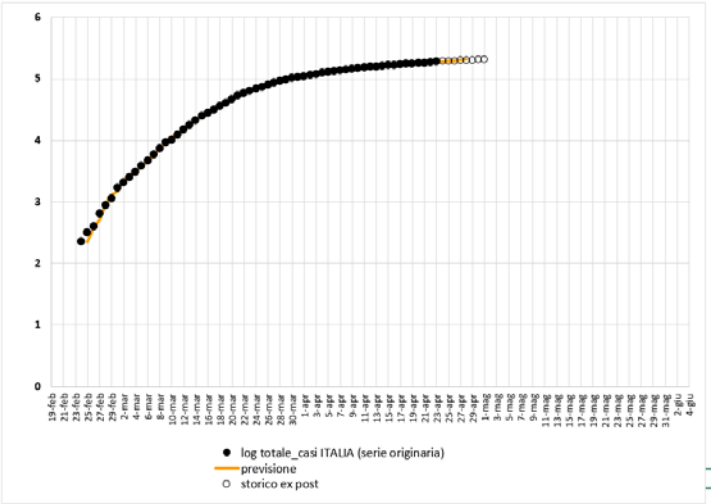


Verifica ex-post del modello per 5 giorni dal 23 al 28 aprile



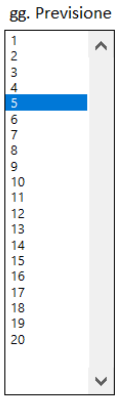
δ1	0,75789
δ2	1
SSQ fitting	0,04862

data	giorno	previsione	storico	forecast MAPE ex post 0,1% forecast-APE (errore assoluto percentuale)
23/4	0	5,2794953	5,27869	
24/4	1	5,2856852	5,28554	0,0%
25/4	2	5,294698	5,29082	0,1%
26/4	3	5,3008878	5,29595	0,1%
27/4	4	5,3099006	5,29976	0,2%
28/4	5	5,3160905	5,30429	0,2%



Modello H-W per Totale Casi Positivi (scala log) – TOSCANA

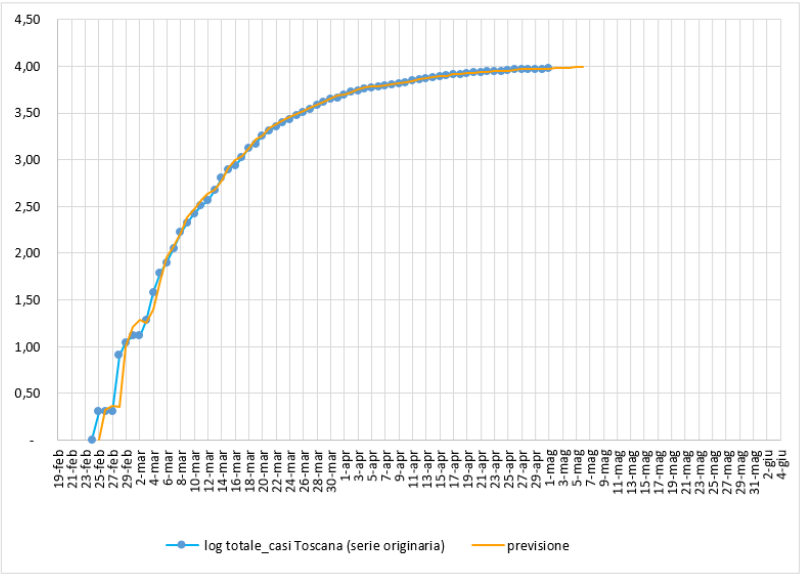
Previsione a 5 giorni



δ1	0,8601001
δ2	0,2611453

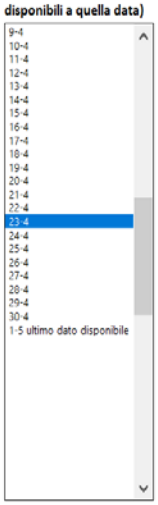
SSQ 0,5056192

data	giorno	previsione	osservato
1/5	0	3,9746727	3,975202
2/5	1	3,9788258	
3/5	2	3,9812415	
4/5	3	3,9853946	
5/5	4	3,9878102	
6/5	5	3,9919633	

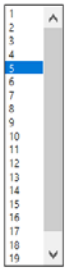


Verifica ex-post del modello per 5 giorni dal 23 al 28 aprile

giorno in cui è stata fatta la previsione (in base a tutti dati disponibili a quella data)



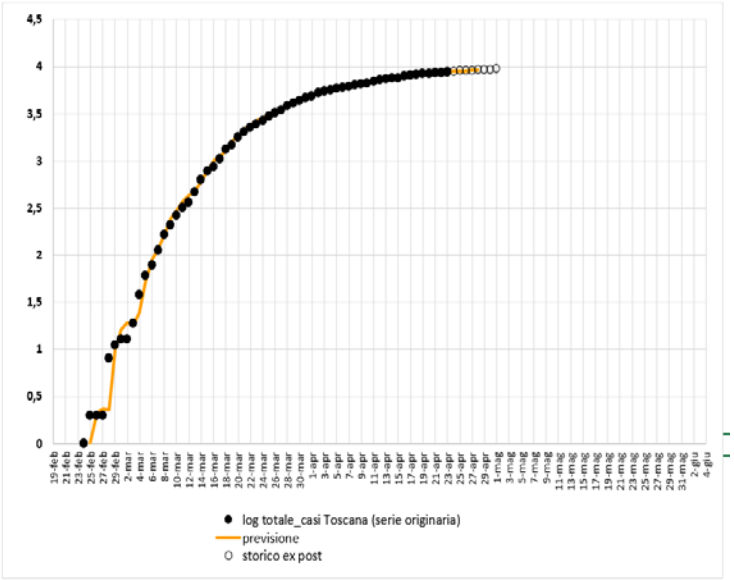
numero di giorni che si cerca di prevedere



δ1	0,86017
δ2	0,26105

SSQ fitting 0,50558

data	giorno	previsione	storico	forecast-APE (errore assoluto percentuale)
				forecast MAPE ex post 0,0%
23/4	0	3,9469703	3,94349	
24/4	1	3,9501662	3,94827	0,0%
25/4	2	3,9541497	3,95497	0,0%
26/4	3	3,9573456	3,96128	0,1%
27/4	4	3,9613291	3,9628	0,0%
28/4	5	3,964525	3,96525	0,0%



Appendice

Preparazione e funzionamento del Data Base e di programmi di elaborazione flessibili

(di Mauro Maltagliati)

I lettori sono autorizzati ad usare il data base per la elaborazione dei grafici per le serie ivi riportate. Comunque, chiediamo a chi lo utilizza per le serie che per le elaborazioni per produrre lavori (articoli, ecc.) di citare sia il data base che il Manuale.

Manuale per l'uso del file autoDB.xlsm

Il file "autoDB.xlsm" è un file di Excel che contiene delle macro (da cui l'ultima lettera dell'estensione del file) scritte in linguaggio Visual Basic, Applications Edition (VBA). Le macro in Excel sono disabilitate per impostazione predefinita, dato che sono suscettibili di essere utilizzate per la consegna di codice dannoso. Per poter usare tutte le macro del file si deve cliccare su "Abilita contenuto" alla sua apertura. In particolare, in questo file, le macro sono necessarie per avviare la macro di aggiornamento automatico del database, come descritto più avanti.

Per eventuali domande sull'utilizzo, ma anche per segnalare eventuali "bugs", contattare mauro.maltagliati@unifi.it.

Se il file viene utilizzato per elaborazioni , si prega di citare la

All'apertura del file, come al solito in Excel, in basso si trovano le "linguette-scheda" che consentono di navigare tra i fogli elettronici.

In questo file ce ne sono 13, raggruppate per colore a seconda della tipologia.

Aggiornamento database (linguette azzurre)

Le Linguette azzurre riguardano il database vero e proprio, che viene aggiornato cliccando sul pulsante che si trova nel foglio "AGGIORNA".

Cliccando sul pulsante si aziona una macro che carica il file dalla pagina web:

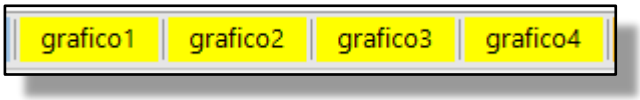
https://github.com/pcm-dpc/COVID-19/blob/master/dati-regioni/dpc-covid19-ita-regioni.csv.

Solitamente i nuovi dati sono presenti dalle ore 19 di ogni giorno. I dati vengono poi riordinati e immessi nel foglio "DBregioni". Come si può notare, la tabella non inizia dalla riga 1 come ci si aspetterebbe. Per trovarla si deve scorrere il foglio fino alla riga 50-70 circa (a seconda dei computer sui quali abbiamo effettuato la prova). Questa "stranezza" è dovuta al fatto che l'operazione di riordinamento della tabella provocava errori nell'esecuzione della macro su alcuni computer su cui il file è stato caricato. In futuro contiamo di risolvere questa piccola imperfezione. Sui computer Mac l'aggiornamento non ha funzionato.

Comunque, una volta che sul proprio computer si individua la riga in cui la tabella dei dati regionali inizia, ad ogni aggiornamento la troveremo sempre nella stessa posizione.

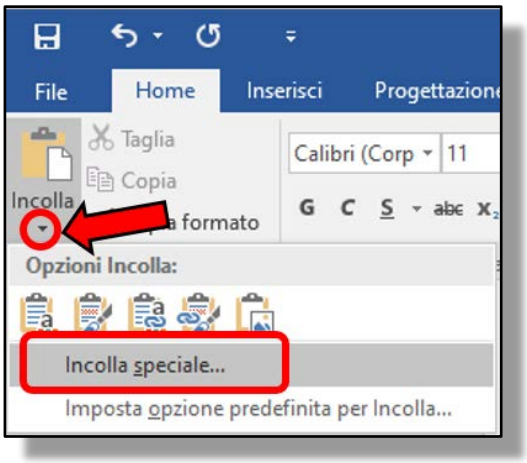
Le tabelle DBItalia e DB_No_Lomb sono ricavate dalla DBRegioni e riportano i dati aggregati per l'Italia intera e per l'Italia con esclusione della Lombardia. Queste tabelle consentiranno di produrre grafici ed elaborazioni per questi due raggruppamenti territoriali (in aggiunta alla possibilità di avere i grafici per le singole regioni).

Grafici semplici (linguette gialle)



Su questi fogli non c'è molto da dire, sono grafici il cui contenuto informativo è facilmente comprensibile. L'unica opzione per l'utente è quella di scegliere la circoscrizione geografica tra le 23 opzionabili: le 21 regioni (le province del Trentino sono separate nel database originale), l'Italia e infine l'Italia con esclusione della Lombardia. La selezione avviene cliccando sulla voce desiderata nel menu a tendina.

In questi fogli l'unica ulteriore operazione che si può compiere è copiare i grafici per incollarli su un'altra applicazione. Se si intende trasferirli in Word, si consiglia (all'interno del documento di Word) di scegliere l'opzione "Incolla speciale- Immagine (Enhanced Metafile)". È un consiglio (valido in generale) che consente di trattare il grafico importato come una semplice immagine, pur perdendo il "collegamento" con Excel.



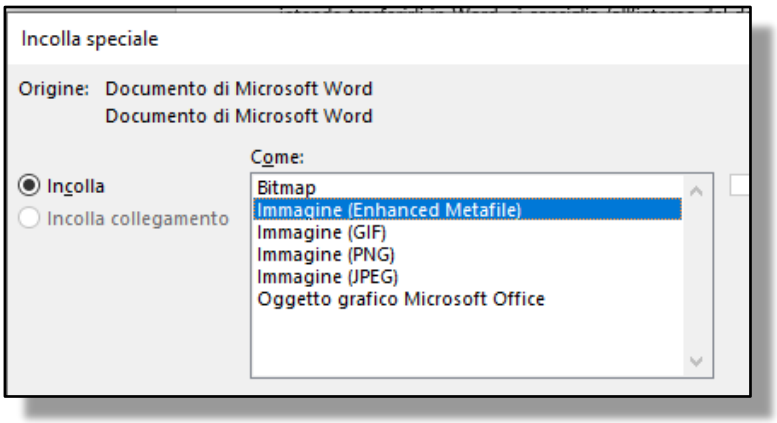
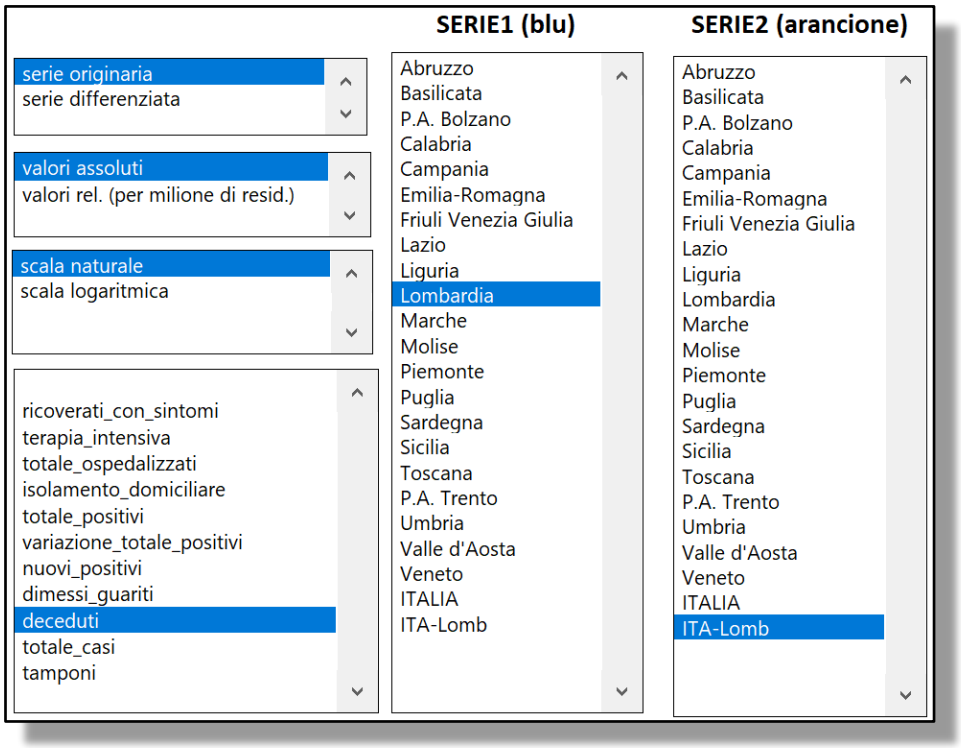


Grafico di confronto dei time plot per due diverse zone geografiche (linguetta arancione "DUE zone")

Su questo foglio è possibile mettere a confronto la stessa serie storica per due differenti aree geografiche. Le due aree sono selezionate attraverso i due menu gemelli ("SERIE1 (blu)" e "SERIE2 (arancione)") e il tipo di serie mediante i quattro menu sulla sinistra.

La combinazione "serie originaria-valori assoluti-scala naturale" dei primi tre menu rappresenta il dato originale caricato dal sito indicato in precedenza. Le altre combinazioni derivano da semplici elaborazioni. Anche per questo "foglio di lavoro"

l'unica operazione a disposizione dell'utente (a parte la scelta della serie e delle zone geografiche) è poter copiare il grafico e incollarlo in un altro programma.

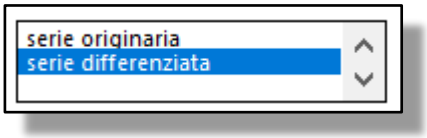


Nella selezione delle serie si deve tener conto di alcune informazioni fondamentali:

le seguenti serie del database originale si riferiscono a valori cumulati:

- ricoverati_con_sintomi
- terapia_intensiva
- totale_ospedalizzati
- isolamento_domiciliare
- totale_positivi
- dimessi_guariti
- deceduti
- totale_casi
- tamponi

Solo su queste serie, pertanto, ha senso passare da "serie originaria" a "serie differenziata".

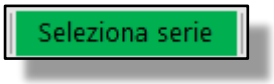


La serie originaria variazione_totale_positivi si riferisce, come dice il nome, alla differenza di totale_positivi con il giorno precedente. Ovvero: variazione_totale_positivi (tempo t)=totale_positivi (tempo t)-totale positivi (tempo t-1).

La serie originaria nuovi_positivi si riferisce alla differenza di totale_casi con il giorno precedente. Ovvero: nuovi_positivi (tempo t)=totale_casi (tempo t)-totale_casi (tempo t-1).
Conseguentemente, la combinazione "serie differenziata"-"totale_positivi" coincide con la "serie originaria"-"variazione_totale_positivi"; analogamente, serie differenziata"-"totale_casi" coincide con la "serie originaria"-" nuovi_positivi"

Inoltre valgono le seguenti relazioni:
ricoverati_con_sintomi+ terapia_intensiva= totale_ospedalizzati;
totale_ospedalizzati+ isolamento_domiciliare= totale_positivi;
totale_positivi+ deceduti+ dimessi_guariti= totale_casi.

Elaborazioni a fini previsivi di una serie scelta (etichette verdi)



Foglio "Seleziona serie"

Analogamente al foglio di lavoro "DUE zone", mediante il foglio "Seleziona serie" possiamo scegliere una (sola) serie da analizzare. Nei tre fogli successivi la serie selezionata viene elaborata in tre modi differenti.

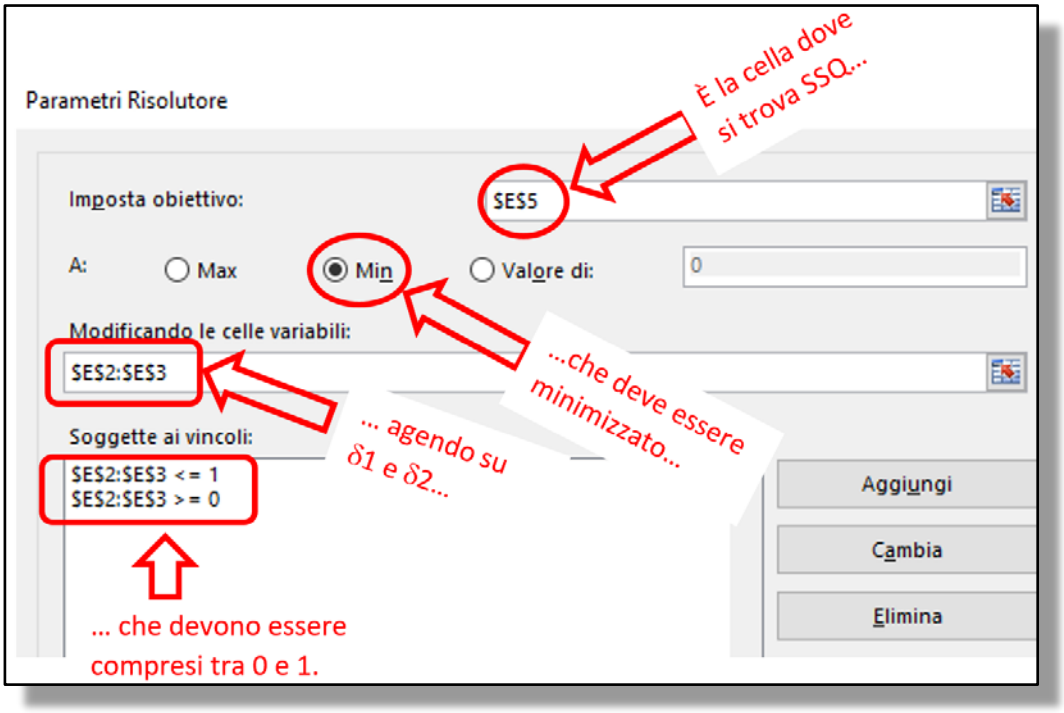
Foglio "Holt-Winters"

In questo foglio viene stimato un modello Holt-Winters di tipo semplificato: è presente il "livello" della serie e il suo "trend", ma non la "stagionalità". A questo proposito, non escludiamo di implementare il modello completo in un prossimo futuro, dato che, soprattutto ultimamente sembra presente un certo "effetto-domenica," o più in generale "effetto-giorno festivo" di cui sono affetti i dati prodotti dall'ISS. Alternativamente potremmo considerare di poter analizzare una serie di medie mobili (a 3 o 5 termini ad esempio), per smorzare qualunque effetto dovuto alla variabilità della tempestività con cui i fenomeni vengono misurati dalle serie storiche ufficiali.

Nel box riquadrato in rosso viene ricordato il modello adottato. Riassumendo:

1. la previsione per il livello della serie al tempo $t+1$ deriva dalla somma di due sole componenti: livello m_t più trend T_t;
2. ad ogni istante t
   - il livello m_t è la media ponderata (o combinazione lineare convessa) secondo δ_1 di due parti: m_{t-1} e T_{t-1};
   - il trend T_t è la media ponderata (o combinazione lineare convessa) secondo δ_2 di due parti: incremento del livello rispetto al tempo precedente (m_t-m_{t-1}) e trend precedente (T_{t-1}).

I parametri $\delta 1$ e $\delta 2$ sono individuati minimizzando la SSQ tra valori osservati e valori *fitted*. Per fare questo si può utilizzare lo strumento Risolutore di Excel, impostandolo secondo quanto è riassunto nella finestra qui sotto:



Lanciando il Risolutore si dovrebbero ottenere i parametri "ottimali", che saranno in generale diversi per ogni serie analizzata.

Nel foglio compaiono le previsioni per i successivi k giorni, dove k è scelto nel menu a discesa. Il giorno 0 è l'ultimo giorno per il quale ci sono i dati nell'archivio. Pertanto per il giorno 0, oltre il valore "da modello", si conosce anche il valore "osservato".

Per la stima di δ_1 e δ_2 sono usati tutti i valori disponibili, dal 24 febbraio in poi. In futuro, probabilmente verrà aggiunta l'opzione di limitarsi a utilizzare gli ultimi n giorni, con n scelto dall'utente.

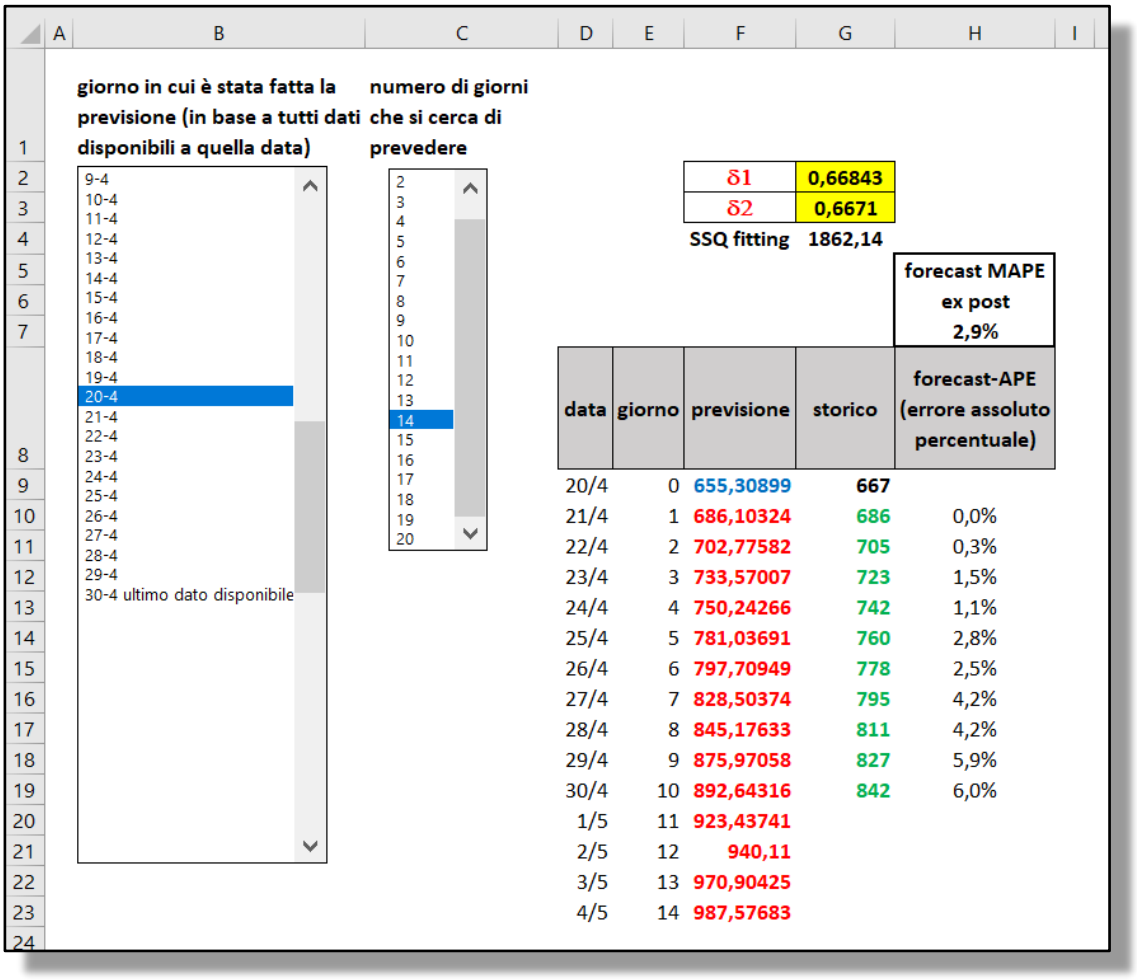
Foglio H-Wexpost

Anche in questo foglio la serie selezionata in "Seleziona serie" viene analizzata secondo lo stesso modello Holt -Winters di tipo semplificato visto in precedenza. La sola (importante) differenza è che in questo foglio ci si può posizionare fittiziamente in un giorno precedente all'ultimo per il quale sono disponibili i dati, e poi, coerentemente, stimare δ_1 e δ_2 con i dati che erano disponibili fino a quel giorno. Prolungando poi la previsione per k giorni, si può confrontare la previsione con i valori che si sono successivamente verificati.

Per chiarire meglio il funzionamento (che comunque è segue l'usuale logica del test di *goodness of forecast* condotto a posteriori) facciamo un esempio, corrispondente alla schermata riportata qui sotto (l'esempio si riferisce a "deceduti in Toscana (valori cumulati)".

Supponiamo di avere a disposizione i dati fino al 30 aprile. Fissiamo la data "fittizia" dati al 20 aprile: in altri termini, "facciamo finta" di trovarci al 20 aprile, disponendo quindi dei soli dati disponibili fino a quel giorno (incluso). Lanciando il Risolutore (in questo caso, come si vede dalla figura, le soluzioni per la minimizzazione

della cella SSQ che si trova in G4 sono contenute nelle celle G2 e G3), abbiamo trovato i parametri δ_1 e δ_2 ottimi pari, rispettivamente, a 0,66843 e 0,6671. Sulla base di questi parametri possiamo calcolare le stime per i 14 giorni successivi (dato che nel secondo menu a discesa abbiamo selezionato "14"); fino al 4 maggio dunque. Dato che in realtà disponiamo dei valori fino al 30 aprile, possiamo confrontare le previsioni coi valori che abbiamo osservato dal 21 aprile al 30 aprile (in verde). Sulla base di questi 10 giorni che ci permettono un confronto previsione-storico, possiamo verificare il MAPE (Mean Absolute Percentage Error) di previsione. Ovviamente, le serie di valori "cumulati" (come nel caso dell'esempio qui riportato) avranno performance molto migliori rispetto alle serie di valori "giornalieri". Analoghe considerazioni possono essere fatte per scala naturale-scala logaritmica. Alla luce di ciò dovremo calibrare le nostre frustrazioni ed entusiasmi nel trovare MAPE enormi o ridotti.



Rette mobili

Il foglio rette mobili produce un grafico in cui, ad ogni tempo, viene rappresentato un segmento ricavato da una retta dei minimi quadrati (variabile esplicativa è il tempo). Ciascuna retta è stimata sui dati della solita serie storica selezionata in "Seleziona serie".

Per ogni tempo *t*:

- la retta è calcolata sui m valori della serie, di cui l'ultimo è quello del tempo t; m è scelto mediante il primo menu, e va da 3 a 15;
- il segmento della retta dei m.q. inizia al tempo t;
- il segmento della retta dei m.q. finisce al tempo al tempo $t+k$, dove k è selezionato dal secondo menu: ad esempio "segmento 8" significa che la lunghezza di ciascun segmento è 8, ovvero che va da t a $t+8$.