Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per l'Economia e l'Impresa

RELAZIONE FINALE

Un modello per l'evoluzione del COVID-19: un'applicazione con dati italiani

Relatore: Prof.ssa Luisa Bisaglia

Dipartimento di Scienze Statistiche

Laureando: Mattia Bruno

Matricola: 1172412

Anno Accademico 2019/2020

Indice

1 Introduzione **2**

2 Fatti stilizzati **4**

2.1 Presentazione dei dati 4

2.2 Limiti dei dati 5

2.3 Analisi descrittive 9

3 Metodologia **17**

3.1 Background: modelli SIR 17

3.2 Modello autoregressivo Poisson log-lineare 20

3.2.1 Proprietà del processo e simulazioni 21

3.2.2 Generalizzazione e processo INGARCH 24

3.3 Inferenza 26

3.4 Procedura di stima 28

4 Applicazione e risultati **29**

4.1 Dati nazionali 30

4.1.1 Stima sul campione completo 30

4.1.2 Stima dinamica 32

4.1.3 Stima del numero di riproduzione 34

4.2 Dati regionali 35

4.3 Estensioni e sviluppi futuri 36

5 Conclusioni **39**

A Codice R **41**

Bibliografia **51**

Capitolo 1

Introduzione

L'epidemia di COVID-19 è dovuta a un nuovo coronavirus da sindrome respiratoria acuta grave, denominato SARS-CoV-2. I primi casi accertati sembrerebbero risalire al dicembre 2019, identificati a Wuhan, città della provincia di Hubei in Cina. Per quanto riguarda l'Italia, i primi casi isolati sono relativi a due turisti cinesi in visita a Roma a fine gennaio, mentre il virus si diffonde in tutto il Paese in un secondo momento, a partire dai focolai d'infezioni di Vo' (Veneto) e Codogno (Lombardia). Qui, i primi contagi certificati risalgono al 20 febbraio 2020, anche se alcuni studi sembrano dimostrare che il virus circolasse nel Nord Italia già da prima (si veda ad esempio Cerqua e Di Stefano (2020)).

Successivamente, l'epidemia arriva a colpire tutto il mondo, tanto che l'11 marzo 2020 la *World Health Organization* (WHO) dichiara il coronavirus pandemia mondiale. La rapida diffusione del virus provoca una crisi in diversi Paesi, i quali non sono preparati a gestire una situazione così fuori dall'ordinario, stante le scarse conoscenze disponibili sul fenomeno e vista la mancanza di cure specifiche e di un vaccino. In Italia in particolare, la situazione d'emergenza riguarda il sistema sanitario, messo sotto pressione dell'elevato numero di contagiati gravi da ricoverare in terapia intensiva, come dimostra il caso della regione Lombardia. Tutto ciò, induce il Governo Italiano ad emanare misure di contenimento via via sempre più restrittive, quali distanziamento sociale, quarantena e chiusura delle attività commerciali, che

provocano gravi conseguenze economiche e sociali.

A fine luglio, essendo il COVID-19 un fenomeno ancora in corso, i dati rivestono un ruolo di vitale importanza per monitorare l'andamento dell'epidemia. In questa fase, in Italia, i dati sono fondamentali affinchè non si sviluppi una seconda ondata epidemica ed eventualmente per gestirla; mentre all'inizio della pandemia lo erano per verificare gli effetti dei provvedimenti adottati. Per entrambe le cause, è utile disporre di un modello *data-driven* che possa mostrare in maniera dinamica l'evoluzione del contagio.

A tal fine, questo lavoro prende le mosse dal paper "Arianna Agosto, Alexandra Campmas, Paolo Giudici e Andrea Renda. *Monitoring COVID-19 contagion growth in Europe.* CEPS Working Papers, 30 marzo 2020", nel quale viene presentato un modello di Poisson autoregressivo (PAR), che viene applicato ai nuovi casi giornalieri di alcuni Paesi europei. L'obiettivo della tesi è quello di estendere la metodologia proposta ai dati italiani fino al 31 luglio, con in aggiunta una breve applicazione ad alcune regioni d'interesse. Lo scopo non è quindi tanto quello di prevedere il numero di nuovi casi giornalieri, che in questo periodo risulta altamente imprevedibile per vari fattori; quanto quello di analizzare e monitorare la diffusione dell'epidemia, attraverso un modello che possa tenere conto di *shock* dovuti per esempio allo sviluppo di nuovi *cluster* d'infezioni.

Il lavoro è strutturato come segue. Nel Capitolo 2 si introducono i dati italiani e si effettuano alcune analisi descrittive su di essi. Nel Capitolo 3 si descrive il modello utilizzato dal punto di vista teorico. Nel Capitolo 4 se ne presenta un'applicazione empirica con i nuovi casi giornalieri italiani e quelli di alcune regioni d'interesse. Infine, il Capitolo 5 conclude.

Capitolo 2

Fatti stilizzati

2.1 Presentazione dei dati

Per l'Italia, la principale fonte ufficiale di informazioni quantitative riguardo il COVID-19 è il Dipartimento della Protezione Civile, il quale pubblica quotidianamente, attorno alle ore 18, vari dati in forma di serie storiche alla pagina `https://github.com/pcm-dpc/COVID-19` . A partire dal 24 febbraio 2020, questi dati vengono prima inseriti dalle singole autorità regionali in una piattaforma web apposita e poi diffusi a diversi livelli di aggregazione.

Per le analisi, le serie storiche principali sono quelle dei casi totali, del totale dei deceduti, del totale dei dimessi/guariti, degli attualmente positivi e dei soggetti ricoverati in terapia intensiva. Tutti i dati delle variabili elencate sono presenti sia a livello nazionale che regionale, mentre al livello di aggregazione provinciale è disponibile solamente il numero dei casi totali. Le prime tre serie rappresentano grandezze cumulate per cui sono non negative e dovrebbero essere non decrescenti, mentre le ultime due si riferiscono a persone alle prese con il COVID-19 in un giorno specifico, quindi restano non negative ma possono essere decrescenti. In particolare, il numero di attualmente positivi viene derivato sottraendo al numero dei casi confermati quello dei deceduti e dei dimessi/guariti in questo modo:

$$\text{attualmente positivi} = \text{casi confermati - deceduti - dimessi/guariti}$$

Invece, differenziando le serie se ne ottengono le variazioni giornaliere, le quali sono di maggiore interesse perchè descrivono in maniera più chiara l'evolversi dell'epidemia. Di conseguenza, i *nuovi attualmente positivi*, che rappresentano la variazione giornaliera degli attualmente positivi, vengono definiti come i *nuovi casi confermati* al netto dei *nuovi deceduti* e dei *nuovi dimessi/guariti*. Si noti che tale misura non rappresenta i nuovi positivi rispetto al giorno precedente poichè questi vengono determinati dai *nuovi casi confermati*, i quali rappresentano la principale variabile d'interesse da modellare per quanto riguarda i contagi. I nuovi casi giornalieri dovrebbero essere non negativi, il che ha importanti implicazioni per il modello da adottare. I *nuovi attualmente positivi* non rappresentano invece una grandezza osservata direttamente, ma derivano completamente dalle altre tre variabili. Anch'essi sono informativi ai fini delle analisi in quanto riescono a indicare il livello di pressione sul sistema sanitario in un giorno specifico.

Prima di iniziare le analisi, è opportuno comprendere le criticità che presentano i dati che si hanno a disposizione.

2.2 Limiti dei dati

Il processo generatore dei dati del Dipartimento della Protezione Civile è instabile, in quanto all'incertezza insita nel fenomeno epidemico si aggiunge quella risultante dal processo di raccolta dei dati. Questi ultimi non sono privi di errori, difatti risultano "rumorosi" e talvolta manifestano incongruenze, in quanto ci sono, per esempio, alcuni casi di regioni in cui le variazioni giornaliere dei deceduti sono negative. Queste caratteristiche riflettono diversi fattori, quali errori di registrazione e comunicazioni tardive o parziali delle informazioni da parte delle Regioni. Infatti, esiste un *trade-off* tra tempistività e accuratezza dei dati: per avere a disposizione i dati ogni giorno così come avviene, si rinuncia inevitabilmente a qualcosa in termini di qualità. Inoltre, è chiaro che in un periodo di emergenza come quello in cui ci si è trovati, cercare di salvare delle vite ha la priorità rispetto ad avere informazioni statistiche impeccabili.

Partendo dalle problematiche relative ai dati riguardanti i casi confermati, va subito messo in evidenza il fatto che il totale dei casi confermati non è equivalente a quello dei contagiati reali, bensì al totale delle persone che sono risultate positive al tampone. Infatti, queste sono minori del totale degli infetti reali, per cui il numero dei contagiati viene sottostimato, in quanto è stata esclusa una buona parte degli asintomatici. Ciò è dovuto al fatto che le persone sottoposte a tampone non provengono da un disegno di campionamento specifico ma, per gran parte dell'epidemia, sono state selezionate secondo un criterio "di convenienza", che privilegiava sintomatici e contatti degli infetti.

A partire dal 18 maggio, con l'inizio della cosiddetta "Fase 2", sicuramente è aumentato il numero di asintomatici sottoposti a tampone. Purtroppo, il dato dei casi confermati è difficile da interpretare, perchè non si sa cosa rappresenti esattamente questo conteggio, visto che si conosce poco o nulla sui criteri tramite i quali vengono selezionate le persone da testare, specie per quel che riguarda gli asintomatici. Per tali ragioni, risulta difficile determinare il numero totale delle persone che si sono contagiate in Italia: per sopperire a ciò è stata avviata un indagine sulla sieroprevalenza da parte di Ministero della Salute e Istat.[1]

Sulla base delle considerazioni precedenti, è evidente come il dato dei casi confermati risenta del numero di tamponi effettuati perchè, banalmente, i casi positivi trovati aumentano al crescere del numero di tamponi eseguiti. Inoltre, si nota una notevole eterogeneità fra le regioni per quanto riguarda le politiche di somministrazione dei test, in quanto ogni Regione ha adottato regole proprie in tal senso, regole che probabilmente sono anche cambiate nel tempo. Basti pensare alle differenze tra il Veneto, nel quale si sono effettuati tamponi "a tappeto", e la Lombardia, che nella fase critica permetteva di eseguire il tampone solo nei casi più gravi. Pertanto, bisogna prestare attenzione a confrontare i contagi in aree diverse, perchè non si sa quali strategie sono state adottate per misurare il fenomeno.

Un altro punto critico riguarda il fatto che i test non sono perfettamente affidabili, infatti si può verificare che un tampone dia esito negativo anche se

[1] si veda `https://www.istat.it/it/archivio/242676`

il soggetto testato risulta positivo al virus. Infine, è opportuno notare che sulla carta la data in cui un caso sintomatico viene dichiarato positivo dovrebbe corrispondere a quella dell'inizio dei sintomi. Nella realtà, chiaramente non è così, in quanto passa diverso tempo dai primi sintomi all'esecuzione del tampone, anchè perchè gli operatori sanitari sono in grado di effettuare un numero limitato di tamponi. Peraltro, vi può essere asincronismo anche tra la data in cui un caso viene dichiarato positivo e la data in cui il tampone viene effettuato. Oltre che alle tempistiche di comunicazione dei risultati, questo ritardo è dovuto al tempo che il tampone necessita per essere analizzato: questa operazione poteva richiedere più giorni nel periodo più critico dell'epidemia, a causa del sovraccarico di lavoro per i laboratori. Per questi motivi, principalmente nel periodo più critico dell'epidemia, i dati sui casi possono risultare differiti rispetto alla situazione reale.

Alla luce di tutte queste considerazioni, i dati sui deceduti e sui ricoverati in terapia intensiva sembrerebbero preferibili per analizzare l'epidemia rispetto a quelli sui contagi. Però, bisogna tenere conto che anche questi dati presentano alcune criticità.

Il numero di decessi pare anch'esso sottostimato, come dimostra il rapporto sulla mortalità dei primi cinque mesi del 2020, condotto congiuntamente dall'Istituto nazionale di statistica (Istat) e dall'Istituto Superiore di Sanità (ISS).[2] Infatti, questo mostra che il numero di decessi di marzo-aprile 2020 è maggiore della media di questi mesi nei cinque anni precedenti e la differenza fra deceduti del 2020 e media dello stesso periodo dei cinque anni precedenti è maggiore del numero di decessi fornito dalla Protezione Civile. Questo scostamento tra eccesso di mortalità e numero di decessi COVID-19 è notevole ed è dovuto a:

- decessi causati direttamente dal COVID-19, ma non certificati. Tali decessi comprendono tutte le persone che sono morte a causa del coronavirus, ma non sono state considerate infette perchè non sottoposte a tampone, neanche *post-mortem*;

[2] si veda `https://www.istat.it/it/archivio/245415`

- decessi causati indirettamente dal COVID-19, i quali non sarebbero avvenuti in condizioni normali e che sono presumibilmente conseguenza delle condizioni critiche in cui ha dovuto operare il sistema sanitario in alcune regioni.

Per quanto concerne invece i ricoveri in terapia intensiva, un limite riguarda il fatto che questo dato è accurato solo nel caso in cui vi siano posti disponibili nelle terapia intensive. Difatti, in un momento d'emergenza, in cui le strutture sanitarie si saturano, vengono escluse dal conteggio tutte le persone che avrebbero bisogno di ricovero in terapia intensiva, ma per le quali non vi è più posto. Inoltre, è opportuno notare che la variazione di questa quantità rappresenta il saldo giornaliero tra nuovi ingressi in terapia intensiva e uscite (con esito positivo o negativo), dal quale non si può derivare il numero di nuovi ingressi, in quanto non si hanno informazioni riguardo a quante delle persone ricoverate decedono o vengono dimesse giornalmente.

I dati sui ricoverati in terapia intensiva descrivono una situazione leggermente differita rispetto a quelli sui casi confermati poichè rappresentano solo gli infetti che sono arrivati a versare in condizioni gravi. Stesso discorso vale per i dati sui decessi, i quali per ovvie ragioni presentano uno sfasamento temporale molto più importante. Perciò, i dati sui contagi, al netto di tutte le criticità esposte in precedenza, sono quelli che in teoria riflettono più tempestivamente l'andamento dell'epidemia. Un modello per le serie storiche dei nuovi casi giornalieri riesce infatti a cogliere più prontamente *shock* dovuti a sviluppo di nuovi *cluster* d'infezioni o cambiamenti nelle misure di contenimento. D'altra parte, i dati sui contagi sono necessari per stimare il numero di riproduzione di base R_0, misura fondamentale dell'andamento del virus, di cui si parlerà nel capitolo successivo.

Dunque, nel seguito, preso atto degli importanti limiti appena descritti, si sceglie di modellare la serie dei nuovi casi giornalieri. Prima però, si presentano delle analisi descrittive volte a comprendere in maniera completa il fenomeno, cercando di andare oltre un'analisi basata esclusivamente sul numero di contagi. A tal proposito tutti i grafici sono stati realizzati con RStudio versione 1.3.1073, il codice utilizzato è riportato in Appendice A.

2.3 Analisi descrittive

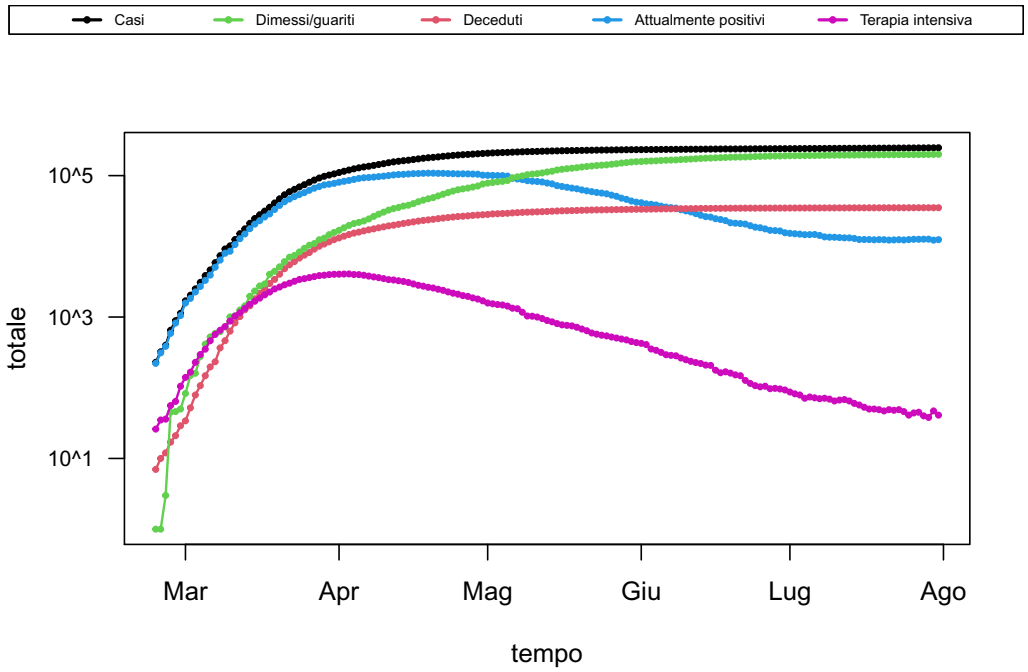

Figura 2.1: Serie storiche del numero totale di casi, dimessi/guariti, deceduti, attualmente positivi e ricoverati in terapia intensiva (scala semilogaritmica)

In Italia, al 31 luglio, il numero totale di casi confermati risulta 247.537, di cui 199.974 dimessi/guariti, 35.141 deceduti e 12.422 attualmente positivi, tra i quali 41 ricoverati in terapia intensiva. La figura 2.1 riassume l'andamento complessivo dell'epidemia mostrando l'evoluzione del numero totale di casi confermati, dimessi/guariti, deceduti, attualmente positivi e ricoverati in terapia intensiva. Il grafico è in scala semilogaritmica perchè questa aiuta a cogliere meglio le tendenze di lungo periodo rispetto a una scala lineare. Si nota infatti che le curve dei contagiati e dei decessi hanno rallentato notevolmente la loro crescita, che in un primo momento era esponenziale, e che fortunatamente una buona parte dei casi totali risulta guarita.

In figura 2.2 sono rappresentate, invece, le variazioni giornaliere di casi confermati, dimessi/guariti e deceduti. Le serie originali sono descritte dalle linee più sottili, mentre le linee più spesse sono le medie mobili semplici a 7

giorni, le quali sostanzialmente rappresentano la variazione media settimanale. Si osserva che i nuovi casi giornalieri hanno raggiunto il picco verso fino marzo e da lì hanno iniziato a decrescere lentamente. Le variazioni dei decessi ricalcano quelle dei contagiati su una scala minore. Il lisciamento ha la funzione di mettere in risalto i trend, riducendo la componente di rumore (dovuta in parte agli errori di misurazione descritti in precedenza) e rimuovendo effetti stagionali. Infatti, la serie storica dei nuovi casi giornalieri è caratterizzata da stagionalità, come si nota in particolare dai lag 8 della PACF in figura 2.4. Questo effetto settimanale è dovuto al fatto che nei weekend i laboratori che analizzano i tamponi sono chiusi e questo si riflette sui nuovi casi giornalieri della domenica e del lunedì, i quali risultano ogni settimana minori rispetto agli altri giorni.

Invece, in figura 2.3 viene rappresentato, a partire dal 21/04/2020, l'andamento giornaliero del tasso di positività, definito come rapporto tra i casi positivi e i soggetti testati. Si osserva che questo tasso è diminuito nel tempo, mentre negli ultimi due mesi è rimasto pressochè costante, denotando una diffusione del contagio stabile.

Bisogna anche notare che l'epidemia ha avuto un'evoluzione diversa nelle varie zone del Paese, colpendo alcune regioni in modo particolare rispetto alle altre. Tant'è vero che più del 70% dei casi si concentra in Lombardia, Emilia-Romagna, Piemonte e Veneto (queste regioni hanno circa il 40% degli abitanti italiani), mentre circa la metà dei decessi è avvenuta in Lombardia.

Le variazioni giornaliere di casi e decessi denotano grande eterogeneità fra le regioni, sia dal punto di vista della scala che da quello dell'evoluzione temporale, come si evince dal grafico in figura 2.5. In questo vengono rappresentati, per le otto regioni più popolose d'Italia, gli andamenti dei casi e dei decessi giornalieri su due differenti scale. Siccome i dati regionali risultano ancora più erratici di quelli nazionali, vengono presentate solo le medie mobili semplici a 7 giorni, al fine di delineare in modo chiaro il segnale.

I confronti vanno effettuati con grande cautela per quel che riguarda i contagi, poichè, come evidenziato in precedenza, le strategie di tampone sono state diverse tra le regioni ed entro le regioni, quindi vi è disomogeneità spazio-temporale nella misurazione dei contagi. Detto ciò, si può constatare

che la tendenza dei decessi giornalieri si rivela sfasata temporalmente rispetto a quella dei contagi. La Campania invece è l'esempio che i dati talvolta manifestano incongruenze, come si può vedere dal numero di casi giornalieri, il quale diventa negativo in un frangente a causa di alcuni ricalcoli.

Per quanto riguarda i ricoveri in terapia intensiva, la dinamica delle variazioni giornaliere è mostrata nel grafico in alto in figura 2.6, dove si nota che dalla linea blu, che corrisponde al 3 aprile, il saldo tra ingressi e uscite diventa negativo. Inoltre, dal grafico in basso in figura 2.6 si vede che, sempre dal 3 aprile (linea blu), il rapporto tra i decessi giornalieri e i ricoverati in terapia intensiva 8 giorni prima rimane costante attorno al 14% per il resto del tempo (fonte dell'analisi: Facebook Enrico Rettore). L'interpretazione di queste due evidenze empiriche è che dai primi di aprile sia diminuita la pressione sul sistema sanitario e che quindi le terapie intensive siano state in grado di accogliere chiunque avesse bisogno di quel tipo di ricovero. Conseguentemente, il 14% rappresenta la percentuale di decessi tra i ricoverati delle terapie intensive, se queste sono in grado di lavorare in condizioni normali.

Per quel che concerne le modalità di trattamento dei soggetti attualmente positivi, la figura 2.7 ne descrive il cambiamento nel tempo. Infatti, rispetto al picco della curva degli attualmente positivi, si osserva che sono diminuite in maniera significativa le proporzioni di infetti ricoverati con sintomi e in terapia intensiva. Questo è in parte dovuto al fatto che una frazione importante dei soggetti risultati positivi al tampone nell'ultimo periodo siano asintomatici e non abbiano necessitato di cure ospedaliere. Infatti, recentemente, molti dei tamponi sono stati eseguiti come conseguenza dell'attività di tracciamento dei contatti e di quella di *screening*. Tale interpretazione è supportata dai report pubblicati dell'Istituto Superiore di Sanità, i quali da giugno contengono una sezione che illustra per quale motivo sono stati testati i casi diagnosticati. In tabella 2.1, si vede che circa il 68% dei casi positivi dal 6 al 19/07/2020 deriva dal tracciamento dei contatti e dallo *screening*.[3]

Quindi, pare lecito ipotizzare che i casi confermati rappresentino una proxy dei sintomatici sino all'inizio della "Fase 2", da quando comprendono anche una quota rilevante di asintomatici.

[3]si veda il report esteso alla pagina `https://www.epicentro.iss.it`

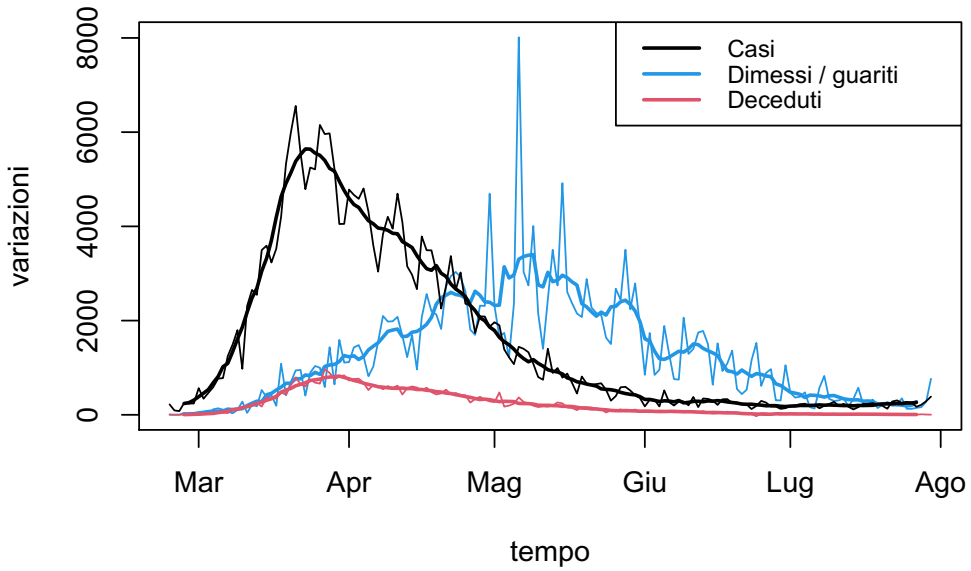

Figura 2.2: Variazioni giornaliere: serie storiche originali (linee sottili) e medie mobili semplici a 7 giorni (linee spesse)

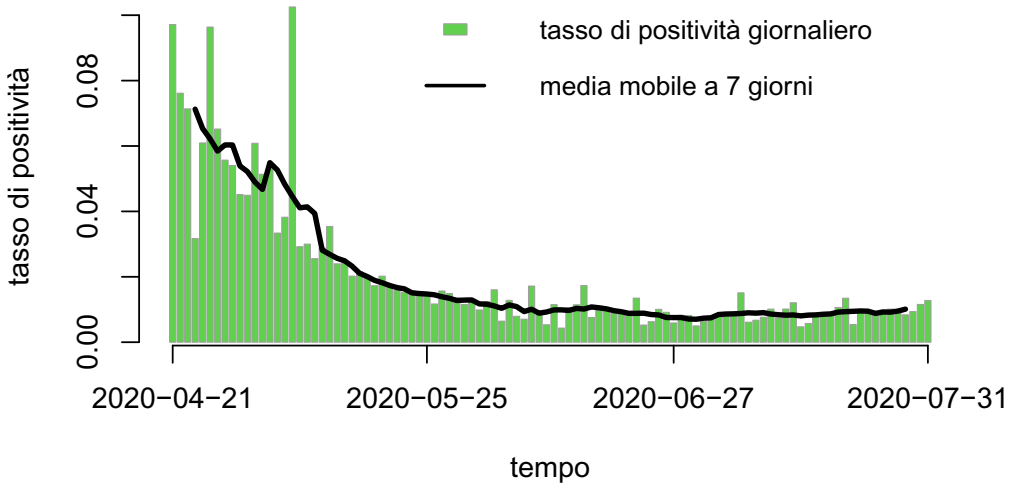

Figura 2.3: Evoluzione del tasso di positività giornaliero e della rispettiva media mobile semplice a 7 giorni. Il tasso di positività giornaliero viene definito come rapporto tra i nuovi casi positivi e i soggetti testati in un determinato giorno

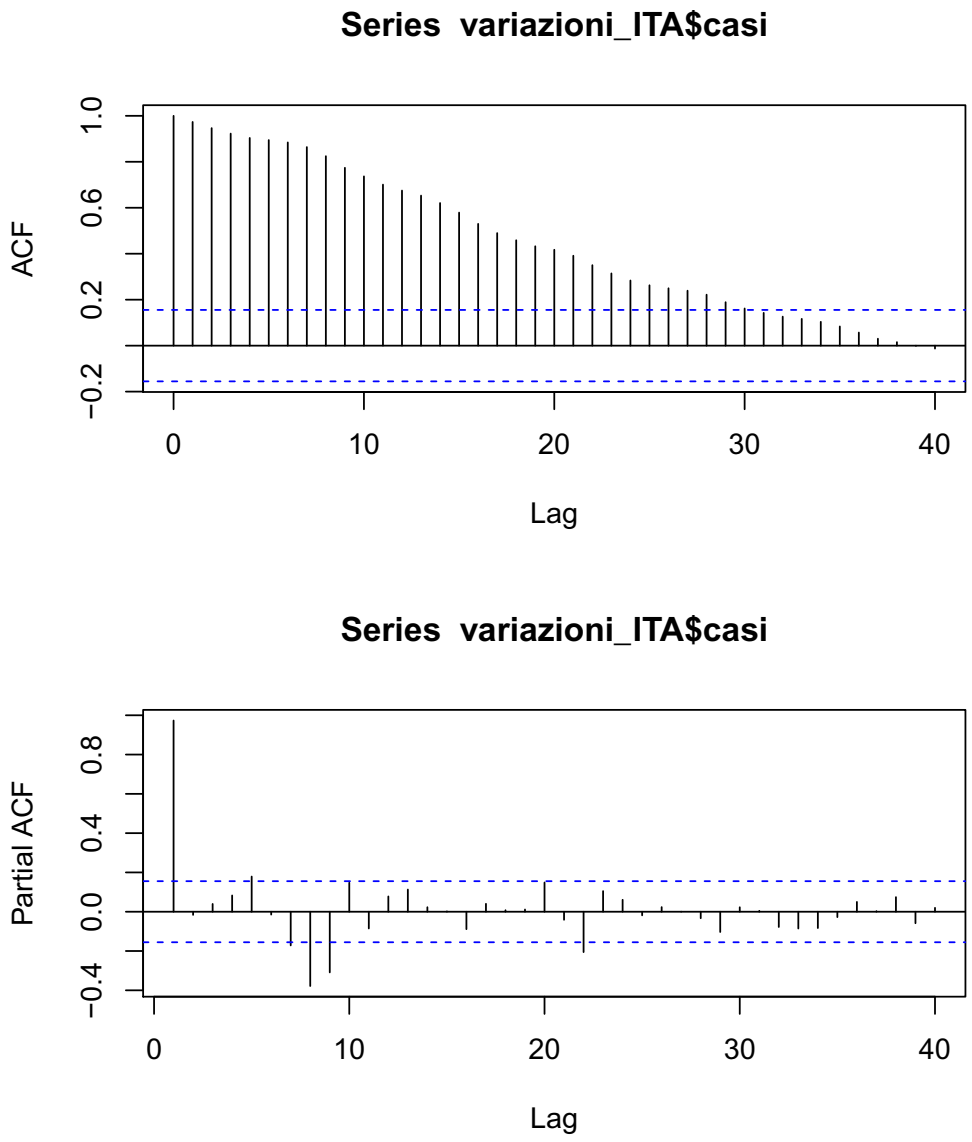

Figura 2.4: Funzioni di autocorrelazione empiriche della serie dei nuovi casi giornalieri

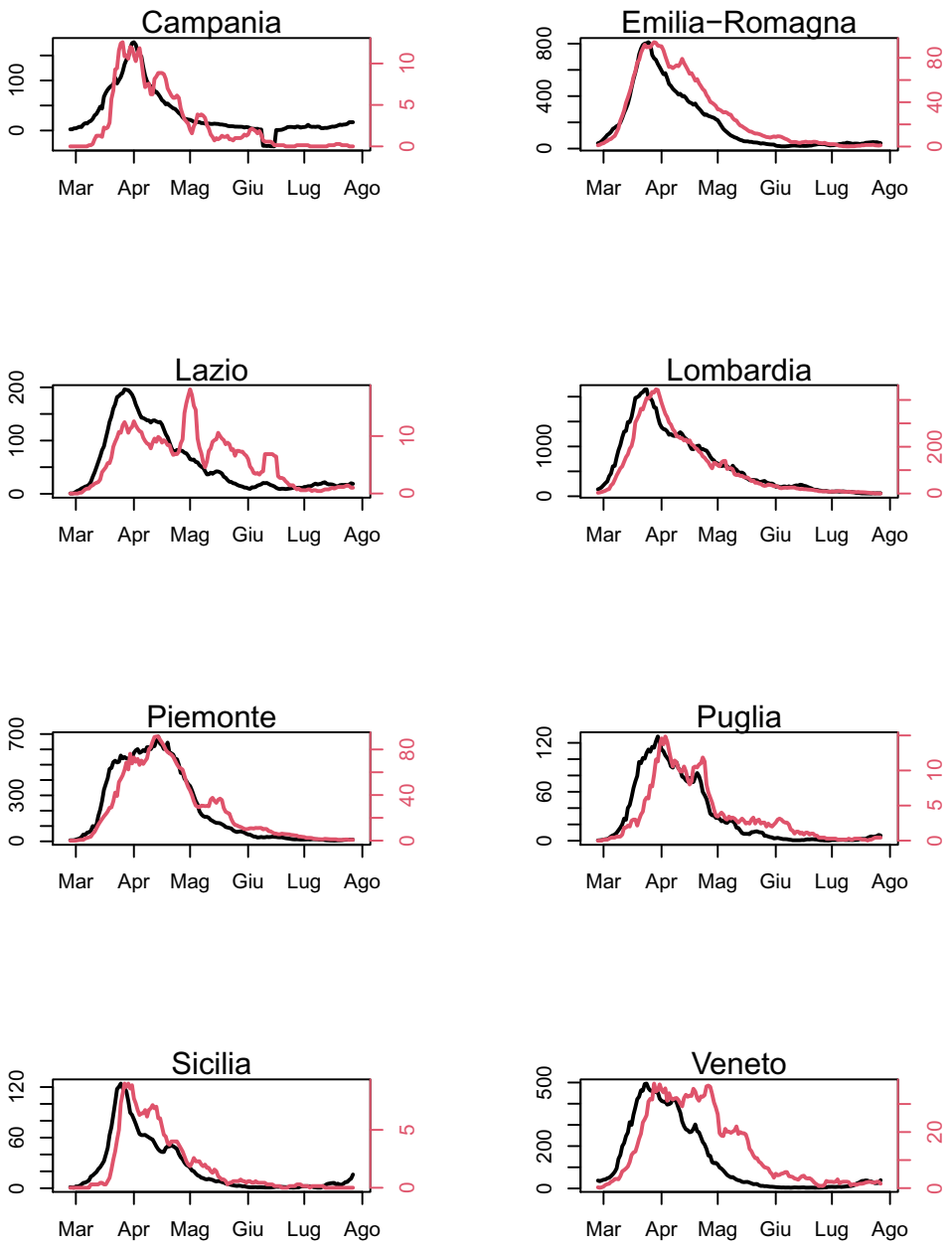

Figura 2.5: Medie mobili semplici a 7 giorni delle variazioni giornaliere regionali dei casi (in nero sulla scala a sinistra) e dei decessi (in rosso sulla scala a destra)

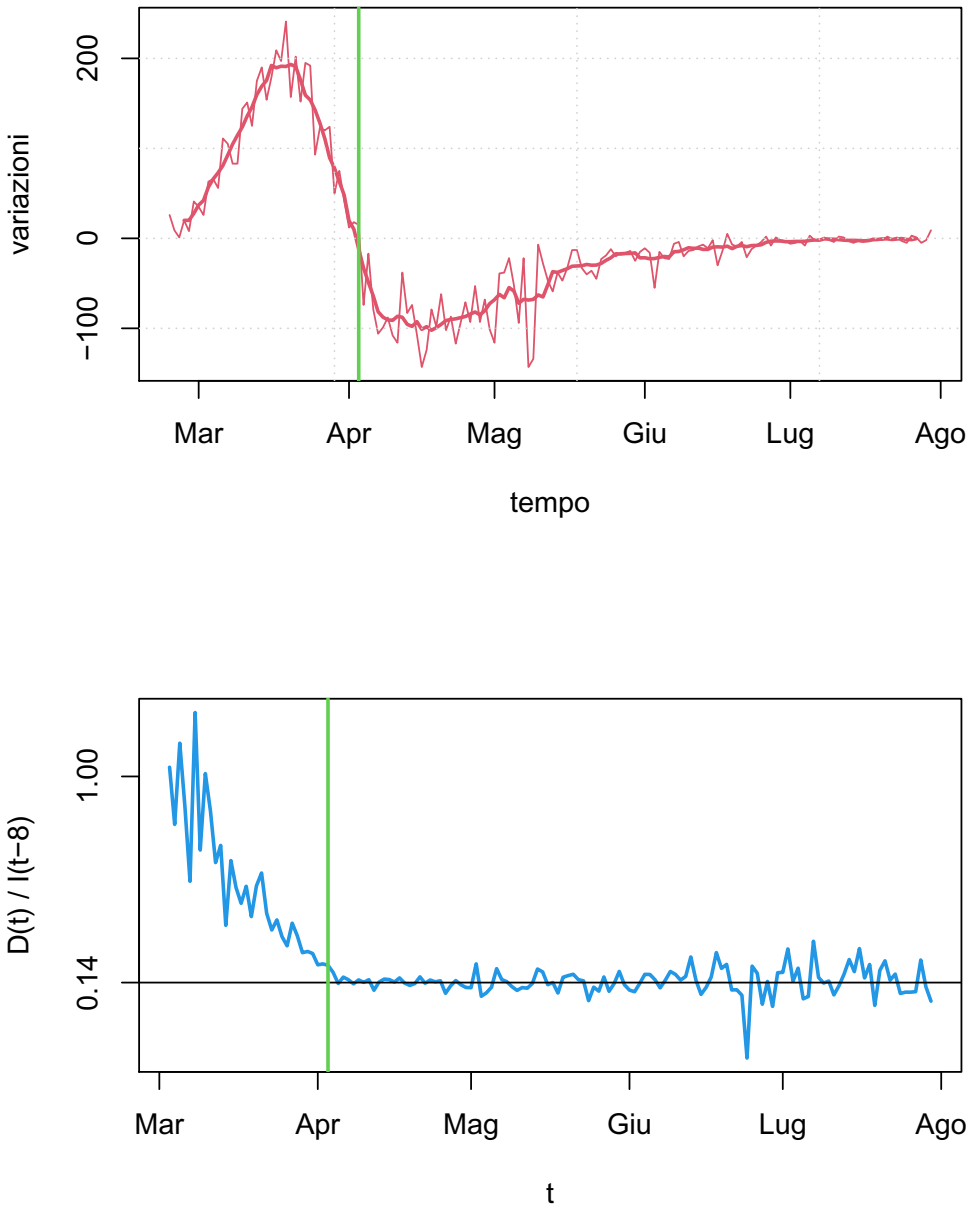

Figura 2.6: In alto: variazioni giornaliere dei ricoveri in terapia intensiva e medie mobili semplici a 7 giorni.
In basso: evoluzione del rapporto tra i decessi giornalieri e i ricoveri in terapia intensiva di 8 giorni prima

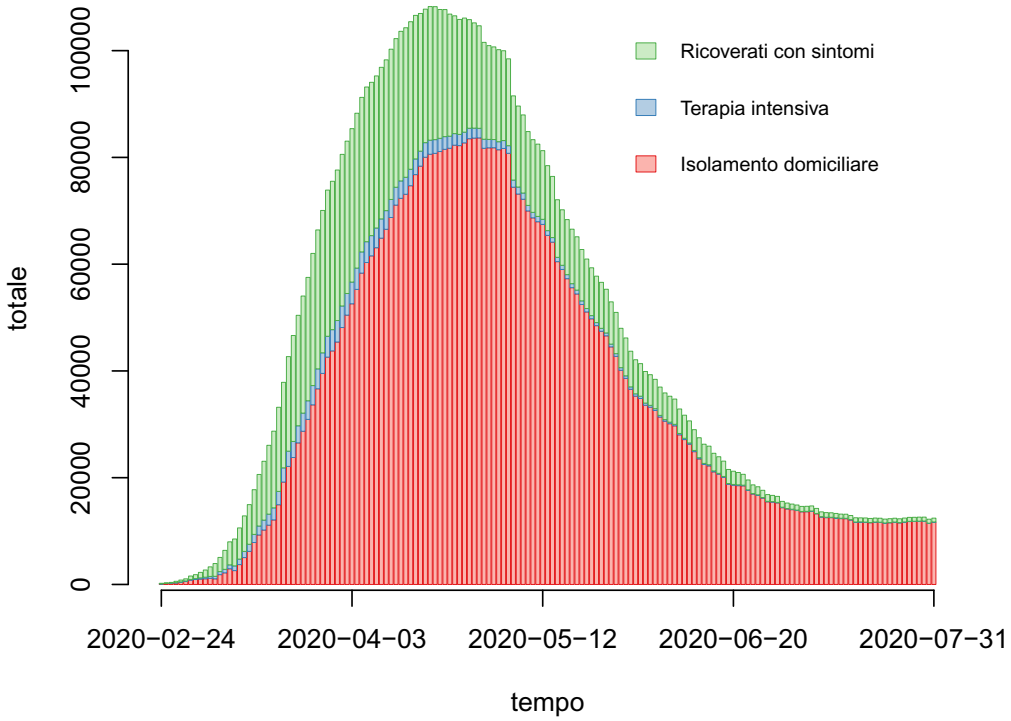

Figura 2.7: Suddivisione delle modalità di trattamento dei soggetti attualmente positivi

Tabella 2.1: Distribuzione del motivo per cui sono stati testati i casi diagnosticati dal 6 al 19/07/2020 (N=2.746 casi)

Motivo del test	Casi	
	N	%
Contact tracing	663	24.1
Caso sintomatico	809	29.5
Screening	1.204	43.8
Non noto	70	2.6
Totale	2.746	100

Capitolo 3

Metodologia

Il modello utilizzato da Agosto et al. (2020) mira a descrivere l'evoluzione del contagio, tenendo in considerazione la dipendenza temporale tra i nuovi casi giornalieri. L'approccio adottato estende la classe di modelli epidemiologici SIR (*Susceptible Infected Recovered*) proposta da Kermack e McKendrick (1927), introducendo una componente autoregressiva nella dinamica della crescita del totale dei casi. Pertanto, prima di procedere nello specifico, si presentano i modelli esponenziali impiegati all'interno della metodologia SIR.

3.1 Background: modelli SIR

Nella letteratura, la misura fondamentale dell'andamento dell'epidemia è il *numero di riproduzione di base* R_0, parametro che rappresenta il numero medio di infezioni secondarie prodotte da ciascun individuo infetto in una popolazione mai venuta a contatto con il virus. In base al valore assunto da questo indice, si possono distinguere tre fasi diverse dell'epidemia:

- $R_0 > 1$: ogni individuo infetto contagia in media più di una persona, l'epidemia è in una fase d'espansione;
- $R_0 = 1$: ogni individuo infetto contagia in media una persona, la diffusione dell'epidemia rimane stabile;

- $R_0 < 1$: ogni individuo infetto contagia in media meno di una persona, l'epidemia è in una fase di contrazione.

Gu et al. (2020) hanno mostrato che R_0 si può calcolare come segue:

$$R_0 = \delta * \mathbb{E}[T] \tag{3.1}$$

La variabile casuale $T > 0$ rappresenta il tempo d'incubazione della malattia, ovvero il periodo per il quale l'individuo contagiato è infettivo. T segue una distribuzione Gamma, il cui valore atteso è $\mathbb{E}[T] = 7.5$ per il COVID-19, come hanno dimostrato Shao et al. (2020). Di conseguenza, la durata media dell'infettività è di 7.5 giorni. Invece, δ rappresenta il tasso di riproduzione istantaneo: indica il numero di persone che un individuo infetta in media per ciascuno dei giorni in cui è contagioso.

La stima del parametro δ è basata sul modello esponenziale impiegato nell'ambito della metodologia SIR, il quale assume che i nuovi casi giornalieri y_t abbiano distribuzione $y_t \sim \text{Poisson}(\lambda_t)$ con intensità determinata da:

$$\lambda_t = \lambda_0 \exp(\gamma t) \tag{3.2}$$

Il parametro γ viene solitamente stimato attraverso una regressione lineare del logaritmo dei nuovi casi giornalieri sul tempo:

$$\log(y_t) = \kappa + \gamma t + \varepsilon_t \,, \quad \varepsilon_t \sim N(0, \sigma^2) \tag{3.3}$$

dove $\kappa = log(\lambda_0)$ è l'intercetta e t rappresenta il tempo, espresso in giorni dall'inizio dell'epidemia. Chiaramente, solo in una prima fase dell'epidemia, si puo assumere che il numero dei nuovi casi cresca esponenzialmente ad un tasso γ costante, dopodiché questo parametro varia nel tempo. Per cogliere meglio il progressivo cambio di pendenza della curva del logaritmo dei contagi giornalieri, la cui inclinazione diventa negativa dopo il picco, avrebbe maggior senso stimare il coefficiente γ su più finestre temporali (per esempio di lunghezza 8 giorni, pari alla durata media dell'infettività). Questo perchè, localmente, $\log(y_t)$ ha un andamento approssimativamente lineare nel tempo.

Se si ottiene una stima globale dei coefficienti di regressione, si può calcolare quella dei casi giornalieri attesi in questo modo:

$$\hat{\lambda}_t = \exp(\hat{\kappa} + \hat{\gamma} t) \tag{3.4}$$

Se invece, si calcolano delle stime locali γ_j, nella finestra temporale j-esima di lunghezza l osservazioni, si ha:

$$\hat{\lambda}_t = \hat{\lambda}_{t-l-1} \exp(\hat{\gamma}_j * t), \tag{3.5}$$

con $\hat{\lambda}_{t-l-1}$ ottenuta dalla finestra temporale precedente, mentre $\hat{\lambda}_0 = e^{\hat{k}_1}$, dove $\hat{k}_1$ è l'intercetta stimata nel primo periodo.

A questo punto, il tasso d'infettività δ può essere ottenuto come rapporto tra i nuovi casi stimati al tempo t e il totale di quelli degli 8 giorni precedenti, coerentemente con la durata media dell'infettività pari a 7.5 giorni. Quindi:

$$\hat{\delta} = \frac{\hat{\lambda}_t}{\sum_{i=1}^{8} \hat{\lambda}_{t-i}} \tag{3.6}$$

Di conseguenza, dalla formula (3.1) si può determinare $\hat{R}_0 = 7.5 * \hat{\delta}$.

Il modello esponenziale con parametro γ costante, nonostante sia parsimonioso, è un modello "statico" per cui non riesce ad adattarsi bene all'evoluzione del contagio, in quanto questa è caratterizzata da *shock* temporali. Quest'ultimi sono relativi a variazioni di breve periodo del numero dei contagiati, causate da nuovi focolai, cambiamenti nelle politiche di contenimento o nei criteri di somministrazione dei tamponi. Invece, il modello di regressione locale (3.5) descrive meglio le dinamiche di breve termine, ma è poco parsimonioso perchè necessita della stima di un parametro per ogni finestra temporale considerata.

Va evidenziato inoltre, che gli effetti delle misure restrittive sono persistenti a lungo nel tempo, perciò bisogna considerare anche la dipendenza di lungo periodo tra i conteggi dei nuovi casi. Questa dipendenza è dovuta anche al fatto che il tempo d'incubazione della malattia varia a seconda del soggetto e che non vi è uniformità spaziale e temporale nelle procedure

adottate per la misurazione dei contagi giornalieri.

Sulla base di tali motivazioni, è utile sviluppare un modello che, attraverso un numero limitato di parametri, possa apprendere in maniera dinamica dalle osservazioni passate, catturando la dipendenza di breve e lungo termine dagli *shock* temporali.

3.2 Modello autoregressivo Poisson log-lineare

Un modello con tali caratteristiche è il modello autoregressivo di Poisson log-lineare, utilizzato nell'ambito del contagio finanziario. Questo tipo di modello può essere adattato all'epidemia COVID-19, includendo una dinamica autoregressiva nel modello SIR, che gli permette di adattarsi alle variazioni temporali dei casi giornalieri, tenendo conto della dipendenza di breve e lungo termine. Si noti che questo è un modello *data-driven* che, in quanto autoregressivo "auto-impara" dai suoi stati passati; quindi si basa fortemente sui dati, piuttosto che sulla teoria epidemiologica sottostante il fenomeno. Cio è perfettamente in linea con lo scopo di questo lavoro, ovvero monitorare la diffusione dell'epidemia e valutare gli effetti delle eventuali politiche di contenimento adottate.

Dal punto di vista teorico, il processo è stato proposto da Fokianos e Tjøstheim (2011). Si tratta di un modello lineare generalizzato (GLM) di Poisson con funzione di legame logaritmica, nel quale viene introdotto un meccanismo di *feedback* autoregressivo. Con riferimento al COVID-19, assumendo che i nuovi casi giornalieri y_t abbiano, condizionatamente all'informazione al tempo $t-1$, distribuzione Poisson, la formulazione è la seguente:

$$y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t)$$
$$\log(\lambda_t) = \omega + \alpha \log(1 + y_{t-1}) + \beta \log(\lambda_{t-1}), \tag{3.7}$$

dove la serie storica $y_t \in \mathbb{N}$ è osservata e $\omega \in \mathbb{R}$, $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ sono parametri da stimare. Nel modello, ω è l'intercetta, mentre α e β esprimono la dipendenza del numero atteso di nuovi casi λ_t dalla storia del processo.

Si noti che l'inclusione di $\log(1 + y_{t-1})$ al posto di $\log(y_{t-1})$ permette di far fronte a conteggi uguali a zero. Fokianos e Tjøstheim (2011) hanno dimostrato che questa scelta non comporta differenze significative nei risultati.

La somma dei parametri α e β rappresenta la persistenza del processo, ovvero la misura di quanto l'impatto degli *shock* si protrae nel tempo. Più $\alpha + \beta$ è vicino a 1, maggiore è la persistenza degli *shock* e più lentamente la funzione di autocorrelazione globale tende ad annullarsi.

Nello specifico, α esprime la dipendenza di breve termine dall'osservazione dell'istante temporale precedente, mentre β la dipendenza di lungo termine da tutta la storia del processo. La presenza di β permette di tener conto degli effetti di memoria lunga nella serie dei contagiati, in maniera più parsimoniosa rispetto all'inclusione di numerosi ritardi di y_t nella forma $\log(1 + y_{t-i})$.[1]

Una volta ottenuta una stima dei parametri, si può calcolare quella dei casi giornalieri attesi in questo modo:

$$\hat{\lambda}_t = \exp\big(\hat{\omega} + \hat{\alpha}\log(1 + y_{t-1}) + \hat{\beta}\log(\hat{\lambda}_{t-1})\big)\,, \tag{3.8}$$
$$\text{con } \hat{\lambda}_1 = \exp(\hat{\omega} + \hat{\alpha}\log(1 + y_0) + \hat{\beta}\log(\lambda_0))\,, \quad y_0, \lambda_0 \text{ fissati.}$$

Da qui si può stimare il tasso di riproduzione istantaneo δ così come nella formula 3.6 e di conseguenza $\hat{R}_0 = 7.5 * \hat{\delta}$.

3.2.1 Proprietà del processo e simulazioni

Le condizioni di stazionarietà ed ergodicità del processo sono ambito di ricerca corrente. In un recente lavoro, (Sim, 2016, Proposizione 5.4.7) ha derivato le seguenti condizioni sufficienti:

$$\{|\beta| \vee |\alpha + \beta|\} < 1 \tag{3.9}$$

[1] Ciò corrisponde al passare da un processo ARCH ad uno GARCH nei modelli per la varianza condizionata

Sotto tali vincoli, seguendo Agosto et al. (2020), si può ricavare in maniera ricorsiva la seguente espressione:

$$\log(\lambda_t) = \omega \frac{1-\beta^t}{1-\beta} + \beta^t \log(\lambda_0) + \alpha \sum_{i=0}^{t-1} \beta^i \log(1 + y_{t-i-1}) \,, \tag{3.10}$$

la quale esprime il logaritmo dell'intensità in funzione di tutte le osservazioni passate, fissati y_0 e λ_0.

Si noti che l'ipotesi che la distribuzione di riferimento sia, condizionamente, Poisson implica l'assunzione $\mathbb{E}(y_t|\mathcal{F}_{t-1}) = \mathrm{Var}(y_t|\mathcal{F}_{t-1}) = \lambda_t$; invece la distribuzione marginale di y_t non è Poisson. Per quanto riguarda la media non condizionata della trasformata della serie $z_t = \log(1 + y_t)$, definendo $\nu_t = \log(\lambda_t)$, vale l'approssimazione $\mathbb{E}(z_t) \approx \mathbb{E}(\nu_t) \approx \mu$, dove $\mu = \frac{\omega}{1-\alpha-\beta}$

Risulta, invece, difficile ricavare un'espressione esplicita per la funzione di autocorrelazione del processo perchè la specificazione log-lineare non permette di trovare una relazione ricorsiva per la covarianza.

Pertanto, per comprendere meglio le caratteristiche del processo, si ricorre alla simulazione. Sulla base della funzione (3.10), fissati i valori $y_0 = 10$ e $\lambda_0 = 10$, si simulano cinque serie per altrettante combinazioni diverse dei parametri α e β, con $\omega = 0.1$. Vengono generate serie da 1000 osservazioni, delle quali le prime 800 vengono scartate, per assicurarsi che il processo abbia raggiunto la regione di stazionarietà. I grafici delle serie simulate e quelli delle rispettive funzioni di autocorrelazione globale sono rappresentati in figura 3.1. Dall'alto verso il basso, si considerano le seguenti configurazioni:

1. $\alpha = 0.83, \beta = 0.15$: $\alpha+\beta$ prossimo a 1 quindi vi è elevata persistenza, siccome $\alpha \gg \beta$ prevale la dipendenza di breve termine;

2. $\alpha = 0.15, \beta = 0.83$: elevata persistenza, prevale la dipendenza di lungo termine;

3. $\alpha = 0.5, \beta = 0.1$: scarsa persistenza, α prevale su β;

4. $\alpha = -0.13, \beta = -0.85$: $\alpha + \beta$ prossimo a -1, $|\beta|$ prevale su $|\alpha|$;

5. $\alpha = 1.65, \beta = -0.67$: $\alpha + \beta$ prossimo a 1, $\alpha > 1$ e $\beta < 0$.

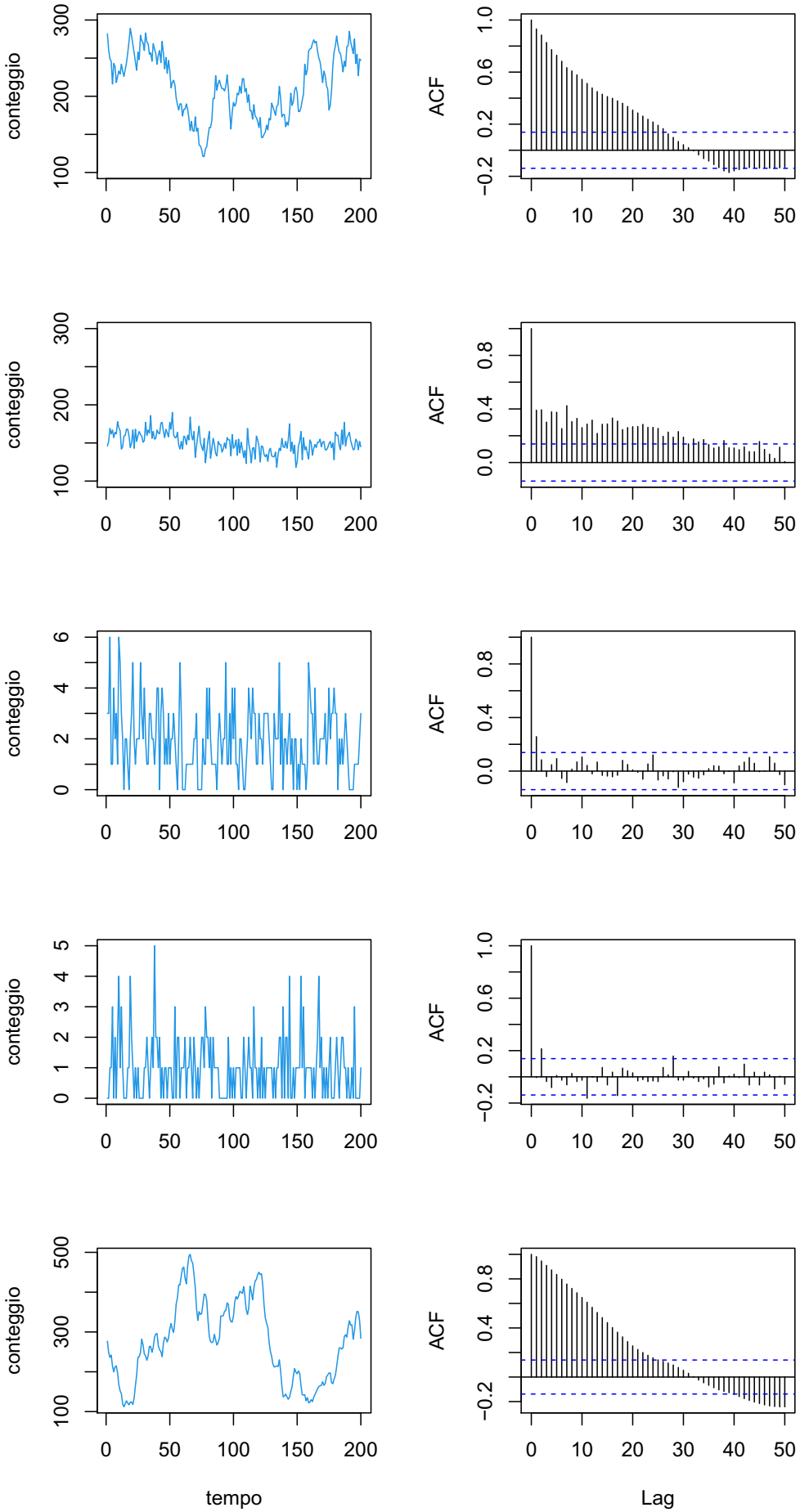

Figura 3.1: Dall'alto verso il basso, serie simulate e ACF per le combinazioni: (1) $\alpha = 0.83$, $\beta = 0.15$; (2) $\alpha = 0.15$, $\beta = 0.83$; (3) $\alpha = 0.5$, $\beta = 0.1$; (4) $\alpha = -0.13$, $\beta = -0.85$; (5) $\alpha = 1.65$, $\beta = -0.67$

Come già evidenziato in precedenza, si nota dalle figure 3.1(1), 3.1(2) e 3.1(5) che quando la somma di α e β è vicina a 1, la funzione di autocorrelazione globale tende ad annullarsi lentamente. Restando in questo caso, se α prevale su β, si osserva dal grafico in 3.1(1) che la serie simulata ha un'alta variabilità, poichè il suo comportamento è fortemente influenzato dalle variazioni temporali di breve periodo. Diversamente, nel grafico in 3.1(2) (sulla stessa scala di quello in 3.1(1)), si vede che, se $\alpha + \beta \approx 1$ e $\beta \gg \alpha$, la serie è caratterizzata da un'elevata persistenza e manifesta la dipendenza da tutta la sua storia passata.

Se invece $\alpha + \beta$ non è prossimo a 1 e quindi vi è scarsa persistenza, l'ACF diventa subito non significativa, mentre la serie simulata risulta molto "rumorosa", come si evince dalle figure 3.1(3) e 3.1(4). Nel caso (4) si nota che la serie simulata presenta molte osservazioni uguali a zero. Ciò è dovuto al fatto che entrambi i parametri di riferimento sono negativi e quindi, la media condizionata λ_t della distribuzione Poisson da cui si simula, si attesta sempre su valori prossimi a zero.

Infine, nell'ultimo caso ci si rifà alle considerazioni del caso (2): il processo è dominato dal parametro α, quindi la serie simulata è guidata dagli *shock* di breve periodo, come si osserva dalla figura 3.1(5). Si noti anche che da un certo lag in poi la serie presenta autocorrelazione negativa.

3.2.2 Generalizzazione e processo INGARCH

Il modello appena presentato nel dettaglio è quello che si prende a riferimento per quanto riguarda questo lavoro. Tuttavia, si è interessati a caratterizzare, nella loro formulazione più generale, questi tipi di modello.

A tal fine, si assume che la serie storica di conteggi $Y_t \in \mathbb{N}$ abbia media condizionata $\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = \lambda_t$. Si noti che non si fa alcun tipo di assunzione nè sulla distribuzione nè sulla varianza condizionata. Quindi, denotando con $\boldsymbol{X_t}$ il vettore delle covariate r-dimensionale $\boldsymbol{X_t} = (X_{t,1}, \ldots, X_{t,r})^T$, si può scrivere la formulazione generale:

$$g(\lambda_t) = \omega + \sum_{k=1}^{p} \alpha_k \, \tilde{g}(Y_{t-k}) + \sum_{l=1}^{q} \beta_\ell \, g(\lambda_{t-\ell}) + \boldsymbol{\eta}^T \boldsymbol{X_t} \,, \tag{3.11}$$

dove $g\colon \mathbb{R}^+ \to \mathbb{R}$ è la funzione di legame, $\tilde{g}\colon \mathbb{N}_0 \to \mathbb{R}$ è una funzione di trasformazione e $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)^T$ è il vettore di parametri delle covariate.

Inoltre, in molti casi è di interesse inserire nel modello le cosiddette *covariate di intervento*, relative a eventi straordinari, quali per esempio cambi radicali delle politiche di contenimento in un contesto epidemico. Seguendo Liboschik et al. (2017), le covariate di intervento sono della forma $\delta^{t-\tau}\,\mathbb{1}(t \geq \tau)$, dove τ rappresenta il tempo in cui si verifica l' *intervento* e il tasso di decadimento δ è una costante nota.

Ritornando al caso (3.11) senza variabili di intervento, si ipotizzi $Y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t)$ e quindi $\mathrm{Var}(Y_t|\mathcal{F}_{t-1}) = \lambda_t$. Si vede subito che, assumendo $g(x) = \log(x)$, $\tilde{g}(x) = \log(x+1)$, $p = q = 1$ e $\boldsymbol{\eta} = \mathbf{0}$, ci si riconduce al modello PAR log-lineare descritto in precedenza.

Un altro modello specifico di particolare interesse è il cosiddetto *integer-valued GARCH model*, abbreviato come INGARCH. Continuando a riferirsi ad una distribuzione condizionatamente Poisson, assumendo $g(x) = \tilde{g}(x) = x$, $p = q = 1$ e $\boldsymbol{\eta} = \mathbf{0}$, si può caratterizzare il processo INGARCH(1,1):

$$\lambda_t = \omega + \alpha\, Y_{t-1} + \beta\, \lambda_{t-1}\,, \tag{3.12}$$

dove $\omega > 0$, $\alpha \geq 0$ e $\beta \geq 0$ sono vincoli sui parametri per garantire la positività di λ_t. Si osservi che, per la stessa ragione, qualora si decida di includere delle covariate, esse e i corrispondenti coefficienti dovrebbero essere non-negativi. Ferland et al. (2006) hanno dimostrato che il processo è stazionario se $0 < \alpha + \beta < 1$ e, in questo caso, si può arrivare ad una rappresentazione analitica della media non condizionata $\mathbb{E}(Y_t) = \mathbb{E}(\lambda_t) = \mu = \frac{\omega}{1-\alpha-\beta}$ e della varianza non condizionata $\mathrm{Var}(Y_t) = \mu\Big(1 + \frac{\alpha^2}{1-(\alpha+\beta)^2}\Big)$

Si vede che $\mathrm{Var}(Y_t) > \mathbb{E}(Y_t)$ se $\alpha \neq 0$, quindi il modello tiene conto della sovradispersione, fenomeno tipico delle serie storiche di conteggio. Inoltre, gli autori hanno dimostrato che $\mathrm{Cov}[Y_t, Y_{t-h}] > 0$ se $0 < \alpha + \beta < 1$ e quindi il processo non può essere utilizzato per modellare correlazioni negative.

In questa tesi, si propende per l'utilizzo del modello di Poisson autoregressivo (3.7), principalmente perchè la specificazione log-lineare accomoda la natura sub-esponenziale del processo epidemico. Poi dal punto di vista teo-

rico, questo modello fornisce una struttura più generale, la quale, al contrario di (3.12), permette di modellare correlazioni negative e di includere covariate in maniera più diretta. Nel caso specifico dell'applicazione al COVID-19, si sceglie di includere solo i primi lag di Y_t e λ_t e di non inserire covariate, al fine di ottenere un modello più semplice possibile, in particolare dal punto di vista interpretativo.

3.3 Inferenza

Dal punto di vista inferenziale, il modello viene stimato attraverso il metodo della quasi-massima verosimiglianza (QMLE) condizionata. In questa sezione ci si riferisce al modello PAR log-lineare (3.7), ma il procedimento può essere facilmente generalizzato.

Rifacendosi alla distribuzione $y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t)$, si può scrivere la funzione di probabilità condizionata:

$$P(Y_t = y_t|\mathcal{F}_{t-1}) = \frac{\lambda_t^{y_t} \exp(-\lambda_t)}{y_t!}$$

Si assuma che $\theta = (\omega, \alpha, \beta)^T$ sia il vettore dei parametri da stimare e che $\theta_0 = (\omega_0, \alpha_0, \beta_0)^T$ rappresenti i veri valori dei parametri. Dato il valore iniziale $\lambda_0 = \exp(\nu_0)$, la funzione di verosimiglianza condizionata per le osservazioni $(y_1, \ldots, y_n)$ è:

$$L(\theta) = \prod_{t=1}^{n} \frac{\lambda_t^{y_t}(\theta) \exp(-\lambda_t(\theta))}{y_t!}$$

Definendo $\nu_t = \log(\lambda_t)$, il modello (3.7) si può riscrivere nella seguente forma:
$\nu_t = \omega + \alpha \log(1 + y_{t-1}) + \beta\nu_{t-1}$
Quindi la log-verosimiglianza è data, al netto di una costante additiva, da:

$$l(\theta) \propto \sum_{t=1}^{n} \Big(y_t\, \nu_t(\theta) - \exp(\nu_t(\theta)) \Big)$$

Definito come $l_t(\theta)$ lo score di una singola osservazione, si può ottenere la

funzione punteggio:

$$S_n(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \sum_{t=1}^{n} \frac{\partial l_t(\theta)}{\partial \theta} = \sum_{t=1}^{n} \Big(y_t - \exp(\nu_t(\theta)) \Big) \frac{\partial \nu_t(\theta)}{\partial \theta} \,,$$

dove $\frac{\partial \nu_t(\theta)}{\partial \theta}$ è il vettore tridimensionale delle derivate parziali, che può essere calcolato dalle seguenti ricorsioni:

$$\begin{cases} \frac{\partial \nu_t}{\partial \omega} = 1 + \beta \frac{\partial \nu_{t-1}}{\partial \omega} \\ \frac{\partial \nu_t}{\partial \alpha} = \log(1 + y_{t-1}) + \beta \frac{\partial \nu_{t-1}}{\partial \alpha} \\ \frac{\partial \nu_t}{\partial \beta} = \nu_{t-1} + \beta \frac{\partial \nu_{t-1}}{\partial \beta} \end{cases}$$

Infine, la matrice Hessiana condizionata è data dalla seguente formula (3.13):

$$\begin{aligned} H_n(\theta) &= -\frac{\partial^2 l(\theta)}{\partial \theta \, \partial \theta^T} = -\sum_{t=1}^{n} \frac{\partial^2 l_t(\theta)}{\partial \theta \, \partial \theta^T} \\ &= \sum_{t=1}^{n} \exp(\nu_t(\theta)) \Big(\frac{\partial \nu_t(\theta)}{\partial \theta} \Big) \Big(\frac{\partial \nu_t(\theta)}{\partial \theta} \Big)^T - \sum_{t=1}^{n} \big(y_t - \exp(\nu_t) \big) \frac{\partial^2 \nu_t(\theta)}{\partial \theta \, \partial \theta^T} \end{aligned}$$

Lo stimatore di quasi-massima verosimiglianza $\hat{\theta}_n$ di θ, qualora esista, è soluzione del seguente problema di ottimizzazione non lineare:

$$\hat{\theta}_n := \underset{\theta \in \Theta}{\operatorname{argmax}} \, l(\theta) \,,$$

il quale va risolto tramite un metodo di approssimazione numerica.

L'inferenza sui parametri del modello è basata sulla normalità asintotica della stima di quasi-massima verosimiglianza. Difatti, sotto le condizioni $|\beta_0| < 1 \vee |\alpha_0 + \beta_0| < 1$ che determinano la stazionarietà e l'ergodicità del processo, esiste unico lo stimatore $\hat{\theta}_n$, il quale è consistente ed asintoticamente normale:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N_3(0, \Sigma) \,,$$

dove θ_0 è il vero valore del parametro e $\Sigma/n = H(\theta)^{-1}G(\theta)H(\theta)^{-1}$ è la cosiddetta matrice di covarianza *sandwich*. $H(\theta)$ può essere stimata in maniera consistente da $H_n(\hat{\theta})$ così come in (3.13). Invece, uno stimatore consistente di $G(\theta)$ è dato da $G_n(\hat{\theta})$, dove

$$\begin{aligned} G_n(\theta) = \mathrm{Var}(S_n(\theta)) &= \sum_{t=1}^{n} \mathrm{Var}\Big(\frac{\partial l_t(\theta)}{\partial \theta} | \mathcal{F}_{t-1}\Big) \\ &= \sum_{t=1}^{n} \exp(\nu_t(\theta)) \Big(\frac{\partial \nu_t(\theta)}{\partial \theta}\Big) \Big(\frac{\partial \nu_t(\theta)}{\partial \theta}\Big)^T \end{aligned}$$

Se l'assunzione della distribuzione condizionatamente Poisson è valida, si ottiene uno stimatore di massima verosimiglianza ordinario. Infatti, si verifica facilmente che $H(\theta) = G(\theta)$, per cui

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N_3(0, G(\theta)^{-1})$$

Se invece il modello risulta non correttamente specificato, si può ottenere uno stimatore di quasi-massima verosimiglianza. In questo caso, $H(\theta) \neq G(\theta)$, per cui la varianza di $\hat{\theta}_n$ è definita dalla matrice di covarianza *sandwich*.

3.4 Procedura di stima

Per stimare il modello utilizzo il software statistico R versione 4.0.2. In particolare, impiego la funzione *tsglm* del pacchetto *tscount*, implementato appositamente per questo tipo di processi. Nello specifico, (attraverso l'opzione *init.method = "firstobs")*, ν_t viene inizializzato da $log(1 + y_1)$, dove y_1 è la prima osservazione, e le derivate parziali di ν_t vengono inizializzate con valore zero. I parametri vengono stimati senza nessun vincolo (opzione *final.control = list(constrained = NULL))*. Bisogna esaminare con attenzione le stime nei casi in cui non rispettano le condizioni $\{|\beta| \vee |\alpha + \beta|\} < 1$, perchè ciò significa che il processo diverge e quindi le stime sono poco affidabili.

Capitolo 4

Applicazione e risultati

In questo Capitolo presento un'applicazione empirica del modello, il quale può essere adattato in maniera specifica per ciascun Paese o regione.
Il modello si adatta dinamicamente alla curva dei nuovi casi giornalieri, tramite la stima di tre parametri che variano nel tempo. I coefficienti α e β costituiscono il fulcro del modello, poichè forniscono importanti elementi per quanto riguarda l'interpretazione dell'andamento dell'epidemia.

In un primo passo, si stima il modello sul campione completo a disposizione, arrivando a delle stime dei coefficienti che sono le più aggiornate possibili. Infatti, riferendosi ai dati all'ultimo istante temporale disponibile, queste stime fotografano la situazione del momento, mostrando in quale fase si trova l'evoluzione dell'epidemia.

Successivamente, in un secondo passo, si stima il modello in maniera dinamica, utilizzando delle *expanding windows*. Nello specifico, seguendo Agosto et al. (2020), si stima prima il modello sulle 15 osservazioni iniziali, poi sulle prime 16 e così via, aggiungendo ogni volta un'osservazione, fino ad arrivare alla stima sul campione completo.[1] In questo modo, si può analizzare l'evoluzione temporale delle stime dei coefficienti e cogliere così i cambiamenti strutturali del processo epidemico. Se le stime sul campione completo rappresentano una fotografia dell'epidemia, le stime tramite *expanding windows* ne sono il film e permettono quindi di comprendere la dinamica del contagio.

[1] Un numero di osservazioni pari a 15 è la condizione minima per ottenere delle stime dei parametri attendibili

Infine, in un terzo passo, vengono applicati i risultati ottenuti nel contesto del modello SIR, stimando il numero di riproduzione di base R_0 così come illustrato nel capitolo precedente. Ciò permette di interpretare meglio il modello proposto e di trarre ulteriori conclusioni.

Le analisi illustrate in questi tre passi vengono svolte prima per quanto riguarda i dati nazionali e in seguito per alcune regioni italiane d'interesse.

4.1 Dati nazionali

In primis, si considerano i dati sui contagi giornalieri italiani, comunicati dalla Protezione Civile, che vanno fino al 31 luglio. Per comprendere al meglio l'applicazione del modello, si confrontano i risultati ottenuti per l'Italia con quelli della Florida, Stato degli USA che a fine luglio si trova in una fase del contagio completamente diversa da quella italiana. I dati relativi alla Florida provengono dal *Florida Department of Health*: si fa riferimento al lasso temporale che va dal 4 marzo, data dei primi contagi, al 31 luglio.

Dal grafico in figura 4.1 si osserva che, a fine luglio, i contagi giornalieri in Florida sono in una fase di *plateau* e non hanno ancora iniziato a descrescere in maniera significativa. Si noti anche che i casi giornalieri in Florida sono su una scala maggiore rispetto all'Italia, nonostante la Florida abbia circa 21 milioni di abitanti e l'Italia ne abbia circa 60 milioni. Va comunque tenuto presente che i criteri di misurazione dei contagi sono diversi nei due Paesi e quindi i confronti vanno effettuati con grande cautela.

4.1.1 Stima sul campione completo

Le tabella 4.1 riporta, rispettivamente per Italia e Florida, i risultati della stima del modello sull'intero campione di osservazioni.

Si nota che nel caso dell'Italia il parametro α prevale su β, indicando che i nuovi casi di *oggi* sono più legati a quelli osservati *ieri*, piuttosto che alla storia del processo. Difatti, il parametro β, che esprime la dipendenza di lungo termine, non è statisticamente significativo. Al contrario, in Florida β predomina su α, il che denota che c'è un trend ancora in corso.

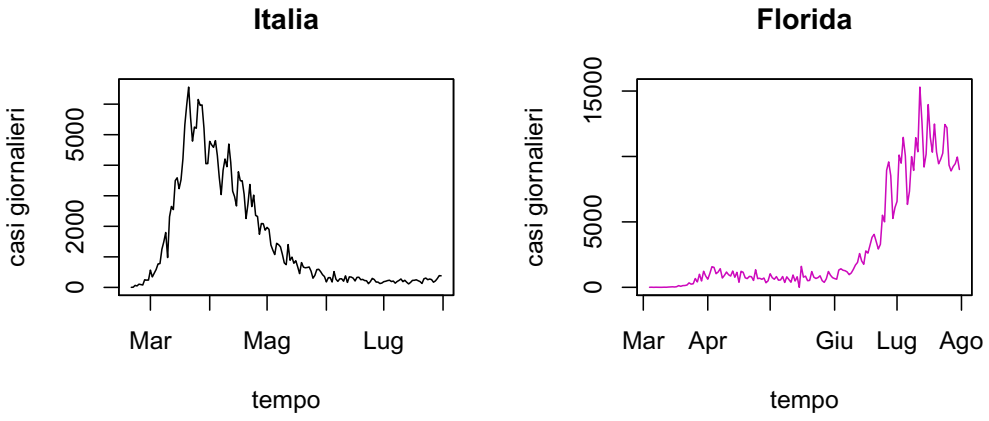

Figura 4.1: Nuovi contagi giornalieri osservati in Italia e in Florida

Sempre per la Florida, si nota che il coefficiente α non è significativo; mentre, per entrambi gli Stati, il parametro ω risulta non significativo.

Seguendo Agosto et al. (2020), si può fornire un'interpretazione dei parametri del modello. Per l'Italia, la stima del coefficiente α mostra che se i casi attesi di *ieri* λ_{t-1} sono prossimi a zero, 100 casi osservati *ieri* portano a circa 40 nuovi casi attesi per *oggi*. Invece dalla stima di β si può constatare che se ci si attende 100 casi per *ieri*, ma si osservano zero casi, allora i nuovi casi attesi per *oggi* sono circa 2.

Il parametro ω rappresenta invece l'intercetta del modello. In generale, un valore elevato di ω indica che i dati manifestano un'alta variabilità che non può essere spiegata dal modello e che la serie dei nuovi casi giornalieri presenta scarsa dipendenza temporale.

Tabella 4.1: Risultati della stima del modello per l'Italia e la Florida

	Italia			Florida		
Parametro	Stima	S.E.	(p-value)	Stima	S.E.	(p-value)
ω	0.224	0.328	(0.495)	0.287	0.300	(0.338)
α	0.787	0.269	(0.003)	0.307	0.258	(0.234)
β	0.186	0.271	(0.493)	0.663	0.276	(0.016)

4.1.2 Stima dinamica

In seconda battuta, ci si focalizza sull'evoluzione temporale delle stime ottenute. La figura 4.2 mostra la dinamica delle stime dei cofficienti α e β, ottenute a partire dal campione dei primi 15 dati, aggiungendo via via un'osservazione.

Per l'Italia, si osserva che inizialmente β è maggiore di uno e quindi, in una prima fase, la crescita dei nuovi casi giornalieri è esplosiva. Se all'inizio β prevale su α, in seguito il divario tra le stime dei due coefficienti tende a ridursi, come conseguenza delle misure di contenimento sempre più restrittive che sono state imposte dal Governo. Nel grafico in alto, la linea rossa verticale rappresenta l'11 marzo, data in cui il Governo ha emanato il decreto che estendeva il *lockdown* a tutto il territorio italiano. Dopo tre settimane da questa data, si osserva che i due coefficienti si invertono e che α diventa maggiore di β, indicando il superamento del picco dell'epidemia. Con il passare del tempo, si vede che α continua a dominare su β, a dimostrazione che i nuovi casi giornalieri sono guidati dalla dipendenza di breve termine.

Invece, per quanto riguarda la Florida, si vede che, nonostante il *gap* si sia ridotto, β continua a prevalere su α nel tempo. Questo suggerisce che il picco del contagio non è stato ancora superato e quindi il trend persiste, dato che la componente di lungo termine prevale su quella di breve termine.

In generale, questa procedura di stima dinamica può essere funzionale ai *policymaker* per identificare le tendenze di fondo del contagio. Infatti, osservando l'evoluzione delle stime di α e β nel tempo, si possono individuare eventuali effetti delle politiche di contenimento.

Per l'Italia, risulta evidente che, nell'ultima fase, le variazioni dei coefficienti sono impercettibili, poichè le stime si basano su dati relativi ad un periodo molto ampio. Pertanto, per cogliere meglio cambiamenti strutturali del processo epidemico, dovuti per esempio allo sviluppo di nuovi focolai, si può considerare la stima dinamica su finestre temporali più brevi, ad esempio tramite il metodo *rolling window*.

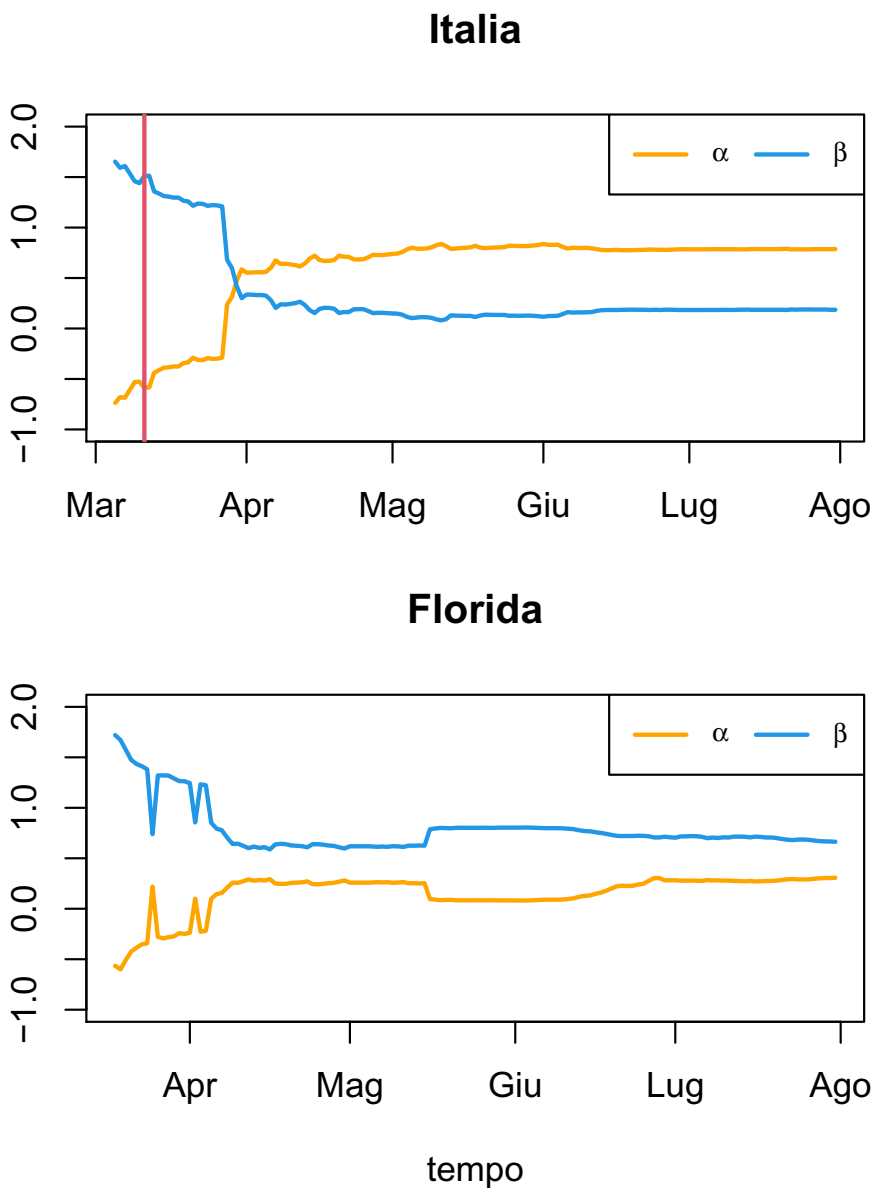

Figura 4.2: Evoluzione temporale dei parametri α e β per Italia e Florida.

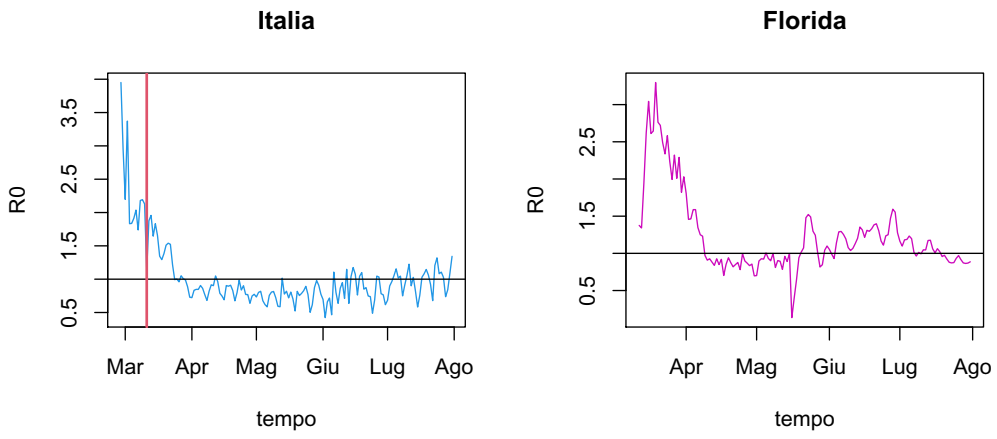

Figura 4.3: Dinamica temporale di R_0 per Italia e Florida

4.1.3 Stima del numero di riproduzione

Al terzo passo, si introducono le stime ottenute all'interno del modello SIR, al fine di stimare il numero di riproduzione di base nel tempo. Nella figura 4.3 si rappresenta l'evoluzione temporale di R_0 per Italia e Florida, dove nel grafico a sinistra la linea rossa indica sempre la data dell'11 marzo. Per l'Italia, si può vedere che le misure restrittive hanno permesso di ridurre R_0, che in una fase iniziale si attestava ad un valore attorno a 3. Difatti, nei mesi successivi questo parametro si è stabilizzato su valori sotto la soglia 1, salvo poi riaumentare nell'ultimo periodo a causa di alcuni focolai.

Per la Florida, si osserva che R_0 è stato a lungo sensibilmente superiore a 1 o comunque su valori attorno a tale soglia, denotando una situazione del contagio che viene definita come critica.

Nel complesso, il monitoraggio dell'andamento delle stime di R_0 nel tempo costituisce un importante strumento al servizio dei *policymaker*, che consente di decidere se introdurre misure restrittive o rilassare quelle già esistenti. La stima di R_0 tramite il modello autoregressivo permette di considerare il fatto che il tempo d'incubazione ha una media di 7.5 giorni, ma può variare da 1 a 21 giorni a seconda dell'individuo.

4.2 Dati regionali

In aggiunta, si eseguono le analisi illustrate nei tre passi precedenti su base regionale. Si considerano sempre i dati della Protezione Civile, relativi ai contagi in alcune regioni d'interesse, segnatamente Veneto, Sicilia, Lombardia, Piemonte, Emilia-Romagna e Lazio. Un'analisi più circostanziata a livello territoriale permette infatti di monitorare meglio l'eventuale sviluppo di nuovi *cluster* d'infezioni.

Dalla tabella 4.2 si denota che in Veneto, Lombardia e Piemonte sono significative sia la componente di breve che quella di lungo periodo; mentre per Sicilia, Emilia-Romagna e Lazio prevale la dipendenza di breve termine nel numero di nuovi contagi.

Tabella 4.2: Risultati della stima del modello per le regioni italiane

	Veneto		Sicilia	
Parametro	Stima	S.E. (p-value)	Stima	S.E. (p-value)
ω	0.219	0.135 (0.104)	0.067	0.190 (0.726)
α	0.558	0.142 (0.000)	0.601	0.257 (0.019)
β	0.405	0.149 (0.007)	0.381	0.247 (0.122)
	Lombardia		Piemonte	
Parametro	Stima	S.E. (p-value)	Stima	S.E. (p-value)
ω	0.214	0.163 (0.190)	0.270	0.257 (0.294)
α	0.476	0.118 (0.000)	0.502	0.240 (0.037)
β	0.494	0.125 (0.000)	0.455	0.254 (0.073)
	Emilia-Romagna		Lazio	
Parametro	Stima	S.E. (p-value)	Stima	S.E. (p-value)
ω	0.122	0.146 (0.406)	0.099	0.113 (0.378)
α	0.761	0.151 (0.000)	0.696	0.117 (0.000)
β	0.218	0.151 (0.149)	0.281	0.117 (0.016)

Esaminando invece le dinamiche delle stime in figura 4.4, si osserva che c'è eterogeneità tra gli andamenti temporali dei coefficienti nelle varie regioni. L'esempio eclatante è il Lazio, dove α prevale su β in maniera stabile per tutta la durata dell'epidemia: ciò significa che il processo dei casi giornalieri

presenta scarsa memoria e quindi il numero di nuovi contagi è dettato principalmente dalle contingenze giornaliere degli stessi. Invece, le altre regioni presentano un'evoluzione dei coefficienti un po' più simile a quella nazionale. Nell'esempio del Piemonte, il divario tra i due parametri rimane alto per più tempo rispetto agli altri casi, indicando un picco della curva dei contagi giornalieri ritardato in confronto alle altre regioni.

Invece, il grafico della Sicilia rivela che nell'ultimo periodo β si sta avvicinando ad α, probabilmente a causa dell'innalzamento della curva dei contagi, dovuto a casi provenienti dall'estero. Sempre per la Sicilia, le stime del numero di riproduzione di base in figura 4.5, nell'ultimo periodo evidenziano l'innalzamento di R_0 a valori ben al di sopra di 1, a causa dei motivi sopracitati. Anche in Veneto si è verificato un aumento di R_0, a causa del numero di contagi massivo avvenuto tra migranti, in una caserma di Treviso.

Diversamente, nelle altre regioni, il valore di R_0 si mantiene intorno ad 1. Va notato che le stime di R_0 risultano molto variabili poichè i dati regionali utilizzati sono più "rumorosi" di quelli nazionali.

4.3 Estensioni e sviluppi futuri

Il modello presentato può essere facilmente applicato ad altri Stati e regioni. La struttura generale dello stesso permette di inserire covariate spaziali, temporali o di altro tipo. Inoltre, si può considerare l'introduzione nel modello di differenti ritardi di y_t e λ_t, sebbene la struttura presa come riferimento in questo lavoro consenta di tenere conto sia della non stazionarietà in media e varianza che anche di componenti stagionali.

Il modello può anche essere applicato ai dati sui ricoveri in terapia intensiva e sui nuovi decessi giornalieri. A tal proposito, Farcomeni et al. (2020) utilizzano un aproccio che combina un *modello lineare generalizzato con effetti misti* (GLMM) di Poisson e un modello PAR, per prevedere i posti letto occupati da pazienti COVID-19 nelle terapie intensive delle regioni italiane.

Per concludere, il modello presentato in questa tesi può anche essere utilizzato per misurare l'impatto del COVID-19 sull'economia e sui mercati finanziari.

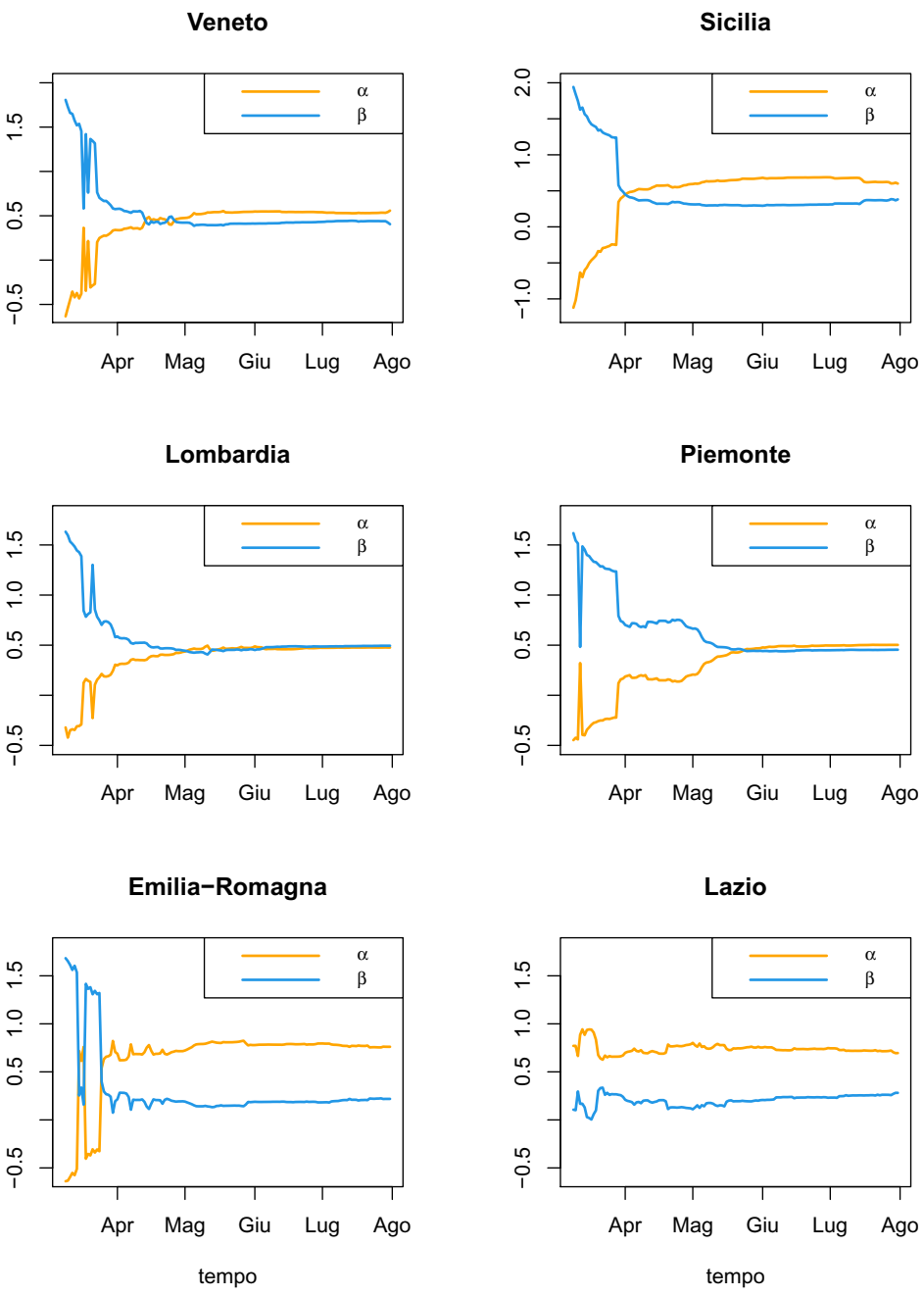

Figura 4.4: Evoluzione temporali dei parametri α e β a livello regionale

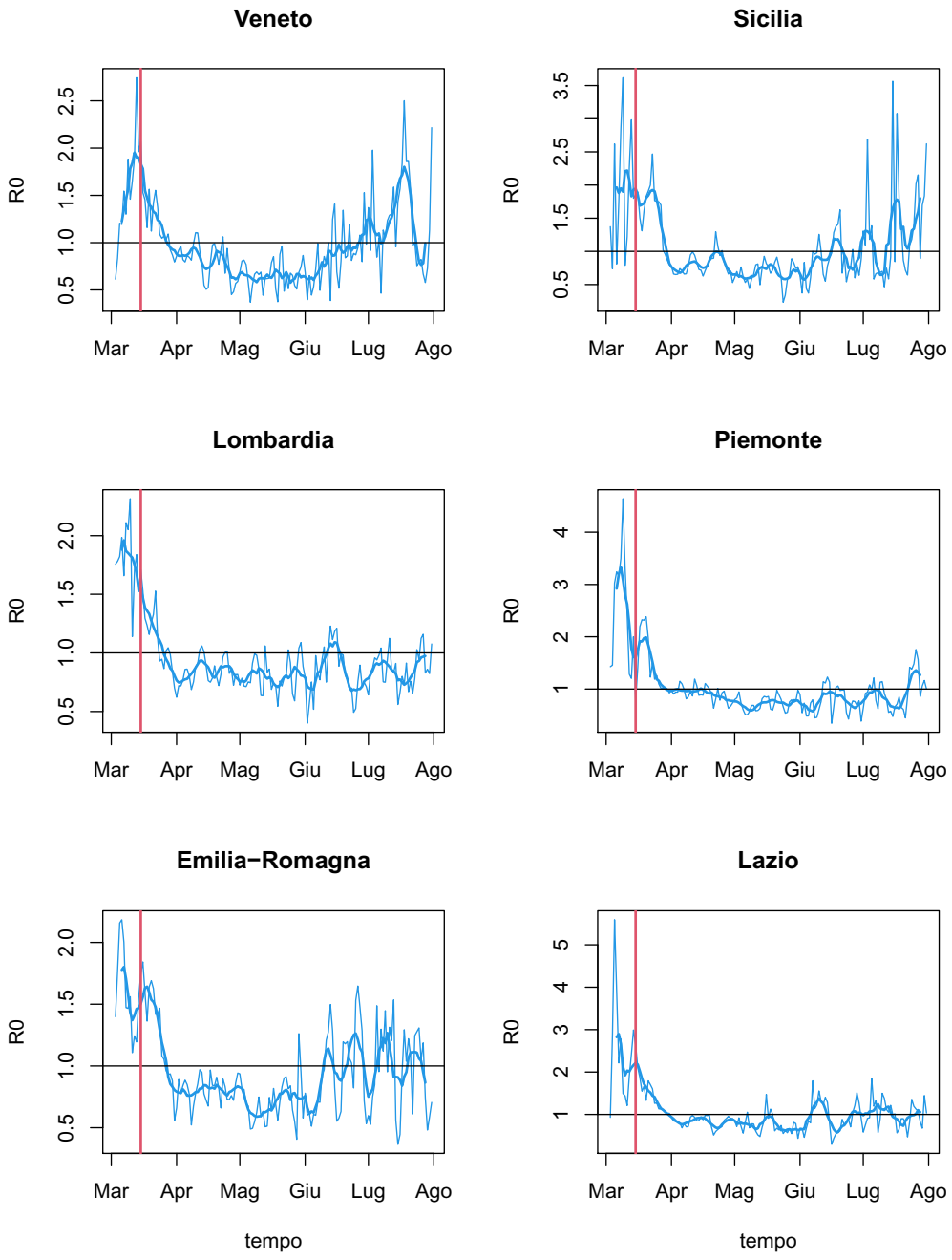

Figura 4.5: Dinamica temporale di R_0 a livello regionale. Evoluzione delle stime (linee sottili) e medie mobili a 7 giorni (linee spesse), la linea rossa rappresenta la data dell'11 marzo

Capitolo 5

Conclusioni

In questo lavoro si è presentato un modello dinamico per i dati sui contagi COVID-19 in Italia, che può essere aggiornato quotidianamente, allo scopo di permettere al Governo di monitorare, in tempo reale, l'evoluzione dell'epidemia nella fase *post-lockdown*.

In primo luogo, si sono puntualizzati i limiti che presentano i dati ufficiali italiani. In particolare, i contagiati sono una misurazione imprecisa, poichè la rilevazione di essi non avviene tramite un disegno di campionamento specifico e i criteri di somministrazione dei tamponi non sono uniformi nello spazio e nel tempo. Ciò nonostante, il modello illustrato, attraverso la sua dinamica autoregressiva, riesce a tenere conto di parte di questi errori di misurazione, considerando eventuali cambiamenti nelle strategie di rilevazione mediante tampone. Nello specifico, si è mostrato come questo modello permetta di tenere in considerazione la dipendenza di breve e lungo termine dei casi giornalieri, in modo da riuscire a cogliere sia lo sviluppo di nuovi focolai che cambiamenti nelle politiche di contenimento.

In seguito, si è applicato il modello per ottenere una stima dinamica dei parametri e di R_0. Tali stime testimoniano l'efficacia del *lockdown* nel ridurre il numero dei contagi. Pertanto la metodologia adottata, magari abbinata all'utilizzo di altri modelli, può fornire importanti indicazioni riguardo gli effetti delle misure restrittive introdotte. Quest'ultime, possono riguardare per esempio la chiusura delle discoteche oppure l'obbligo di indossare le

mascherine in determinati orari.

D'altra parte, siccome nel periodo estivo l'età media dei casi rilevati si è abbassata, i dati sui contagi possono offrire maggiori informazioni rispetto a quelli relativi a ricoveri in terapia intensiva e decessi. Pertanto, si può pensare che un'applicazione del modello a livello regionale, possa anticipare lo sviluppo di una seconda ondata epidemica, suggerendo eventualmente l'attuazione di politiche di contenimento specifiche per ciascuna regione.

La situazione della Florida, così come quella di altri Paesi, dimostra che il COVID-19 resta un'emergenza a livello globale e quindi occorre mantenere alta la guardia per non vanificare tutti i sacrifici economici e sociali fatti in questi mesi.

Appendice A

Codice R

```
library(astsa)
library(dplyr)
library(zoo)
library(tscount)
library(xtable)

# caricamento dati Protezione Civile

dati_ITA <- read.csv("https://raw.githubusercontent.com/
   pcm-dpc/COVID-19/master/dati-andamento-nazionale/dpc-
   covid19-ita-andamento-nazionale.csv")
dati_regioni <- read.csv("https://raw.githubusercontent.
   com/pcm-dpc/COVID-19/master/dati-regioni/dpc-covid19-
   ita-regioni.csv")

# creazione dei data frame che comprendono le serie
   storiche d'interesse e le corrispondenti variazioni
   giornaliere, sia a livello nazionale che regionale

totali_ITA <- data.frame(data = as.Date(dati_ITA$data),
    casi = dati_ITA$totale_casi,
    deceduti = dati_ITA$deceduti,
```

```
    dimessi_guariti = dati_ITA$dimessi_guariti,
    attualmente_positivi = dati_ITA$totale_positivi,
    terapia_intensiva = dati_ITA$terapia_intensiva,
    soggetti_testati = dati_ITA$casi_testati
    )

variazioni_ITA <- data.frame(
    data = as.Date(dati_ITA$data),
    casi = dati_ITA$nuovi_positivi,
    deceduti = diff(c(0, dati_ITA$deceduti)),
    dimessi_guariti = diff(c(0, dati_ITA$dimessi_guariti)
       ),
    attualmente_positivi = dati_ITA$variazione_totale_
       positivi,
    terapia_intensiva = diff(c(0, dati_ITA$terapia_
       intensiva)),
    soggetti_testati = diff(c(0, dati_ITA$casi_testati))
    )

totali_regioni <- data.frame(
    data = as.Date(dati_regioni$data),
    regione = dati_regioni$denominazione_regione,
    casi = dati_regioni$totale_casi,
    deceduti = dati_regioni$deceduti,
    dimessi_guariti = dati_regioni$dimessi_guariti,
    attualmente_positivi = dati_regioni$totale_positivi,
    terapia_intensiva = dati_regioni$terapia_intensiva
    )

variazioni_regioni <- totali_regioni %>%
group_by(regione) %>%
mutate( deceduti = diff(c(0, deceduti)),
    dimessi_guariti = diff(c(0,dimessi_guariti)),
    terapia_intensiva = diff(c(0, terapia_intensiva))
    )
```

```
variazioni_regioni <- as.data.frame(variazioni_regioni)
variazioni_regioni$casi <- dati_regioni$nuovi_positivi
variazioni_regioni$attualmente_positivi = dati_regioni$
   variazione_totale_positivi
variazioni_regioni <- variazioni_regioni %>% arrange(
   regione)

n <- nrow(variazioni_ITA)

# analisi descrittive

# figura 2.1
op <- par(no.readonly = T)
plot(totali_ITA$data,totali_ITA$casi,log="y",type="o",pch
   =20,cex=0.6,lwd=1.5,ylab = "totale", xlab="tempo",yaxt
   ="n",ylim = c(1,max(totali_ITA$casi,totali_ITA$
   attualmente_positivi,totali_ITA$deceduti, totali_ITA$
   dimessi_guariti)))
lines(totali_ITA$data,totali_ITA$attualmente_positivi,
   type="o",pch=20,cex=0.6, col=4,lwd=1.5)
lines(totali_ITA$data,totali_ITA$deceduti,type="o",pch=20
   ,cex=0.6, col=2,lwd=1.5)
lines(totali_ITA$data,totali_ITA$dimessi_guariti,type="o"
   ,pch=20,cex=0.6, col=11,lwd=1.5)
lines(totali_ITA$data,totali_ITA$terapia_intensiva,type="
   o",pch=20,cex=0.6, col=6,lwd=1.5)
axis(2L,10^c(1,3,5),sprintf("10^%d",c(1,3,5)), las=2L,cex
   .axis=0.8)
par(fig = c(0, 1, 0, 1), oma = c(0, 0, 0, 0), mar = c(0,
   0, 0, 0), new = TRUE)
plot(0,type = "n",axes=F,ann=F)
legend("top",c("Casi","Dimessi/guariti","Deceduti","
   Attualmente positivi","Terapia intensiva"),xpd=TRUE,
   horiz=TRUE,lwd = 2,lty = 1,col=c(1,3,2,4,6),pch=20,cex
   =0.55,text.width = c(0.11,0.13,0.12,0.09,0.13))
```

```
# figura 2.2
par(op)
plot(variazioni_ITA$data,variazioni_ITA$dimessi_guariti,
   type="l",lwd=1,col=4, ylab="variazioni", xlab="tempo")
lines(variazioni_ITA$data,rollmean(variazioni_ITA$dimessi
   _guariti,
k=7, fill=NA),lwd=2,col=4)
lines(variazioni_ITA$data,variazioni_ITA$casi,lwd=1,col=1
   )
lines(variazioni_ITA$data,rollmean(variazioni_ITA$casi, k
   =7, fill=NA),lwd=2,col=1)
lines(variazioni_ITA$data,variazioni_ITA$deceduti,lwd=1,
   col=2)
lines(variazioni_ITA$data,rollmean(variazioni_ITA$
   deceduti, k=7, fill=NA),lwd=2,col=2)
legend("topright",c("Casi","Dimessi / guariti","Deceduti"
   ),lwd=2,lty = 1,pt.cex=0,col=c(1,4,2),cex=0.86)

# figura 2.3
posrate <- variazioni_ITA$casi[58:n]/variazioni_ITA$
   soggetti_testati[58:n]
ppr <- barplot(posrate, col=3,border=8, xlab="tempo",
   ylab="tasso di positivita")
axis(side=1, at=ppr[c(1,92-57,125-57,n-57)], labels=
   variazioni_ITA$data[c(58,92,125,n)])
lines(ppr,rollmean(posrate, k=7, fill=NA),lwd=3)
legend("topright", c("tasso di positivita giornaliero","
   media mobile a 7 giorni"),bty="n",fill=c(3,NA),border=
   c(8,NA), lty=c(NA,1), lwd=c(NA,2),x.intersp=c(0.5,1.9)
   ,y.intersp=c(1,2),cex=0.9,box.lwd=1.5)

# figura 2.4
par(mfrow=c(2,1))
acf(variazioni_ITA$casi,lag=round(n/4))
pacf(variazioni_ITA$casi, lag=round(n/4))
```

```
# figura 2.5
par(op)
attach(variazioni_regioni)
r24 <- variazioni_regioni[(regione=="Lombardia"|regione==
   "Piemonte"|regione=="Emilia-Romagna"|regione=="Veneto"
   |regione=="Lazio"| regione == "Campania"|regione=="
   Puglia"|regione=="Sicilia"),]
row.names(r24) <- seq(1:(8*n))
detach(variazioni_regioni)
par(mar = c(5, 4, 4, 4) + 0.3)
par(mfrow=c(4,2))
for(i in seq(1, 8*n, by=n)){
plot(variazioni_ITA$data, rollmean(r24$casi[i:(i+n-1)],k=
   7,fill=NA),type="l",ann="F",lwd=2)
mtext(r24[i,]$regione)
par(new=TRUE)
plot(variazioni_ITA$data,rollmean(r24$deceduti[i:(i+n-1)
   ], k=7, fill=NA),type="l",lwd=2,col=2,axes=F,ann=F)
axis(side=4,at = pretty(range(r24$deceduti[i:(i+n-1)])),
   col=2, col.ticks=2, col.axis=2)
}

# figura 2.6
par(op)
par(mfrow=c(2,1))
plot(variazioni_ITA$data,variazioni_ITA$terapia_intensiva
   ,type="l",lwd=1,col=2,ylab="variazioni", xlab="tempo",
   yaxt="n")
lines(variazioni_ITA$data,rollmean(variazioni_ITA$terapia
   _intensiva, k=7, fill=NA),lwd=2,col=2)
abline(v=totali_ITA$data[40], lwd=2,col=3)
axis(2, c(-100,0,200))
grid()
```

```
plot(totali_ITA$data[9:n],variazioni_ITA$deceduti[9:n]/
   totali_ITA$terapia_intensiva[1:(n-8)],type="l",lwd=2,
   col=4,ylab="D(t) / I(t-8)", xlab="t", yaxt="n")
abline(h=.14)
abline(v=totali_ITA$data[40], col=3,lwd=2)
axis(2, c(0.14,1))

# figura 2.7
par(op)
par(lwd = 0.2)
treat <- matrix(c(totali_ITA$isolamento_domiciliare,
   totali_ITA$terapia_intensiva,totali_ITA$ricoverati_con
   _sintomi),nrow=3,byrow=T)
timeDate <- seq.Date(from = as.Date('2020-02-24'), to =
   as.Date('2020-07-31'), by = 1)[c(1,40,79,118,n)]
colors = RColorBrewer::brewer.pal(3, "Pastel1")
options(scipen=10)
ptr <- barplot(treat,col=colors, border=RColorBrewer::
   brewer.pal(3, "Set1"),ylab = "totale",xlab="tempo")
axis(side=1, at=ptr[c(1,40,79,118,n)], labels=timeDate)
legend("topright",c("Ricoverati con sintomi","Terapia
   intensiva","Isolamento domiciliare"), bty="n",cex = 0.
   75, fill = rev(colors),border = rev(RColorBrewer::
   brewer.pal(3, "Set1")), y.intersp=c(1.75,1.75,1.75))


# simulazioni dal modello autoregressivo di Poisson log-
   lineare, sulla base della funzione (3.10)

parsim <- function(n,h,lambda0,y0,omega,alpha,beta){
    lambda <- rep(0,n)
    y <- rep(0,n)
    lambda[1] <- lambda0
    y[1] <- y0
    j <- 1
```

```
    for(i in 2:n){
      lambda[i] <- exp(omega * (1-beta^j)/(1-beta) +
      beta^j *log(lambda0) + alpha * t(beta^(0:(j-1))
      %*% log(1+y[j:1]))
      y[i] <- rpois(1,lambda[i])
      j <- j+1
      }
    return(y[(h+1):n])
    }

# figura 3.1
par(mfrow=c(1,2))
set.seed(1339)
x1 <- parsim(1000,800,10,10,0.1,0.83,0.15)
plot(x1,type="l", col=4, ylab="conteggio", xlab = "tempo"
   , ylim=c(100,300))
acf(x1,lag=round(200/4), main="")
set.seed(185)
x2 <- parsim(1000,800,10,10,0.1,0.15,0.83)
plot(x2,type="l", col=4, ylab="conteggio", xlab = "tempo"
   , ylim=c(100,300))
acf(x2,lag=round(200/4), main="")
set.seed(826)
x3 <- parsim(1000,800,10,10,0.1,0.5,0.1)
plot(x3,type="l", col=4, ylab="conteggio", xlab = "tempo"
   )
acf(x3,lag=round(200/4), main="")
set.seed(831)
x4 <- parsim(1000,800,10,10,0.1,-0.13,-0.85)
plot(x4,type="l", col=4, ylab="conteggio", xlab = "tempo"
   )
acf(x4,lag=round(200/4), main="")
set.seed(11)
x5 <- parsim(1000,800,10,10,0.1,1.65,-0.67)
plot(x5,type="l", col=4, ylab="conteggio", xlab = "tempo"
```

```
   )
acf(x5,lag=round(200/4), main="")


# applicazione empirica del modello

# per i dati nazionali si considerano anche i casi
   giornalieri dal 20/02/2020 al 23/02/2020 (fonte: World
    Health Organization)

y <- zoo(c(c(3,6,67,48,105),variazioni_ITA$casi[-1]),c(
   seq.Date(from = as.Date('2020-02-20'),to = as.Date('
   2020-02-23'), by = 'days'),variazioni_ITA$data))
plot(y)


# stima dinamica tramite "expanding windows"

# le stime dei coefficienti sono equivalenti con distr="
   poisson" e con distr="nbinom"

n <- length(y)
k=14
cf <- data.frame(matrix(0, n-k, 3), row.names=index(y)[(k
   +1):n])
colnames(cf) <- c("gamma","alpha","beta")

for(i in 1:(n-k)){
    cf[i,] <- coefficients(tsglm(y[1:(i+k)],model=list(
       past_obs=1, past_mean=1),link="log", distr="
       poisson",init.method="firstobs",start.control=list
       (method="fixed"),final.control = list(constrained
       = NULL)))
    cat("processing observation, ", i, "\n")
}
```

```
# grafico evoluzione temporale dei parametri

plot(as.Date(rownames(cf)),cf$alpha,ylab="",xlab="tempo",
   type="l",ylim=c(-1,2), lwd=2,col="orange",main="Italia
   ")
lines(as.Date(rownames(cf)),cf$beta, col=4,lwd=2)
legend("topright",c(expression(alpha),expression(beta)),
   lwd=c(2,2),lty=c(1,1),col=c("orange",4),horiz=T,cex=0.
   75,pt.cex=10)
abline(v=index(y)[21],col=2,lwd=2)


# stima sul campione completo
# per tener conto maggiormente della sovradispersione,
   gli std.error derivano dalla matrice di covarianza
   sandwich, ottenuta sulla base della verosimiglianza
   Poisson, ipotizzando un modello binomiale negativo non
    correttamente specificato

final.tsglm <- tsglm(y,model=list(past_obs=1,past_mean=1)
   ,link="log",distr="nbinom",init.method="firstobs",
   start.control=list(method="fixed"),final.control=list(
   constrained=NULL))
summary(final.tsglm)

# creazione tabella contenente i risultati della stima

estimate <- as.data.frame(cbind(summary(final.tsglm)$
   coefficients[,-c(3,4)],
2*(1-pnorm(summary(final.tsglm)$coefficients[,1]/summary(
   final.tsglm)$coefficients[,2]))))[-4,]
colnames(estimate)[3] <- "p-value"
xtable(estimate,digits=3)
```

```
# stima di R0
R0 <- rep(0, n-8)
for(i in 1:(n-8)){
    R0[i] = 7.5*(fitted(final.tsglm)[i+8]) / sum(fitted(
       final.tsglm)[i:(i+7)]) }
# grafico dinamica temporale di R0
plot(index(y)[9:n],R0, type="l",ylab="R0", xlab="tempo",
   main="Italia" ,col=4,lwd=1)
abline(h=1,col=1,lwd=1)
abline(v=index(y)[21], col=2, lwd=2)


# caricamento dei dati sui casi giornalieri in Florida,
   importati attraverso la libreria "COVID19"

library("COVID19")
flo <- covid19(country="USA",level=2, start = "2020-03-04
   ",end="2020-07-31",verbose=F)
flo <- flo[flo$administrative_area_level_2=="Florida",]$
   confirmed
y <- diff(c(0,flo))
y <- zoo(y,seq.Date(from = as.Date('2020-03-04'),to = as.
   Date('2020-07-31'), by = 'days'))

# pulizia dei dati regionali sui nuovi casi giornalieri (
   sono presenti alcuni valori negativi che non
   permettono di stimare correttamente i modelli)

variazioni_regioni$casi[variazioni_regioni$casi<0] <- c(2
   ,1,29,0,3,3,6,0,113,2,0,0,1)
y <- zoo(variazioni_regioni$casi[variazioni_regioni$
   regione=="Lombardia"],variazioni_ITA$data)

# si ripete l'intero procedimento di cui sopra per i dati
    della Florida e delle regioni d'interesse
```

Bibliografia

[1] A. Agosto, A. Campmas, P. Giudici e A. Renda. *Monitoring COVID-19 contagion growth in Europe*. CEPS Working Papers, 30 marzo 2020. https://www.ceps.eu/ceps-publications/monitoring-covid-19-contagion-growth-in-europe/.

[2] A. Cerqua e R. Di Stefano. *When did coronavirus arrive in Europe?* 2020. https://www.researchgate.net/publication/340568094_When_did_coronavirus_arrive_in_Europe.

[3] A. Farcomeni, A. Maruotti, F. Divino, G. Lasinio e G. Lovison. *An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian Regions*. 2020. https://arxiv.org/abs/2005.11975.

[4] R. Ferland, A. Latour e D. Oraichi. «Integer-Valued GARCH Process». In: *Journal of Time Series Analysis* 27 (2006), pp. 923–942.

[5] K. Fokianos e D. Tjøstheim. «Log-linear Poisson autoregression». In: *Journal of Multivariate Analysis* 102 (2011), pp. 563–578.

[6] C. Gu, W. Jiang, T. Zhao e B. Zheng. *Mathematical recommendations to fight against COVID-19*. 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract%20id=3551006.

[7] W. Kermack e A. McKendrick. «A contribution to the mathematical theory of epidemics». In: *Proceedings of the Royal Society of London. Series A*. 115 (1927), pp. 700–721.

[8] T. Liboschik, K. Fokianos e R. Fried. «tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models». In: *Journal of Statistical Software* 82 (2017), pp. 1–51.

[9] F. Peracchi. *The Covid-19 pandemic in Italy*. EIEF, 2020. `http://www.eief.it/eief/index.php/forecasts`.

[10] N. Shao, J. Cheng e W. Chen. *The reproductive number R0 of COVID-19 based on estimate of a statistical time delay dynamical system*. Cold Spring Harbor Laboratory Press, 2020. `https://www.medrxiv.org/content/10.1101/2020.02.17.20023747v2`.

[11] T. Sim. «Maximum Likelihood Estimation in Partially Observed Markov Models with Applications to Time Series of Counts». Ph.D. thesis. Télécom ParisTech, Paris, 2016. `https://pastel.archives-ouvertes.fr/tel-01458087/document`.