UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Scuola di Economia e Statistica

Corso di laurea in Scienze Statistiche ed Economiche


# IL CORONAVIRUS: MODELLI STATISTICI PER L'ANALISI DELLA DIFFUSIONE NEL TERRITORIO LOMBARDO

Relatore:
**Ch.mo prof. Aldo Solari**

Tesi di Laurea di:
**Matteo Baldanza**
**Matr. N. 826018**

**ANNO ACCADEMICO**
**2019/2020**

Ai miei genitori e
alla mia famiglia

*"Se torturi i numeri a lungo,*
*confesseranno qualsiasi cosa"*
*Gregg Easterbrook*

# Indice

**Elenco delle figure** 5

**Introduzione** 6

**1 Analisi dei dati** 8

1.1 L'analisi descrittiva dei dati . . . . . 8

1.2 La prevalenza: Italia e Hubei . . . . . 10

1.3 Incidenza: Italia e Hubei . . . . . 12

1.4 Analisi delle varie misure e problematiche annesse . . . . . 14

1.5 La mortalità in Italia . . . . . 15

**2 La stima di $R_0$** 19

2.1 Introduzione ai modelli epidemiologici . . . . . 19

2.2 Il parametro $R_0$ . . . . . 20

2.2.1 Alcune delucidazioni su $R_0$ e differenze con $R_t$ . . . . . 22

2.3 Generation-Time . . . . . 23

2.4 Stima di $R_0$ tramite il tasso di crescita esponenziale . . . . . 25

2.4.1 Introduzione . . . . . 25

2.4.2 L'equazione di Lotka-Eulero . . . . . 26

2.4.3 Derivazione di $R_0$ dall'equazione di Lotka-Eulero . . . . . 28

2.4.4 $R_0$ utilizzando campione di osservazioni cinesi e italiane . . . . 29

2.5 Stima di $R_0$ tramite la funzione di massima verosimiglianza . . . . . 31

2.5.1 Derivazione della stima di massima verosimiglianza . . . . . 31

2.5.2 $R_0$ utilizzando campione di osservazioni cinesi e italiane . . . . 35

**3 I modelli epidemiologici** 36
3.1 Modello $SIR$ . . . 36
3.1.1 In termini teorici . . . 36
3.1.2 Relazione $R_0$ con il numero di infetti . . . 38
3.1.3 Applicazione modello SIR in Lombardia . . . 39
3.1.4 L'andamento di $R_t$ in Lombardia . . . 41
3.2 Modello $SEIR$ . . . 45
3.2.1 In termini teorici . . . 45
3.2.2 Applicazione del modello in Lombardia . . . 47

**4 La struttura dei mezzi pubblici in Lombardia** 50
4.1 La matrice OD . . . 50
4.1.1 La matrice OD Lombarda . . . 51
4.2 Analisi della matrice . . . 53
4.3 I collegamenti dei mezzi pubblici Lombardi . . . 56

**5 Modello stocastico $SEII_aR$** 58
5.1 Introduzione al modello . . . 58
5.2 La costruzione del modello . . . 59
5.2.1 Le equazione stocastiche . . . 59
5.2.2 Mobilità e dinamica globale delle infezioni . . . 62
5.3 La scelta dei parametri . . . 63
5.4 Il modello applicato alla rete Lombarda . . . 64

**Conclusioni** 66

**Appendice** 68

**Bibliografia** 91

# Elenco delle figure


1.1 Curve epidemiche Italia e Hubei, in rosso i picchi . . . . . 11
1.2 Curva incidenza Hubei, in rosso il suo picco . . . . . 12
1.3 Curva incidenza Italia, in rosso il suo picco . . . . . 13

2.1 Intervallo Seriale Sars vs Covid-19 . . . . . 25

3.1 Modello SIR . . . . . 37
3.2 SIR Lombardia . . . . . 40
3.3 Differenze adattamento modelli ai dati . . . . . 42
3.4 Andamento effettivo $R_t$ in Lombardia . . . . . 44
3.5 Modello SEIR . . . . . 45
3.6 SEIR Lombardia . . . . . 48

4.1 Mezzi di trasporto più utilizzati in Lombardia . . . . . 54
4.2 Motivi di spostamento in Lombardia . . . . . 55
4.3 Mezzi di trasporto più utilizzati in Lombardia per motivo . . . . . 55
4.4 La rete dei trasporti dei mezzi pubblici Lombardi . . . . . 56

5.1 Funzionamento modello stocastico di tipo $SEII_aR$ . . . . . 60
5.2 Andamento incidenza tramite modello stocastico . . . . . 64

# Introduzione

Malattie ed epidemie da sempre sono una delle maggiori cause di morte nel mondo, anche più di guerre e carestie. Gli agenti patogeni si adattano e si evolvono nel corso del tempo, determinando così, la comparsa di nuove malattie o il rigenerarsi di antiche forme di contagio che si pensavano estinte. È chiaro che l'invasione umana con nuovi ecosistemi, fattori quali il surriscaldamento globale, il degrado ambientale, il potenziamento degli spostamenti e comunicazioni su scala internazionale, e i cambiamenti in campo economico, abbiano cominciato a fornire sempre più terreno fertile per la generazione di nuove malattie.

Dati gli eventi accorsi in questo inizio anno, si è deciso di destinare il lavoro di tesi sullo studio del *Covid-19*, meglio conosciuto come coronavirus. La malattia infettiva, causata da un virus mai identificato prima negli esseri umani, ha prodotto sconforto, morti e misure straordinarie (atte al contenimeto e al rallentameno della diffusione) in ogni parte del mondo come mai prima ad ora nel dopo guerra. L'idea di questo lavoro nasce prima dell'arrivo della malattia nel territorio italiano quando ancora i soli casi erano limitati al territorio cinese. Verranno affrontati capitolo per capitolo i passi necessari per la costruzione di un modello epidemiologico in grado di stimare come si sarebbero trasmesse e con che velocità le infezione in Lombardia sulla base degli spostamenti effettuati dalle persone tramite mezzi pubblici. Lo scopo finale è di capire come i decreti abbiano limitato e cambiato la diffusione della malattia cercando di capirne il mese di arrivo nel suolo Lombardo.

Nel capitolo 1 si affronterà l'analisi dei dati. Il primo passo sarà dato dalla costruzione della curva epidemica del virus, fondamentale per tutte le analisi successive, per poi affrontare le varie misure utilizzate nel mondo per l'analisi dei dati attraverso una loro visione critica. Come ultimo passo in questo capitolo si troverà il rischio di morte della malattia andandolo a confrontare con quello dell'influenza, mettendo chiarezza a quanto mal comunicato da media e giornali in questi ultimi mesi.
Il capitolo 2 costituisce una parte importante del lavoro. Si affronterà la spiegazione e la delucidazione del parametro $R_0$, con le relative differenze con $R_t$. Si stimerà tale parametro, fondamentale per capire la velocità di diffusione della malattia, tramite due tecniche statistiche: *"tasso di crescita esponenziale e massima verosimiglianza"*.

Lo studio verrà effettuato in parallelo sia per la Cina che per l'Italia, per trovare e comparare eventuali differenze. Il parametro sarà inoltre necessario all'interno dei modelli.
Nel capitolo 3 saranno costruiti i modelli quali SIR e SEIR per cercare di comprendere come si sarebbe diffusa l'epidemia senza l'intervento dei decreti emessi. Rielaborando le equazioni differenziali del modello SIR e integrandole a tecniche quali regressione lineare e di Poisson si stimerà e si valuterà l'andamento nel tempo del parametro $R_t$.
Nel capitolo 4 verrà affrontato lo studio di una particolare matrice il cui nome è *Orgine-Destinazione*, contenente informazioni relative agli spostamenti degli individui durante un giorno lavorativo. L'analisi sarà effettuata sulla Lombardia, con una breve analisi generale, per l'estrapolazione di alcune informazioni di interesse. Si costruirà poi, tramite le informazioni presenti all'interno della matrice, la rete dei trasporti pubblici nel territorio Lombardo.
Nell'ultimo capitolo si applicherà un modello epidemiologico complesso di tipo stocastico che riflettendo gli spostameti degli individui in Lombardia cercherà di valutare un possibile impatto del virus senza la considerazione dei vari decreti emessi dal governo (hanno limitato il numero degli spostamenti). Tramite questo modello si cercherà di capire quando il Covid-19 è effettivamente arrivato in Italia.
Si riporta nell'appendice il codice utilizzato nell'ambiente di lavoro RStudio.

# Capitolo 1

# Analisi dei dati

## 1.1 L'analisi descrittiva dei dati

La descrizione accurata dell'epidemia costituisce il primo importante passo per comprendere un virus. Un analisi descrittiva dell'episodio può bastare per indirizzare le misure immediate di controllo a sviluppare ipotesi sulla sorgente di infezione e sulle modalità di trasmissione. Per questi motivi il primo passo è quello della costruzione di misure, come quella della curva epidemica (uno tra i più utilizzati), che diano informazione riguardanti la situazione reale.
Ponendo in un grafico il numero di casi (incidenza) e il tempo si osserverà una curva, detta epidemica, nella quale sarà possibile osservare la crescita dell'infezione[1]. Lo scorrere del tempo è visualizzato sull'asse delle $x$ mentre i conteggi dei casi vengono visualizzati sull'asse delle $y$. Il risultato è una rappresentazione visiva dell'insorgenza della malattia nei casi associati a un'epidemia. È utile perché può fornire molte informazioni tra cui:

- Descrizione della diffusione
- Dimensione dei contagi nel tempo
- Valori anomali
- Periodo di incubazione della malattia

Tali aspetti dal punto di vista matematico/statistico possono essere spiegati in base al tipo di analisi che viene effettuata.

La descrizione della diffusione può essere osservata tramite misure come la curtosi, che rappresenta un allontanamento dalla normalità[1] distributiva, rispetto alla quale si verifica un maggiore appiattimento o un maggiore allungamento. In base all'appiattimento della curva si possono comprendere fenomeni come la velocità della diffusione della malattia. Una distribuzione platicurtica è indice di diffusione lenta. Una distribuzione leptocurtica rappresenta grossi problemi, in quanto essendoci diffusioni rapide, il sistema sanitario di un determinato paese potrebbe non farcela.

La dimensione dei contagi nel tempo di una malattia può essere trovata visualizzando i picchi[2] di una determinata curva epidemica. Esistono due tipi di picchi per ogni indicatore, che viene utilizzato. Definendo un individuo infetto come colui che è contaminato da microrganismi infettivi, e che può quindi trasmettere o provocare un'infezione, si possono trovare i relativi picchi di questo tipo di misura definendo $f(t)$ come una funzione che descrive il numero dei casi infetti al tempo $t$:

1 Il picco dei nuovi casi (picco dei casi di incidenza): coincide con il conteggio più alto dei nuovi casi osservati di una malattia in un determinato istante temporale (ad esempio un giorno se ci si riferisce a dati analizzati giornalmente)

2 Il picco epidemico rappresenta il punto di massimo dei casi di prevalenza, ovvero il più alto numero di casi totali presenti in una popolazione di interesse, osservati durante l'epidemia stessa

L'individuazione dei valori anomali assume alta importanza nell'utilizzo delle varie misure. Un cambio di metodologia nella rilevazione degli individui infetti, ad esempio aumentando il numero di tamponi effettuati ogni giorno, può portare a interpretazioni sbagliate. Se nel giorno $t$ vengono effettuati 2000 tamponi (si supponga che nei giorni $t-1$, $t-2$... siano stati sempre lo stesso numero), di cui 1000 positivi, e nel giorno successivo $t+1$ ne vengono effettuati 9000, di cui 2000 non negativi, il confronto diretto di persone trovate positive non avrebbe alcuna rilevanza statistica senza tenere in considerazione il numero di tamponi effettuati. Sarebbe infatti sbagliato affermato che nel periodo $t+1$ siano presenti più casi rispetto ai precedenti. È quindi importante nel caso vengano rilevate osservazioni anomale di cercare di comprendere il motivo che le rende tali (nell'esempio significherebbe indagare il motivo dei più tamponi effettuati) per valutarne una loro esclusione dall'analisi.

Nell'incidenza, che rappresenta i nuovi casi giornalieri e che viene rappresentata da istogrammi (grafico a barre), la distanza temporale tra i vari massimi relativi (barre più alte) è molte volte indice del periodo di incubazione del virus.

---

[1] una curva epidemica è approssimabile a una distribuzione normale [2]

[2] i differenti picchi verranno esposti in maniera più chiara tramite visualizzazione grafica nei successivi due paragrafi

## 1.2 La prevalenza: Italia e Hubei

Un tipo di curva epidemica, che rappresenta uno degli strumenti più utilizzati insieme all'incidenza, è quella data dalla prevalenza (indicatore dei nuovi casi giornalieri). La prevalenza indica il numero di individui (a volte anche espresso in termini percentuali) di una certa popolazione che riscontra una malattia in un determinato momento temporale. Se si verificano nuovi casi di persone infette la curva avrà tendenza crescente. Tuttavia se gli individui con la malattia muoiono (il tasso di mortalità è differente da zero) la prevalenza diminuirà di conseguenza. Anche il numero dei guariti incide su questo calcolo nella stessa maniera del numero delle morti e quindi questo potrebbe costituirne un limite. Pertanto è importante esaminare oltre alla prevalenza altre tipo di curve come l'incidenza.

È bene chiarire subito i due concetti con un esempio. Se si pensa ad una vasca da bagno l'incidenza è l'acqua che viene aggiunta nella vasca ogni lasso di tempo $t$ mentre la prevalenza è il contenuto della vasca all'istante di tempo $t$ e i casi che muoiono/guariscono sono rappresentati dall'acqua che esce per evaporazione.

Si riassumente quanto detto in termini algebrici elementari:

$$X(t) \; = \; N(t) \; - M(t) \; - \; G(t) \tag{1.1}$$

Ove:

- $X(t)$ sono le persone attualmente positive al tempo $t$
- $N(t)$ il numero totale di casi al tempo $t$
- $M(t)$ il totale dei morti al tempo $t$
- $G(t)$ il totale dei guariti al tempo $t$

La curva epidemica basata sulla prevalenza per l'Italia e Hubei[3] è data dalla figura 1.1. Sull'asse delle $y$ si trovano gli attualmente positivi mentre sull'asse delle $x$ si ha l'andamento temporale.

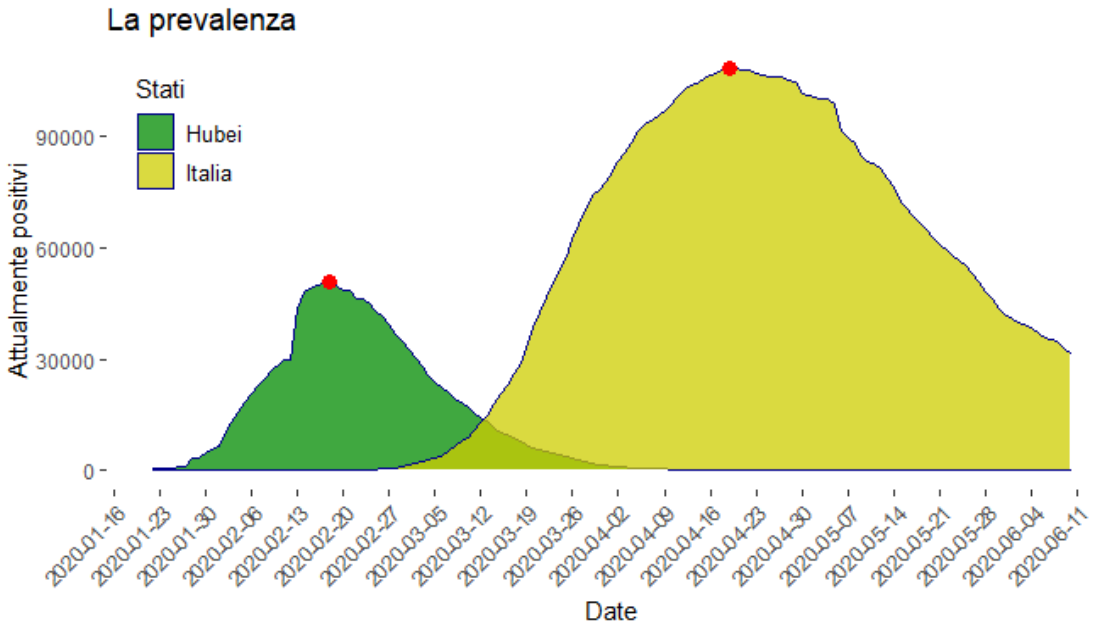


Figura 1.1: Curve epidemiche Italia e Hubei, in rosso i picchi

I periodi di inizio del virus sono differenti cosi come la loro diffusione. Il picco più alto, indice delle persone attualmente positive in un certo istante di tempo (giorno), indica che la dimensione del contagio è stata maggiormente sviluppata nel ù territorio. In parole semplice, in Italia si è visto un raggiungimento di positivi maggiore rispetto ad Hubei. La curva cinese appare inoltre meno piatta, il che vuol dire che, come spiegato in precedenza, la diffusione è stata più veloce ed è terminata prima.

[3]provincia Cina sede del primo focolaio

## 1.3 Incidenza: Italia e Hubei

La prevalenza differisce quindi dall'incidenza della malattia che rappresenta il numero di nuovi casi che si sviluppano in un determinato periodo di tempo[4]. L'incidenza permette di verificare l'andamento nel tempo giornaliero dei nuovi casi all'interno del territorio di riferimento.

Si parte a visionare questa misura per quanto riguarda la provincia cinese. In questo caso sull'asse delle $x$ si ha sempre il tempo mentre sulle $y$ si hanno i valori dei nuovi casi positivi giornalieri. Si aggiunge anche un grafico con finestra mobile pari a 3 giorni per aver una miglior visuale del trend.

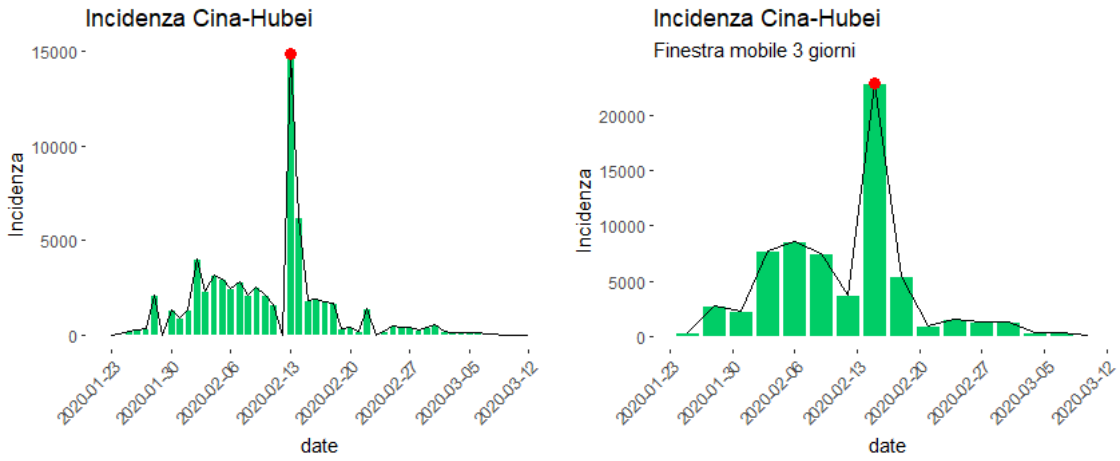


Figura 1.2: Curva incidenza Hubei, in rosso il suo picco

Il picco di nuovi casi facilmente osservabile, rappresenta un valore anomalo che si spiega con un cambio di metodologia di rilevazione dei positivi. Infatti in quel giorno i tamponi sono stato effettuati su larga scala a differenza del giorno precedente, dove i tamponi erano riservati solo ai casi sospetti di individui con sintomi. La considerazione è importante in quanto il cambio di rilevazione comporta errori nell'analisi se non se ne tiene conto. L'andamento dei nuovi contagi è quasi costante nel tempo. Non si osserva una particolare crescita esponenziale, molto probabilmente non rilevata dal metodo di tamponamento limitato e decresce gradualmente fino ad azzerarsi (indice di positivi pari a zero).

[4]il tempo utilizzato per il calcolo dei nuovi casi è di un giorno

Visionando l'incidenza nel territorio italiano si trova invece un andamento differente. Si aggiunge anche in questo caso una finestra mobile di tre giorni per visualizzare al meglio il trend. Il colore nel grafico rappresenta il numero dei tamponi effettuati giornalmente/ogni tre giorni rispettivamente nel grafico senza finestra mobile e con finestra mobile.

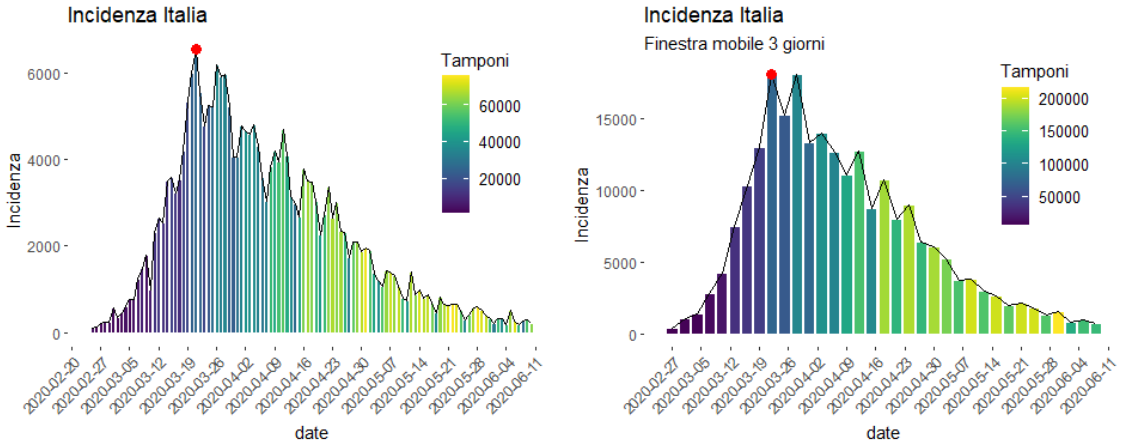


Figura 1.3: Curva incidenza Italia, in rosso il suo picco

Dalla curva si può notare un andamento esponenziale di crescita molto più omogeneo rispetto a quello di Hubei. Si vede inoltre che la crescita sembra arrestarsi intorno al primo aprile, a seguito probabilmente delle misure attuate dal governo. Quello che appare molto interessante è che i vari picchi si trovano in corrispondenza di un intervallo temporale di circa 5-6-7 giorni che corrisponde al periodo medio di incubazione del Covid-19.
Talvolta una variazione dell'incidenza può essere soltanto apparente ed essere dovuta a un artefatto come nel caso in cui vengano utilizzati nuovi metodi diagnostici che consentono di apprendere la malattia già al suo stato precoce. In questo caso anche se l'incidenza reale della malattia rimane costante, attraverso il nuovo strumento diagnostico verrà individuato un numero maggiore di casi rispetto a quanto si era verificato in passato e ciò corrisponderà appunto a un aumento dell'incidenza soltanto apparente. In questo casi si ha un andamento costante dal 15 al 30 marzo circa a fronte però di un numero di tamponi ogni giorno più elevati, il che indica che l'incidenza in realtà non è costante ma in diminuzione in tale periodo.
Un'ultima osservazione ricade sulla parte finale del grafico in cui si nota che a fronte di un numero sempre più elevato di tamponi l'incidenza diminuisce.

## 1.4 Analisi delle varie misure e problematiche annesse

Ogni qual volta si fa affidamento ad un certo tipo di misura bisogna capire ed analizzare che tipo di metodo si sta utilizzando per comprenderne i relativi limiti. Si riassumono le misure più utilizzate in ambito dello studio del Covid-19 (utilizzate quotidianamente per la comunicazione dei dati), le loro caratteristiche e i loro limiti [3]:

- **Frequenza cumulata di nuovi casi:** È l'indicatore maggiormente diffuso e utilizzato. Definendo la frequenza cumulata, associata ad una modalità[5] (in questo specifico caso trattasi di nuovi casi giornalieri), come la somma della sua frequenza assoluta e di quelle delle modalità che la precedono, questo indicatore rappresenta un andamento cumulativo di casi giornalieri. Questo metodo è molto debole dal punto di vista rappresentativo perchè dipende dal numero dei tamponi che vengono effettuati quotidianamente. Inoltre il numero dei nuovi casi del giorno $t$ rappresenta in realtà i casi del giorno $t-1$ in quanto i risultati dei tamponi sono disponibili soltanto il giorno seguente[4]

- **Il numero dei decessi giornalieri:** La qualità del dato è pessima. I risultati di questi numeri comunicati quotidianamente da protezione civile si riferiscono solamente ai casi di morte certa con Covid-19 all'interno degli ospedali. Non vengono quindi considerate tutte le morti delle persone all'interno delle abitazioni, che non sono poche, e le morti dovute al coronavirus non certificate dal tampone positivo a causa del loro numero limitato

- **Il numero totale delle morti:** Tramite frequenza cumulata è impossibile individuare un trend crescente/decrescente dei deceduti, visibile soltanto tramite il calcolo di morti giornaliere che per i motivi descritti in precedenza non sono attendibili. Il numero totale delle morti è quindi poco utile a fornire un immagine precisa dell'andamento della pericolosità della malattia.

- **Numeri registrati su una scala logaritmica:** È una misura che visualizza i dati trasformati in scala logaritmica su un piano cartesiano. L'asse delle $y$ viene etichettata ad esempio da $1, 10, 100, 1000$ mentre l'asse delle $x$ rappresenta il tempo. Questo tipo di misura è utile per la visione e il confronto dei vari trend che esprimono l'andamento degli infetti tra paesi e nazioni ma non viene utilizzato, in quanto poco significativo, per valutare l'andamento della malattia

[5] in statistica una modalità di un carattere è uno dei suoi possibili valori

- **Previsione tramite modelli:** Modelli statistici più o meno avanzati vengono utilizzati per prevedere sia l'andamento dei casi infetti (per capirne i giorni necessari per il raggiungimento del massimo assoluto all'interno della curva epidemica) sia per capire entro quanto tempo i contagi raggiungono soglia zero. Costituiscono una grande importanza all'interno dello studio di una malattia. Una previsione rende consapevoli i paesi dei vari rischi e da quei risultati, capendo la portata della malattia, possono essere prese varie decisioni come la quarantena,chiusura dei mezzi pubblici....
  Il problema di tali modelli è che sono molto imprecisi il che può essere spiegato dall'incertezza dei dati raccolti. È molto difficile prevedere gli infetti futuri sulla base di un numero di infetti, raccolto tramite tamponi, che è per certo sbagliato. Ad oggi infatti nessun modello è stato in grado di predire correttamente quello che poi è successo.
  Alcune volte comunque vengono riportati dai media e telegiornali modelli senza incertezze sulle previsioni e senza spiegarne le problematiche annesse

- **Letalità del virus:** Spesso vengono citate percentuali di letalità del virus generiche che non considerano l'eta, i problemi degli individui non sani e il genere dell'individuo, il che costituisce un informazione non completa. Non si conosce il motivo ma i dati parlano in modo chiaro, apparentemente muoiono più maschi che femmine (*donne* $\approx 41\%$), e l'età colpita maggiormente è quella degli over 65 con problematiche di salute già esistenti. L'età media dei pazienti deceduti e positivi in Italia, i cui dati sono però molto simili nel mondo, è di 80 anni. È stato inoltre trovato su un campione di 3200 deceduti che soltanto il 4% di loro non aveva patologie pregresse[5]

## 1.5 La mortalità in Italia

In Italia in data 29/05/2020 si contano 33 229 decessi. Media, telegiornali e politici hanno e comunicano ancora oggi tantissimi numeri riguardanti il rischio di mortalità, tutti sempre differenti fra di loro, portando a molta confusione. Bisogna quindi capire il motivo di questi risultati differenti anche se è evidente che in base a come viene calcolata la mortalità, i risultati saranno completamente diversi.

Si parte da un esempio semplice per poi fare le stesse considerazione nel territorio Italiano. Per capire meglio il concetto si supponga che si riesca a sapere in qualunque istante di tempo il numero esatto di persone infette, sia con sintomi sia senza (le persone che non presentano sintomi si definiscono asintomatiche). La popolazione di riferimento puramente immaginaria si suppone essere composta da 100 persone all'istante $t_0$. All'istante $t_1$ si scopre che sono presenti 20 persone infette di cui 7 hanno presentano i sintomi della malattia e 13 no. Si considera infine un tempo $t_2$

dove si hanno ancora lo stesso numero di persone infette del tempo $t_1$ con la differenza che tra quelle persone sintomatiche, 5 sono guarite (una persona si definisce guarita quando dopo la presenza dei sintomi ritorna in stato di salute) e 2 sono morte.

In questo esempio si calcola il rischio di morte con i metodi standard che sono i seguenti:

- **Case fatality rate (CFR):** È la percentuale di decessi per una determinata malattia rispetto al numero totale di persone che l'hanno riscontrata e ne hanno manifestato i sintomi riuscendo a guarire. Non vengono quindi considerati in questo calcolo gli asintomatici. Si intuisce quindi come il numero finale sia il più delle volte alto. Questo tipo di indice spesso viene confuso con il rischio di morte ma in realtà è usato soltanto per avere informazioni sui deceduti e/o sul funzionamento delle varie misure restrittive adottate. E' anche utile per capire in quali fasce di età si rilevano un maggior numero di decessi.

  Si può definire il case fatality rate sulla base dell'esempio precedente in questo modo:

  $$Case\ fatality\ rate\ =\ \frac{morti_{t=2}(2)}{guariti_{t=2}(5)}\ =\ 0.4\ =\ 40\%$$

- **Infection fatality rate (IFR):** A differenza del precedente è un indice di rischio di morte ma solo se si è infetti. Rappresenta semplicemente il rapporto tra morti e persone risultate positive al virus, in questo caso quindi vengono inclusi anche gli asintomatici.

  In termini algebrici molto elementari si può calcolare per l'esempio precedente:

  $$Infection\ fatality\ rate\ =\ \frac{morti_{t=2}(2)}{infetti_{t=2}(20)}\ =\ 0.10\ =\ 10\%$$

  In questo caso quindi si ha che una persona infetta ha la probabilità del 10% di morire

- **Crude mortality rate (CMR):** Come il precedente è un indice di rischio di morte ma al posto di essere confrontato con tutti gli infetti in un certo istante di tempo viene confrontato con l'intera popolazione presente all'istante di tempo iniziale.

  $$Crude\ mortality\ rate\ =\ \frac{morti_{t=2}(2)}{popolazione_{t=0}(100)}\ =\ 0.02\ =\ 2\%$$

  Quest'ultimo è l'indice corretto per capire quanto sia alta la letalità di una malattia.

Ora si procede mostrando come i vari indici sono utilizzati all'interno del contesto del territorio italiano. In primo luogo si procede col mostrare l'utilità del **case fatality rate**, cercando di capire se i decreti hanno effettivamente abbassato questo tasso. I tempi considerati sono istante temporale $t = 10/03/2020$ e $t = 10/05/2020$. I risultati sono i seguenti:

$$CFR_{t=10/03/2020} = \frac{morti = 631}{guariti = 1004} = 0.628 \approx 63\%$$

$$CFR_{t=10/05/2020} = \frac{morti = 30560}{guariti = 105186} = 0.290 \approx 29\%$$

Il tasso quasi dimezzato indica come i decreti abbiano aiutato alla riduzione delle morti, ma è totalmente sbagliato definire come rischio di morte uno dei due risultati.

Si procede ora con un confronto diretto tra **infection fatality rate** e **crude mortality rate** entrambi calcolati all'istante temporale $t = 29/05/2020$. Di solito è sempre preferibile il calcolo di questi indici quando la malattia cessa di esistere, perchè rende il risultato *definitivo*. Calcolandolo invece con intervalli temporali si ha un valore sicuramente differente (anche se molto simile), che però permette di avere un'idea molto precisa.

$$IFR_{t=29/03/2020} = \frac{morti = 33229}{infetti = 152844} = 0.22 \approx 22\%$$

$$CMR_{t=10/05/2020} = \frac{morti = 33229}{popolazione = 60{,}36mln} = 0.00056 \approx 0.056\%$$

Il risultato dato dall'infection fatality rate è stato spesso usato dai telegiornali e può creare preoccupazioni, ci dice infatti che una persona infetta ha il 22% di probabilità di morire (chiaramente in quel periodo di tempo). Bisogna però evidenziare il limite di questo calcolo che nell'esempio effettuato in precedenza (quello della popolazione di 100 persone) è stato sorvolato ipotizzando di conoscere con certezza il numero di infetti in ogni instante temporale. Questo come ben noto non è possibile, essendo i tamponi giornalieri limitati, e quindi il suo risultato è privo di significato. Non ha senso utilizzarlo ne per capire se le politiche adottate sono corrette ne per confrontare i rischi di morte con altri paesi, è facile intuire infatti come ogni paese abbia un numero di tamponi differente.

Per questo motivo quello maggiormente utilizzato è il CMR che può essere confrontato anche con differenti malattie per capire le loro differenze. Un confronto doveroso è quello con l'influenza per capirne se effettivamente il rischio di morte è maggiore. Il CMR dell'influenza è il seguente:

$$CMR_{t=2019} = \frac{morti \approx 10000}{popolazione = 60{,}36mln} = 0.00016 \approx 0.016\%$$

Il risultato che ne deriva è che il Covid-19 ha una letalità di 0.035 punti percentuali maggiore.

Come ultima considerazione bisogna osservare che il numero delle morti di coronavirus in Italia pari a 33 229, ma così come nel resto del mondo, è un numero basato sui morti che abbiano effettuato un tampone risultato positivo al virus. A molti deceduti non è stato effettuato quest'ultimo, forse per mancanza di strumenti a sufficienza, ma quello che ne risulta è un numero comunque *sottostimato*. Bisogno inoltre considerare le persone che sono venute a mancare direttamente nelle loro abitazioni che chiaramente non hanno effettuato il test. È quindi formalmente più corretto definire che il coronavirus ha una letalità $\approx 3.5$ volte maggiore di quella dell'influenza, che potrebbe essere leggermente più alta.

# Capitolo 2

# La stima di $R_0$

## 2.1 Introduzione ai modelli epidemiologici

I modelli svolgono un ruolo importante nella comprensione e gestione delle dinamiche di trasmissione dei vari agenti patogeni. Possono essere utilizzati per descrivere nel tempo l'evoluzione della malattia, nonché per esplorare e/o comprendere meglio i fattori che influenzano l'incidenza dell'infezione. La modellizzazione è quindi un passo fondamentale per comprendere quali trattamenti e interventi possono essere più efficaci e quali fattori specifici devono essere considerati quando si cerca di combattere un infezione.

Per comprendere le dinamiche complesse alla base della trasmissione delle malattie, si usano spesso una serie di modelli chiamati *modelli compartimentali*. I modelli compartimentali (di cui ci si occuperà in questo lavoro) sono una tecnica utilizzata per semplificare la modellazione matematica di malattie infettive. La popolazione viene divisa in compartimenti con il presupposto che ogni individuo nello stesso scomparto abbia le stesse caratteristiche[6]. Questi tipi di modelli sono di solito formati da equazioni differenziali (che sono deterministiche), ma possono anche avere una forma stocastica[1], che è una rappresentazione più realistica ma più complicata da analizzare. In questo caso i modelli deterministici modellano la diffusione tramite parametri fissi mentre quelli stocastici introducono variabili casuali (come la binomiale) per spiegare il "successo" o "insuccesso" di un infezione.

I modelli compartimentali possono essere utilizzati per prevedere la grandezza di diffusione della malattia oppure per prevederne il raggiungimento del picco.

[1] modello matematico adatto a studiare l'andamento dei fenomeni che seguono leggi casuali, probabilistiche

Bisogna comunque fare molta attenzione alle caratteristiche individuali delle persone che talvolta sono differenti, ad esempio un individuo può essere asintomatico, può contrarre più facilmente una malattia perchè più anziano...

Si è quindi detto che il funzionamento dei modelli avviene secondo una precisa suddivisione della popolazione. Qui di seguito si riportano le categorie in cui le persone, appartenenti ad una certa popolazione, vengono suddivise:

- **Suscettibili:** Individui che possono essere infettati
- **Esposti:** Individui infettati che a causa del periodo di incubazione dell'agente patogeno, non sono ancora infettivi e non hanno manifestano sintomi
- **Infetti:** Individui infetti da un agente patogeno ed in grado di trasmettere l'infezione ad altri. Un individuo è infetto quando presenta i sintomi
- **Guariti:** L'individuo non è più contagioso o è "rimosso" dalla popolazione (con rimosso si intende la sua morte)

## 2.2 Il parametro $R_0$

Si è spiegato in maniera semplicistica, il funzionamento dei vari tipi di modelli che chiaramente necessitano di alcuni parametri che si definiscono "fondamentali" senza i quali non si possono inizializzare i vari algoritmi. Naturalmente tanto più complesso è un modello tanto più i parametri diventano numerosi e difficili da calcolare. Uno di questi però indipendentemente dal modello utilizzato compare sempre e si definisce **basic reproduction number**, più conosciuto come $R_0$.

$R_0$ è un termine matematico che rappresenta una misura della velocità con cui una malattia infettiva progredisce inizialmente[7]. $R_0$ è caratteristico dell'epidemia nella sua fase iniziale, quindi per capire se le misure di contenimento hanno effetto, viene considerato il numero di riproduzione effettivo, indicato con $R_t$[2], in quanto si riferisce ad un giorno generico t. $R_0$ è sicuramente importante, perché fornisce l'informazione sintetica di quanti casi secondari (casi che si verificano in tempi successivi al periodo di infezione del caso primario[3] che derivano dal contatto con essi) vengono generati. Risulta infatti ovvio che quando un'infezione si diffonde a nuove persone, si riproduce. Il parametro fornisce quindi un informazione sintetica di

[2]si spiegherà poco più avanti la loro relazione nel dettaglio

[3]la terminologia può applicarsi solo a malattie infettive che si diffondono da uomo a uomo, e si riferisce alla persona/gruppo che per prima/i trasmette/ono una malattia ad un gruppo di persone

quanti casi secondari vengono generati, per trasmissione interpersonale, da un caso primario in una popolazione completamente suscettibile. Ne consegue che una epidemia si instaura quando per ogni caso primario si generano più casi secondari e da ognuno di questi vengono generati altri casi. All'opposto, se ogni singolo caso non ne contagia nessun altro, la circolazione dell'infezione è destinata ad estinguersi. Intuitivamente è apprezzabile che il valore è direttamente proporzionale al numero di contatti per giorno del caso primario (più persone incontra, più persone si infettano), alla durata della sua fase di contagiosità (più a lungo rimane contagioso, più è alto il numero delle persone che contagia) e alla probabilità di trasmissione dell'infezione per singolo contatto (un individuo infetto non è detto che trasmetta la malattia a tutti gli individui che incontra). Tutte queste quantità sono difficili da osservare direttamente e in genere ci si basa su stime, sotto diverse assunzioni, che vengono utilizzate per la costruzione dei modelli statistici che si affronteranno nei prossimi capitoli. La spiegazione è che risulta impossibile avere a disposizioni i dati sul numero di infetti giornalieri, per motivi come il numero di tamponi effettuati che risultano limitati. Se questa informazione fosse comunque nota allora il calcolo diretto sarebbe molto semplice, perchè basterebbe dividere gli individui infetti al periodo di tempo $t_x$ con quello $t_{x-1}$ (ove $x$ arco temporale di uno dei primi giorni della diffusione della malattia).

Quando si osserva una trasmissione di contagi nella popolazione generale come il coronavirus, $R_0$ viene stimato in modo empirico, ossia osservando la velocità di crescita del numero totale dei casi giorno dopo giorno (stima tramite tasso di crescita esponenziale). Sapendo la data di insorgenza dei sintomi, il tempo di incubazione e l'intervallo di tempo tra la comparsa dei sintomi nel caso primario e la comparsa dei sintomi nei casi secondari (questo intervallo di tempo si definisce Generation-Time) è possibile ricostruire le diverse generazioni di casi e stimare l'indice di riproduzione.

Se una malattia ad esempio ha un $R_0$ di 1.30[4], una persona che la contrae la trasmetterà ad una media di altre 1.30 persone. La lettura chiave che viene data a questo parametro è la seguente:

- $R_0 < 1$: Se $R_0$ è minore di 1 ogni infezione esistente provoca meno di una nuova infezione. In questo caso, la malattia diminuirà e alla fine si estinguerà
- $R_0 = 1$: Se $R_0$ è uguale a 1, ogni infezione esistente provoca una nuova infezione. La malattia rimarrà viva e stabile, ma non ci sarà un focolaio o un'epidemia
- $R_0 > 1$: Se $R_0$ è maggiore di 1, ogni infezione esistente provoca più di una nuova infezione. La malattia si diffonderà tra le persone e potrebbe esserci un focolaio o un'epidemia.

[4] $R_0$ della comune influenza

Ripetendolo un ulteriore volta, $R_0$ descrive la trasmissione in una popolazione completamente suscettibile. Ne deriva che per ottenere una stima bisogna studiare la malattia nella sua fase epidemica iniziale ove i contagi non vengono ridotti dalla quarantena. È importante sottolineare questa considerazione perchè i vari stimatori di $R_0$ possono essere utilizzati soltanto in questo caso, ovvero se nel campione di riferimento in cui si calcola l'indice:

- Nessuno è vaccinato
- Nessuno è immune alla malattia
- Non sono state prese misure di ordine pubblico per diminuire i contagi

Il motivo è presto spiegato, anche se una sola condizione non viene rispettata $R_0$ viene sottostimato perchè gli individui infetti non riuscirebbero a trasmette il virus a parte della popolazione che in realtà, in mancanza di suddette ipotesi, si sarebbe infettata.

### 2.2.1 Alcune delucidazioni su $R_0$ e differenze con $R_t$

Leader mondiali e sanità pubblica hanno speso gli ultimi mesi e spenderanno i prossimi a parlare di questo $R_0$. Può sembrare un concetto semplice ma in realtà è molto più complesso di quanto spiegato nel paragrafo precedente. È il risultato di metodi di stima molto complessi, e può cambiare radicalmente da un luogo ad un altro in base alle condizioni del territorio e ai comportamenti delle persone. Il parametro nasce da studi demografici dove è usato per descrivere il tasso di nascite[5]. $R$ si riferisce alla riproduzione e 0 alla generazione $0esima$ (paziente zero).
Ripetendo l'esempio in precedenza se si suppone un valore del parametro $R_0 = 1.30$, se 1000 persone fossero infette ci si aspetterà di avere questo tipo di andamento di contagi nel tempo:

$$t_1 \; = \; 1300 \rightarrow t_2 \; = \; 1690 \rightarrow ..... \rightarrow t_{10} \; = \; 42621$$

Il metodo più utilizzato per stimare il parametro $R_0$ è una stima tramite tasso di crescita esponenziale ma ad oggi un consenso su come stimare il parametro non è presente. Per questo motivo c'è la necessità di associare oltre che ad una stima puntuale, un intervallo di confidenza (IC) in modo da avere una misura di incertezza sul valore. Molte delle volta infatti i vari intervalli di confidenza sono molto ampi. La

[5]si userà tale esempio per trovare l'equazione di Lokta-Eulero

SARS ad esempio, è di solito descritta con un $\hat{R}_0$ da 2 a 5. Il morbillo addirittura ha un IC che varia da 3 a 203, un enorme differenza. È evidente che gli scienziati stanno ancora discutendo e rivedendo in che modo possono essere trovare stime sempre più precise.

Il valore di $R_0$ può cambiare a seguito di modifiche nei contatti sociali (per esempio a seguito di interventi di distanziamento sociale) oppure a seguito della riduzione del numero di persone suscettibili. Viene allora solitamente indicato con $R_t$. Tale valore è la stima che viene comunicata giornalmente. La differenza è grossolana e non bisogna farne confusioni. Un andamento di $R_0$ nel tempo, che prende poi il nome di $R_t$ serve ai governi per valutare l'effetto dei decreti che vengono emessi e può essere usato come parametro variabile all'interno di modelli statistici per previsioni future circa gli infetti. La valutazione della malattia viene espressa invece dalla sola stima del parametro $R_0$ che viene sempre effettuata solamente nella fase iniziale di una epidemia per i motivi spiegati in precedenza.

A tal proposito è bene sottolineare che se $R_t$ è più piccolo di 1 non significa che il virus è sconfitto ma supponendo che i numeri siano corretti, che la sua diffusione è stata messa in pausa. In questo caso avendo un valore al di sotto della soglia, significa che ogni, per esempio, 100 malati si infetteranno meno di altri 100 individui. Ogni generazione successiva di infezioni sarà più piccola dell'ultima ma le persone potranno ancora ammalarsi e morire.

Per capire la rapidità della diffusione di una malattia, e per utilizzare gli stimatori di $R_0$ è necessario anche, oltre a questo parametro, la costruzione di un intervallo seriale (noto come *Generation-Time*) per costruire l'andamento delle diverse generazioni di casi nel tempo.

## 2.3 Generation-Time

Il tempo di generazione è definito come l'intervallo di tempo medio tra le infezione di un caso primario e l'infezione di un caso secondario causata dal caso primario.
La comprensione degli intervalli di tempo tra le generazioni successive tra individui infetti è fondamentale per quantificare in modo appropriato la dinamica di trasmissione delle malattie infettive.

Sebbene sia difficile osservare direttamente l'intervallo di tempo, in molti si occupano di derivarne andamenti il più precisi possibili per derivarne distribuzioni che giocano un ruolo chiave nella stima di trasmissione di una malattia (come spiegato

misurata da $R_0$ ). Il tempo di generazione è un periodo non osservabile[6] che di solito viene chiamato anche con il termine *"intervallo seriale"*, relativamente usati per distinguere la durata dall'infezione primaria a quella secondaria misurata in base ai tempi dell'infezione e all'insorgenza dei sintomi. Sia l'intervallo seriale che il tempo di generazione hanno però lo stesso valore medio e per questo ai fini di questo lavoro rappresentano la medesima definizione, sebben come detto in qualche riga precedente hanno concezioni differenti.

In questo paragrafo si illustrano i risultati riguardanti le stime del generation-time per il coronavirus fornite da *International Society for infectious diseases*[8]. Il lavoro svolto per la ricerca della distribuzione si basa sull'osservazione di un campione di 28 coppie formate da infettore-infetto. La derivazione dei risultati ha implicato l'utilizzo di tecniche statistiche complesse che hanno messo a confronto tre tipologie di variabili casuali: *log-normale, gamma e Weibull*

Il risultato che ne è derivato è che il miglior modello per descrivere i vari intervalli di diffusione del virus è stato quello costruito con la log-normale con mediana dell'intervallo stimata attorno a 4.0 giorni con un intervallo di confidenza compreso tra $[3.1, 4.9]$. I parametri stimati della variabili casuale sono risultati:

$$\mu = 4.7 \rightarrow IC[3.7, 6.0]$$
$$\sigma = 2.9 \rightarrow IC[1.9, 4.9]$$

Si ricorda che $f(x) \sim log - normale(\mu, \sigma^2)$ ha funzione di densità:

$$f(x) \; = \; \frac{e^{-\frac{(logx-\mu)^2}{2\sigma^2}}}{x\sqrt{2\pi}\sigma} \; per \; x \; > \; 0 \tag{2.1}$$

La stima della mediana dell'intervallo seriale di 4,0 giorni indica che l'infezione da Covid-19 porta a rapidi cicli di trasmissione da una generazione di casi all'altra. Questo vuol dire che la sua diffusione è rapida e veloce. Se si confrontano i risultati con quelli della SARS (immagine 2.1 ove sull'asse delle $x$ si ha l'andamento temporale e sull'asse delle $y$ si ha la densità di probabilità) l'intervallo di trasmissione medio del coronavirus risulta molto più breve, il che è indice di una pericolosità di diffusione maggiore.

[6] poiché il momento preciso dell'infezione è molto difficile e quasi impossibile da rilevare, il tempo di generazione non è osservabile per due ospiti successivi

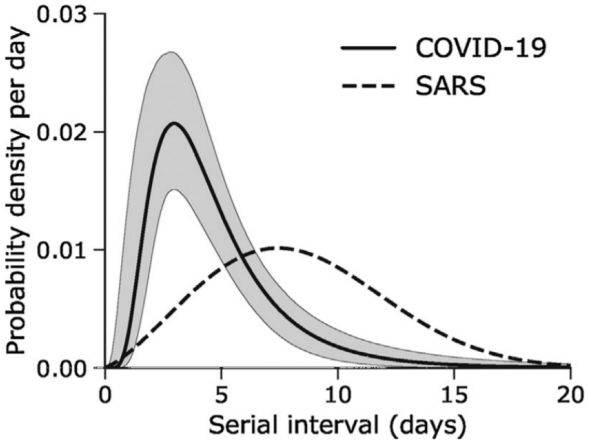


Figura 2.1: Intervallo Seriale Sars vs Covid-19

Una cosa ancora più importante è che la mediana dell'intervallo stimato è più breve delle stime sul periodo medio di incubazione (circa 5 giorni). Questo significa che è probabile che ci sia una trasmissione pre-sintomatica che può verificarsi più frequentemente della trasmissione sintomatica. Il risultato che ne è derivato in sostanza indica che molte trasmissioni non possono essere impedite solo mediante l'isolamento di casi sintomatici, poiché quando i contatti vengono rintracciati potrebbero già aver infettato e generato casi secondari e cosi via.

## 2.4 Stima di $R_0$ tramite il tasso di crescita esponenziale

### 2.4.1 Introduzione

Una stima accurata del valore di $R_0$ è cruciale per capire l'espansione di una malattia. Per le nuove malattie come lo stesso Covid-19, le informazioni disponibili sulla trasmissibilità sono purtroppo limitate al conteggio giornaliero dei nuovi casi. È noto come questi conteggi aumentano quasi in modo esponenziale nella fase iniziale di un'epidemia. Tra i metodi più utilizzati per il calcolo del parametro ci sono due correnti di pensiero differenti per la stima. La prima utilizza il tasso di crescita esponenziale, $r$, che è definito come il rapporto dei casi giornalieri nel tempo rispetto al giorno precedente. Il valore osservato di $r$ viene quindi legato al valore $R_0$ attraverso una semplice equazione lineare[9]:

$$R_0 = 1 + rT_c$$

Ove $T_c$ è il valore atteso della variabile casuale log-normale che definisce il generation-time (equazione 2.1).

Demografi, ecologi e biologi evoluzionisti invece adottano un approccio leggermente diverso, più matematico. Derivano $R_0$ secondo la cosiddetta equazione di **Lotka-Eulero**. In sostanza si semplifica questa equazione ignorando la variabilità del tempo di generazione ove per variabilità si intende la possibile modifica dei tempi di infezione fra due individui[7]e anche lo spazio dei valori che l'intervallo di tempo medio può assumere sulla base dell'intervallo di confidenza.
Il risultato è un'equazione esponenziale:

$$R_0 = e^{rT_c}$$

Qui, $T_c$ è un analogo demografico della media del generation-time epidemiologico.

Avendo due equazioni alternative per mettere in relazione il valore desiderato si affronta la difficoltà di scegliere quello più appropriato. Come è facile capire a seconda della scelta si avranno due risultati totalmente differenti.

Quello che viene fatto nella realtà è una riformulazione dell'equazione di Lotka-Eulero, che viene poi utilizzata da epidemiologici e statistici per la stima del parametro $R_0$.

### 2.4.2 L'equazione di Lotka-Eulero

Si deriva l'equazione di Lotka-Eulero tramite degli esempi semplificativi usando la popolazione umana. Per semplicità, ci si concenta su individui di sesso femminile, supponendo che vi sia sempre una scorta sufficiente di maschi per garantirne la riproduzione. Si misura il tempo e l'età in anni e ci si riferirà al tempo presente come $t = 0$, in modo che gli eventi nel passato si siano verificati quando il tempo t è negativo e gli eventi nel futuro si verificheranno quando il tempo t è positivo. Si parte dal presupposto che la popolazione mostra una crescita esponenziale ad un tasso fisso e che la distribuzione per età della popolazione non cambi nel tempo.

L'equazione di Lotka-Eulero può essere intesa come la combinazione di due concetti che possono essere spiegati in modo molto intuitivo.

[7]è facile pensare in ambito di un virus che questo potendo mutare può variare le tempistiche di contagio tra individui

1- Se si sommano il numero di bambini nati da madri di tutte le età in un determinato momento, si ottiene il numero totale di nascite in quel momento. Il numero di nascite per madri di età $a$ al tempo $t$ è uguale al numero di nascite al momento $t - a$ (il numero di madri, comprese quelle che non sono sopravvissute) moltiplicato per il numero atteso di figli all'anno per le madri di età $a$. Sommando queste nascite in tutte le possibili età delle madri, si ottiene il numero totale di nascite nell'anno $t$ ovvero:

$$b(t) \;=\; \int_{a=0}^{\infty} b(t-a)n(a)\,\mathrm{d}a \tag{2.2}$$

dove $b(t)$ si riferisce al tasso di natalità della popolazione al momento $t$ e $n(a)$ si riferisce al tasso di produzione della prole femminile di una madre all'età $a$.

2- In secondo luogo, perché la popolazione sia in crescita esponenziale con una distribuzione per età stabile, il numero di nascite in un dato momento (ad esempio $t$) è uguale al numero delle nascite di un tempo fa, moltiplicato per la crescita esponenziale della popolazione. Si può quindi dedurne la seguente uguaglianza:

$$b(t) \;=\; b(t-a)e^{ra} \tag{2.3}$$

A questo punto si combinano le espressioni (2.3) e (2.4) per ottere un espressione con $b(t)$ su entrambi i lati:

$$b(t) \;=\; \int_{a=0}^{\infty} b(t)e^{-ra}n(a)\,\mathrm{d}a \tag{2.4}$$

Il parametro $n(a)$ è più familiare nella demografia come il prodotto delle funzioni di sopravvivenza al parto e fecondità, $n(a) = l(a)m(a)$ . Usando questa parametrizzazione più familiare e rimuovendo $b(t)$ da entrambi i lati dell'equazione, si ottiene la cosidetta equazione di *Lotka-Eulero:*

$$\textit{Con quindi} \;:\; n(a) = l(a)m(a)$$

$$1 \;=\; \int_{a=0}^{\infty} e^{-ra}l(a)m(a)\,\mathrm{d}a \tag{2.5}$$

### 2.4.3 Derivazione di $R_0$ dall'equazione di Lotka-Eulero

Mentre l'equazione Lotka-Eulero è una parte fondamentale della demografia, con cui si potrebbe essere interessati a derivare i tassi di crescita della popolazione, il problema qui rilevante è stimare il numero riproduttivo, $R_0$, dai tassi di crescita di una malattia. In precedenza si è definito $n(a)$ come il tasso di produzione della prole femminile da parte di una madre all'età $a$. Risulta che se si integra $n(a)$ per tutta la durata della vita di una madre, si ottiene il suo numero totale di figli. Tale numero altro non è che $R_0$:

$$R_0 \;=\; \int_0^\infty n(a)\,\mathrm{d}a \tag{2.6}$$

Il tasso $n(a)$ può essere normalizzato a una distribuzione $g(a)$, che in questo caso rappresenta l'età in cui l'individuo è in gravidanza:

$$g(a) \;=\; \frac{n(a)}{\int_0^\infty n(a)\,\mathrm{d}a} \;=\; \frac{n(a)}{R_0} \tag{2.7}$$

Ora si considera l'età come il tempo trascorso per prendersi l'infezione, e quindi nella notazione sopra la distribuzione del *generation-time* è equivalente a $g(a)$ . Sostituendo questa espressione con la distribuzione dell'intervallo seriale $g(a)$ nell'equazione di Lotka-Eulero (2.5), si ottiene saltando passaggi algebrici elementari:

$$\frac{1}{R_0} \;=\; \int_{a=0}^\infty e^{-ra} g(a)\,\mathrm{d}a \tag{2.8}$$

Il termine che ora appare nella parte destra di questa equazione è familiare ai matematici come la cosiddetta trasformata di Laplace della funzione $g(a)$. Più specificamente nota agli statistici come la funzione generatrice dei momenti $M(z)$ della distribuzione $g(a)$. Scritta formalmente è definita come:

$$M(z) \;=\; \int_{a=0}^\infty e^{zt} g(a)\,\mathrm{d}a \tag{2.9}$$

Si usa quindi la funzione generatrice dei momenti per semplificare la notazione dell'equazione riformata di Lotka-Eulero (2.8). Qui, l'argomento z prende il valore del tasso di crescita definito in precedenza ma ovviamente cambiato di segno $(-r)$:

$$R_0 = \frac{1}{M(-r)} \tag{2.10}$$

Inutile dire che per trovare un valore di $R_0$, $M(-r)$ deve esistere! Si ricorda che $M(z)$ funzione generatrici dei momenti essendo definita da una funzione integranda può accadere, infatti, che l'integrale sia divergente.

Una funzione generatrice di momenti (se esiste) caratterizza in modo univoco la forma dell'intera distribuzione di probabilità: $M(z)$ determina $g(a)$ e, al contrario, $g(a)$ determina $M(z)$. Il corollario di questo momento che genera l'espressione della funzione è quindi una relazione tra il tasso di crescita $r$ e $R_0$ che caratterizza in modo univoco la forma della distribuzione del generation-time e, al contrario, la forma della distribuzione del generation-time determina la relazione tra il numero riproduttivo e il tasso di crescita.

Per le malattie con un intervallo medio $T_c$ e una deviazione standard della distribuzione $\sigma$, come in questo caso, gli intervalli seriali possono approssimarsi ad una distribuzione normale. Supponendo un intervallo di generazione normalmente distribuito si ottiene la seguente relazione tra il tasso di crescita $r$ e il numero riproduttivo $R_0$:

$$R_0 \; = \; e^{rT_c - \frac{1}{2}r^2\sigma^2} \tag{2.11}$$

Questa relazione è una curva convessa che approssima una curva esponenziale. Si evince anche che una distribuzione più concentrata attorno alla media della distribuzione generation-time, cioè con un valore più basso per $\sigma$, porta a valori più alti per il basic reproduction number.

A questo punto l'ultimo step è quello di crearsi un intervallo di confidenza per il valore di $R_0$ per l'esigenza di associare alla stima puntuale una misura di incertezza. Quello che si fa è utilizzare una regressione di poisson sulla 2.10 per crearsi degli intervalli.

### 2.4.4 $R_0$ utilizzando campione di osservazioni cinesi e italiane

A questo punto ci si serve della curva epidemica ricavata nel capitolo 1, sia per quanto riguarda quello italiano sia quello cinese, e dell'intervallo seriale, per stimare il parametro per entrambe le nazioni.

La prima cosa da fare è scegliere un periodo temporale su cui ricavare le stime. Si è scelto di considerare in entrambi le nazioni i primi 15 giorni di propagazione del virus. La scelta non è stata casuale ma è dovuta al fatto che nei dati cinesi al giorno

20 circa si ha un picco inusuale di casi infetti dovuti all'inizio di tamponamenti del tutto casuali (prima era effettuato solo alle persone infette). Seppur si è pensato di considere come valore anomalo tale osservazione, non avrebbe avuto comunque senso allungare l'intervallo in quanto le disposizioni poi emesse dal governo cinese hanno contribuito ad abbassare il numero di contagi e come ampiamente ripetuto c'è bisogno di un andamento esponenziale, senza restrizioni, per aver il risultato il più preciso possibile. Sulla base di queste premesse si è scelto, per poter fare poi effettuare un confronto, di considerare lo stesso periodo di tempo di 15 giorni anche per l'Italia.

In particolare per questa comparazione si sono considerati solo i dati di *Hubei*, provincia cinese che comprende Wuhan che è la provincia dove si è trasmesso maggiormente il virus. Questa scelta per avere una reale percezione dell'andamento dei contagi.

Si calcolano tramite (2.11) i valori di $R_0$ ottendendo i seguenti risultati:

$$R_0^{Cina} \ : \ 3.39[3.36, 3.42] \ ^{8}$$
$$R_0^{Italia} \ : \ 3.60[3.53, 3.67] \ ^{9}$$

**Considerazioni:**

- Sembrerebbe che l'$R_0$ italiano sia leggermente più alto, ma non si può affermare con certezza il risultato. Si deve però considerare che in Italia le misure attuate dal governo per contenere la diffusione sono state meno severe rispetto a quanto avvenuto in cina e forse questa è una possibile causa del risultato ottenuto

- Si può affermare con certezza invece che la stima del parametro trovato di $R_0$ è in realtà sottostimato ma questo è facilmente intuibile. In sostanza i dati utilizzati (chiaramente anche da OMS[10]e ricercatori) non sono veritieri perchè non comprendono tutte le persone che in un certo periodo hanno realmente riscontrato l'infezione ma soltanto quelle risultate positive ai tamponi. Non potendo fare un controllo totale della popolazioni gli unici dati a disposizione sono questi. E' complicato anche lo studio di un sottocampione per poter poi applicare il risultato alla popolazione intera perchè presenterebbe molti problemi che derivano dalla non conoscenza del Covid-19. Senza precise informazioni uno studio del genere è molto complesso e difficile

[8]periodo di tempo considerato dal 23/01/2020 al 07/02/2020

[9]periodo di tempo considerato dal 22/02/2020 al 08/03/2020:

[10]organizzazione mondiale della sanità

- Ricordando che si è utilizzato come generation-time, una log-normale (equazione 2.1), con parametri fissi stimati si fa notare che facendoli variari secondo i valori degli intervalli di confidenza la stima di $R_0$ varia

Sebben questo metodo di stima sia molto utilizzato per la sua facilità di solito si utilizza anche qualcosa di diverso. Il metodo è quello della funzione di massima verosimiglianza, che in ambito statistico è un approccio classico per la stima dei parametri.

## 2.5 Stima di $R_0$ tramite la funzione di massima verosimiglianza

### 2.5.1 Derivazione della stima di massima verosimiglianza

Si utilizza ora un metodo basato sulla verosimiglianza per stimare $R_0$. Si userà ancora l'intervallo seriale (equazione 2.1) trovato in precedenza. Il metodo in questione prende il nome dai loro inventori White e Pagano[10]. Ci sono due tipi di massimizzazioni, la prima considera l'intervallo seriale noto, mentre il secondo si basa sulla non conoscenza dell'intervallo che viene stimato. Si userà il primo in quanto il generation-time è *"noto"* ma per la sua derivazione serve una spiegazione anche del metodo che stima entrambi.

Prima di illustrare la parte matematico statistica alla base dello stimatore si procede a elencare le sue ipotesi:

- Non vi siano casi importati da altri paesi nella popolazione di riferimento
- Non vi siano dati mancanti
- La popolazione si mescoli uniformente
- La popolazione deve essere chiusa (no nascite e morti)
- Nei casi primari compaiono i sintomi prima dei loro casi secondari (solo per il metodo che stima anche l'intervallo seriale)

Per non complicare troppo le cose si supponga che le condizioni vengano rispettate. Come ben noto chiaramente non tutte le ipotesi sono vere nell'ambito italiano, mentre risultano in parte vere per la Cina se si considera solo il primo periodo di manifestazione del virus. In Italia infatti:

- Il pazienze zero sembra derivare dalla Germania
- Non è verificata l'attendibilità dei dati
- La popolazione è sotto vincoli governativi come la quarantena imposti dai primi giorni

I dati che utilizza lo stimatore non sono altro che i conteggi dei nuovi casi giornalieri (curva epidemica), definiti da un vettore $N = \{N_t\}$ e $t = 0, ..., T$ con t che indica una certa unità di tempo e $N_t$ il numero di nuovi casi all'istante di tempo t . Senza perdita di generalità, si assume che t indicizza il periodo temporale definito dai giorni[11].
Di norma all'inizio di un'epidemia l'intervallo seriale non è ben definito perchè non ci sono osservazioni a sufficienza per poterne ricavare stime. Il primo metodo proposto si basa quindi su una stima diretta sia del numero riproduttivo sia dell'intervallo seriale.

Come possibile modello, si ipotizzi che il numero di casi secondari prodotti da un individuo infetto segua una distribuzione di Poisson, con valore atteso $R_0$, e che l'intervallo seriale sia descritto da una distribuzione multinomiale con vettore di probabilità $p_1, ..., p_k$, con chiaramente $p_1 + ... + p_k = 1$. In sostanza si cerca di modellare le infezioni da casi primari a secondari, da secondari a terziari e cosi via secondo delle probabilità che rappresentano il "successo" di infezione nei casi successivi. L'assunzione di una distribuzione di questo tipo implica che dopo un certo periodo di tempo, definito da un certo $k$, la probabilità di generare un caso secondario sia trascurabile.
All'interno del focolaio la situazione viene descritta nel seguente modo. $N_0$ sono gli individui che inizialmente manifestano l'epidemia. Ognuno di questi casi genera indipendentemente casi secondari secondo una distribuzione di Poisson con media $R_0$. Si definisce poi $X_0$ come il numero totale di casi prodotti dai casi $N_0$ iniziali (il che significa $X_0 \sim Pois(N_0 R_0)$). All'interno del focolaio questi casi $X_0$ infettano altri individui nei k giorni successivi secondo una variabile casuale multinomiale (intervallo seriale definito in precedenza in questo modo).

[11] tuttavia, questo metodo è applicabile a qualsiasi unità di tempo discreta

La notazione è quindi la seguente:

- $N_i$ rappresenta il numero totale di casi il giorno i
- $X_{ij}$ rappresenta il numero di casi presenti nel giorno j, che sono stati generati dai casi $N_i$
- $X_i$ indica il numero totale di casi prodotti dai casi primari nel giorno i (ovvero $X_i = \sum_j X_{i,j}$)

Se k, la lunghezza massima dell'intervallo seriale, fosse per esempio tre, allora si può illustrare il tutto con il seguente schema:

$$\begin{array}{lccccc} N_0 & & & & & \\ N_1 = X_{01} & & & & & \\ N_2 = X_{02} & +X_{12} & & & & \\ N_3 = X_{03} & +X_{13} & +X_{23} & & & \\ N_4 = & X_{14} & +X_{24} & +X_{34} & & \\ N_5 = & & X_{25} & +X_{35} & +X_{45} & \\ \vdots & & & \vdots & \vdots & \end{array}$$

Si noti che questi schemi non forniscono alcuna indicazione del momento in cui si è verificato il passaggio dell'infezione tra l'individuo primario e quello secondario, ma rappresentano solo il momento in cui i casi diventano sintomatici. È da notare che non viene osservato $X_{ij}$. Se si fosse a conoscenza di tale informazione si potrebbe facilmente stimare $R_0$ e il vettore di probabilità $\mathbf{p}$ ($p = \{p_1, p_2, \ldots, p_\kappa\}$ che rappresenta la distribuzione dell'intervallo seriale), tramite massima verosimiglianza. La sua costruzione per la derivazione delle stime risulterebe questa:

$$\begin{aligned} L(R_0, \mathbf{p}|\mathbf{N}, \mathbf{X}) = &\left[\frac{\mathrm{e}^{-N_0 R_0}(N_0 R_0)^{X_0}}{X_0!}\right]\left[\begin{pmatrix} X_0 \\ X_{01}\cdots X_{0,1+k}\end{pmatrix} p_1^{X_{01}}\cdots p_k^{X_{0,k}}\right]\times \\ &\left[\frac{\mathrm{e}^{-N_1 R_0}(N_1 R_0)^{X_1}}{X_1!}\right]\left[\begin{pmatrix} X_1 \\ X_{12}\cdots X_{1,1+k}\end{pmatrix} p_1^{X_{12}} p_k^{X_{1,1+k}}\right]\cdots \\ \times &\left[\frac{\mathrm{e}^{-N_T R_0}(N_T R_0)^{X_T}}{X_T!}\right]\left[\begin{pmatrix} X_T \\ X_{T,T+1}\cdots X_{T,T+k}\end{pmatrix} p_1^{X_{T,T+1}}\cdots p_k^{X_{T,T+k}}\right] \end{aligned} \quad (2.12)$$

Questa configurazione presuppone l'indipendenza degli eventi di trasmissione. Ball e Donnelly[11] forniscono una prova che indica che l'indipendenza degli eventi si interrompe quando viene infettata circa la radice quadrata della popolazione sensibile, il che non risulta un problema ai fini del calcolo. È bene ricordare infatti, che la stima

di $R_0$ deve essere effettuata nel primissimo periodo di un infezione.
Riorganizzando i termini la verosimiglianza si riduce ad una Poisson:

$$L(R_0, \mathbf{p}) = \prod_{t=1}^{T} \frac{\mathrm{e}^{-\mu_t} \mu_t^{N_t}}{N_t!} \tag{2.13}$$

dove $\mu_t = R_0 \sum_{j=1}^{\min(k,t)} N_{t-j} p_j$.
A questo punto si può procedere tramite semplici tecniche di massima verosimiglianza per stimare $R_0$ e $p_j$.

Dato che in questo caso di studio l'intervallo seriale è gia noto non c'è bisogno della costruzione di una massima verosimiglianza che vada a stimare il parametro $\mathbf{p}$. L'analisi in questo lavoro svolta infatti è stata eseguita dopo che si sono presentati molti focolai nel mondo ed è quindi stato possibile tenere tracce tra i contatti tra casi primari e secondari per la stima dell'intervallo seriale con una certa precisione (errore misurato dagli intervalli di confidenza). In questi caso, l'interesse si concentra quindi sulla sola stima di $R_0$.
Di seguito, si descrive lo stimatore per $R_0$ che risulta molto semplice da implementare, infatti può essere derivato dalla massima verosimiglianza (MLE) 2.13.

Definendo il termine comunemente chiamato score (punteggio) come il gradiente, cioè il vettore delle derivate parziali del logaritmo della funzione di verosimiglianza, ne risulta lo score dalla 2.13 la seguente equazione:

$$U_{R_0}(T) = \sum_{t=1}^{T} \frac{(N_t - \mu_t)}{R_0} \tag{2.14}$$

Ove $\mu_t = R_0 \sum_{j=1}^{\min(k,t)} N_{t-j} p_j$.
Uguagliandolo a zero e risolvendo per $R_0$ si ottiene il seguente stimatore di massima verosimiglianza:

$$R_0 = \frac{\sum_{t=1}^{T} N_t}{\sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} p_j N_{t-j}} \tag{2.15}$$

Tramite questo si può procedere al calcolo del valore di $R_0$.

### 2.5.2 $R_0$ utilizzando campione di osservazioni cinesi e italiane

Senza ripetere quanto detto all'inizio del paragrafo 2.4.4 perchè il procedimento e le considerazioni effettuate sono le medesime si osservano direttamente i risultati, ottenuti questa volta tramite massimizzazione di $R_0$ nell'equazione 2.13 (in sostanza si utilizza lo stimatore 2.15):

$$R_0^{Cina} \ : \ 3.00[2.98, 3.03] \ ^{12}$$
$$R_0^{Italia} \ : \ 3.10[3.04, 3.17] \ ^{13}$$

**Considerazioni:**

- Il metodo usato precedentemente sembra che sovrastimi i parametri
- Anche in questo caso ma come ci si poteva aspettare, l'$R_0$ italiano appare leggermente superiore a quello cinese e come spiegato in precedenza è plausibile che sia "vero". Essendo poi davvero vicini i due valori e essendo state prese misure diverse (meno restrittive) il risultato è del tutto possibile
- Anche in questo caso si ricorda che si è utilizzato come generation-time, una log-normale (equazione 2.1), con parametri fissi e quindi facendoli variare secondo gli IC la distribuzione cambia e di conseguenza $R_0$ varia

Un'ultima analisi è volta a capire se i numeri trovati sono significativi. Per farlo basta guardare le stime di $R_0$ svolte da OMS e altri ricercatori. Facendolo risulta che la stime ufficiali hanno un intervallo di confidenza con questi valori[12]:

$$R_0 \ = \ [2, 4] \tag{2.16}$$

L'incertezza su questi valori rimane assoluta come ben spiegato in precenza. Il risultato trovato ci indica $R_0$ medio per tale malattia e trasmette un'idea sulla forza di diffusione del coronavirus.
Questo risultato ci indica la pericolosità della malattia. La stima trovata non indica alta pericolosità che invece viene espressa dalla sua velocità di diffusione che può portare a sua volta al collasso delle strutture ospedaliere. È meglio pensare a questo numero ottenuto come un punto di partenza per il comportamento del virus in assenza di fattori umani o ambientali del mondo reale. Per interpretare e capire come nel tempo varia questo parametro in base ai decreti emessi, si utilizza $R_t$. I governi infatti utilizzano stime di $R_0$ giornaliere per capire come il virus si stia diffondendo. Tale stime sono, per quanto sofisticate, altamente teoriche. È una stima basata su altre stime, e quindi i margini d'errore sono ancora più ampi. Tuttavia, è una delle uniche metriche per capire che tipo di restrizioni sulla popolazione applicare.

[12] periodo di tempo considerato dal 23/01/2020 al 07/02/2020

[13] periodo di tempo considerato dal 22/02/2020 al 08/03/2020:

# Capitolo 3

# I modelli epidemiologici

## 3.1 Modello $SIR$

### 3.1.1 In termini teorici

Il modello SIR è stato utilizzato per la prima volta da Kermack e McKendrick ed è stato successivamente applicato a una varietà di malattie, in particolare a malattie dell'infanzia con immunità permanente al recupero, come morbillo, parotite, rosolia e pertosse. S, I e R rappresentano il numero di individui sensibili, infetti e recuperati e $N = S + I + R$ è la popolazione totale.
Il modello SIR viene utilizzato laddove gli individui si infettano direttamente (quindi non attraverso un vettore di malattia come ad esempio un animale) e il virus non manifesta periodo di incubazione. Un individuo che guarisce viene considerato immune, ovvero non può più essere infettabile. Il contatto tra le persone è modellato per essere del tutto casuale.

Il tasso di infezione delle persone è proporzionale al numero di persone infette e al numero di persone suscettibili. Se ci sono molte persone infette, le probabilità che una persona sensibile venga a contatto con qualcuno che è infetto sono alte. Allo stesso modo, se ci sono pochissime persone suscettibili, le probabilità che una persona sensibile entri in contatto con un infetto sono inferiori (poiché la maggior parte del contatto sarebbe tra le persone non sensibili-infette).
Il modello SIR può essere riassunto tramite il seguente grafico.

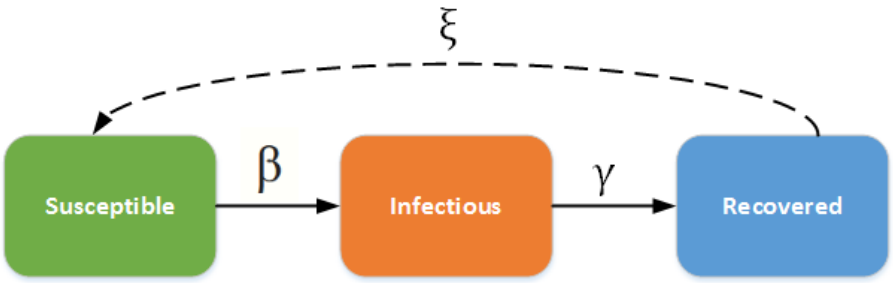


Figura 3.1: Modello SIR

Dove $\beta$ rappresenta il tasso infettivo e controlla il tasso di diffusione. Rappresenta la probabilità di trasmissione della malattia da un soggetto infetto ad un individuo appartenente alla categoria dei suscettibili. Il tasso di recupero è determinato dalla durata media, $D$, dell'infezione. Per il modello SIRS, una tipologia differente del modello SIR, $\xi$ è la velocità con cui gli individui guariti, ritornano alla stato sensibile a causa della perdita dell'immunità. Nel caso del coronavirus l'immunità è quasi permanente (si spiegherà meglio nel prossimo modello) e quindi questo aspetto non verrà affrontato.

Nella sua forma deterministica, il modello $SIR$ può essere scritto tramite le seguenti equazioni differenziali:

- **Suscettibili:**
$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{3.1}$$
- **Infetti:**
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \tag{3.2}$$
- **Guariti:**
$$\frac{dR}{dt} = \gamma I \tag{3.3}$$

Questo sistema non è lineare, tuttavia è possibile derivarne la soluzione analitica in forma implicita[1].

[1]alcuni strumenti numerici includono i metodi Monte Carlo

Si osserva che:

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$$

da cui ne segue che

$$S(t) + I(t) + R(t) \;=\; N \;(\forall\, t)$$

Il modello, come facilmente intuibile, è molto limitato. Le ipotesi alla sua base sono molte come spiegato in precedenza, tra cui l'assunzione che la popolazione in esame deve essere in uno stato chiuso. Questo significa che, oltre a non dover esserci nessun entrata/uscita di individui dalla popolazione di riferimento, non ci devono essere ne morti ne nascite durante il periodo di studio. In una popolazione chiusa senza dinamiche vitali, un'epidemia alla fine si estinguerà a causa di un numero insufficiente di soggetti sensibili per sostenere l'andamento delle infezioni. Ogni persona appartenente al compartimento dei suscettibili verrà sicuramente infettata in una popolazione chiusa. Le persone infette appartenenti a intervalli di generazioni successivi non inizieranno un'altra epidemia a causa dell'immunità permanente della popolazione esistente.

Un esempio semplificativo per la spiegazione di tale fenomeno può essere trovata immaginando un contenitore (paese di riferimento) di particelle[2] (individui). Essendo lo spazio limitato e le particille in continuo movimento, è inevitabile che tutte entreranno in contatto fra di loro o in maniera diretta o per conto di terzi.

Le dinamiche di un'epidemia però, come ad esempio l'influenza, sono spesso molto più veloci delle dinamiche di nascita e morte, quindi queste sono spesso omesse in semplici modelli compartimentali, in quanto non significative. Il caso del Covid-19 può essere trattato come una semplice influenza, anche se il tasso di mortalità è leggermente superiore, e quindi questa omissione di dinamicità non ne causa grossi errori.

Il modello SIR è un modello statistico fondamentale in quanto rimodellando le equazioni differenziali si possono determinare altre tipologie di modelli, più complicati, come ad esempio il modello SEIR che tiene conto del periodo di incubazione.

### 3.1.2 Relazione $R_0$ con il numero di infetti

Nel capitolo precedente si è trovata una stima di $R_0$ fondamentale per l'applicazione di questo tipo di modelli. Nel precedente paragrafo nelle equazioni differenziali non essendo presente quest'ultimo non è stato chiarito il motivo della sua importanza all'interno del modello.

[2] si assuma che abbiano carica neutra per evitare attrazioni e repulsioni

La prima cosa da fare è quella di prendere l'equazione differenziale 3.2, che descrive il numero di infetti nel tempo, e di porla maggiore a zero. In questo modo quello che ne deriva è che un'epidemia avrà luogo se e soltanto se saranno presenti infetti al tempo $t$:

$$\frac{\beta SI}{N} - \gamma I > 0$$

che con banali operazioni algebriche diventa:

$$\frac{\beta SI}{N} > \gamma I$$

Ora $\frac{S}{N} = 1$ in quanto all'inizio dell'epidemia $S \approx N$ e $I$ si semplifica risultando:

$$\frac{\beta}{\gamma} > 1$$

Ponendo poi una semplice condizione la relazione è trovata.

$$R_0 = \frac{\beta}{\gamma} \tag{3.4}$$

Si fa notare che come descritto al capitolo due abbiamo trovato ora, tramite semplici passaggi algebri, che un epidemia si svilupperà (ipotesi iniziale ponendo infetti $> 0$) solo e soltanto se $R_0 > 1$ .

### 3.1.3 Applicazione modello SIR in Lombardia

Sebben il modello SIR non consideri un tempo di incubazione per la malattia, che in questo caso è presente, è utile vedere una sua applicazione per capire la portata dell'infezione. Ci si renderà così conto di come lo scenario avrebbe potuto modificarsi se il tempo d'incubazione del coronavirus fosse stato relativemente piccolo o inesistente. Tralasciando quindi decessi e nascite si possono definire i parametri iniziali del modello ricordando che $\hat{R}_0$ in Italia è risultato pari a 3.12 e 3.60. In questo caso essendo il parametro molto probabilmente una sottostima si è scelto di prendere il valore massimo assumibile dall'intervallo di confidenza trovato nel precedente capitolo. I parametri quindi necessari all'interno del modello sono i seguenti:

- $R_0 = 4$
- $\gamma$ che è data da $\frac{1}{D}$ ove $D$ numero di giorni della durata media dell'infezione dalla comparsa dei sintomi, per il coronavirus pari a 10[13]

- $\beta$ è facilmente calcolabile tramite relazione 3.4
- *Parametri iniziali:* in questo si è scelto di inizializzare il modello con i seguenti parametri:

  $\rightarrow S = N - 1$

  $\rightarrow I = 1$ (rappresenta il paziente zero)

  $\rightarrow R = 0$

La figura sottostante riporta il risultato ottenuto. L'asse delle $x$ rappresenta il tempo ossia il numero dei giorni da quando l'epidemia è iniziata mentre l'asse delle $y$ rappresenta il numero percentuale di persone per ogni categoria in un dato istante temporale.

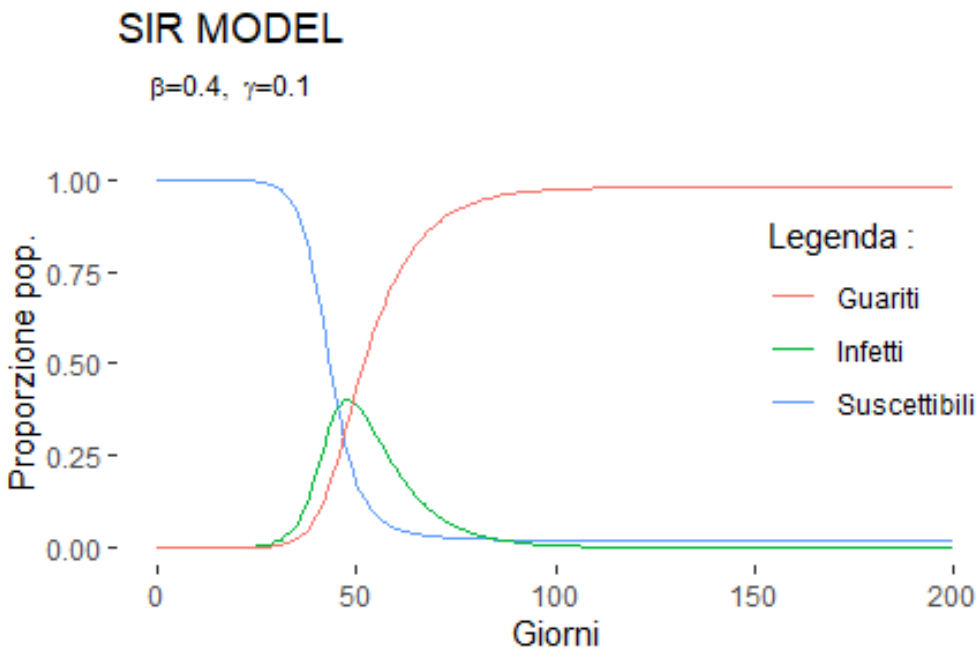


Figura 3.2: SIR Lombardia

L'interpretazione del grafico per ogni linea è la seguente:

- **Linea blu.** Il declino lento della curva, che rappresenta le persone che ancora non sono state infettate, indica la velocità di diffusione della malattia. In questo caso si può concludere che avviene abbastanza lentamente. Intorno al 80 giorno si rileva il valore quasi più basso per tale categoria. Il risultato ci conferma che la lenta diffusione dei contagi all'inizio dell'epidemia, ha reso difficile la comprensione delle vera velocità di diffusione del Covid-19

- **Linea verde.** È essenzialmente la curva epidemica data dalla prevalenza (affrontata nel capitolo 1) della malattia. Raggiunge il picco velocemente al giorno 50 circa per poi cadere lentamente nella seconda parte. Il motivo della discesa più lenta è che ci sono più infetti che persone sane e quindi risulta più difficile che una persona contragga il virus. La pendenza della linea verde riflette il tasso di recupero.
- **Linea rossa.** Rappresenta il numero di persone tolte dalla simulazione perchè sono guarite e hanno sviluppato l'immunita. La malattia si può definire "risolta" quando tale numero è maggiore dei casi infetti, e quindi quando la linea rossa interseca quella verde. In questo caso circa al 55 giorno.

Quindi se si immagina l'assenza di un periodo di incubazione al Covid-19 e l'assenza di casi asintomatici il risultato che ne fuoriesce dal modello è che il picco di infetti viene raggiunto intorno al 50 giorno.

### 3.1.4 L'andamento di $R_t$ in Lombardia

In questo paragrafo si cerca di stimare l'andamento del parametro $R_0$ nel tempo (formalmente $R_t$), per quanto riguarda la Lombardia, per avere una concezione migliore di come i contagi si siano evoluti nel tempo. La stima sarà derivata considerando gli infetti in funzione del tempo, perchè come affermato nella sezione 3.1.2 gli individui malati sono strettamente collegati al basic reproduction number. In particolare gli infetti al tempo $t$ possono essere trovati tramite la seguente relazione:

$$I(t) \approx I_0 \, e^{(R_0 - 1)\,\gamma\, t} \tag{3.5}$$

Ove $I(t)$ infetti al tempo t e $I_0$ infetti al tempo iniziale della malattia.

Per stimare $R_0$ la prima cosa da capire è quale modello utilizzare per aver le miglior performance possibili. In questo caso quelli presi in considerazione sono regressione lineare, regressione log-lineare e regressione poissoniana. La stima del modello, per osservare le variazioni del parametro, è stata effettuata su una finestra mobile di 5 giorni che è sufficiente per stimarne un trend che risulterà crescente o decrescente.

Da un punto di vista grafico, mettendo in relazione gli infetti e il tempo, la miglior soluzione sembra essere un modello di tipo log-lineare, come si può osservare nell'immagine presente nella pagina successiva, in cui si sono raggruppate osservazioni per una finestra temporale di 10 giorni[3].

[3]non si è scelto 5 giorni per non rendere incompresibile il grafico, il risultato è pressochè equivalente

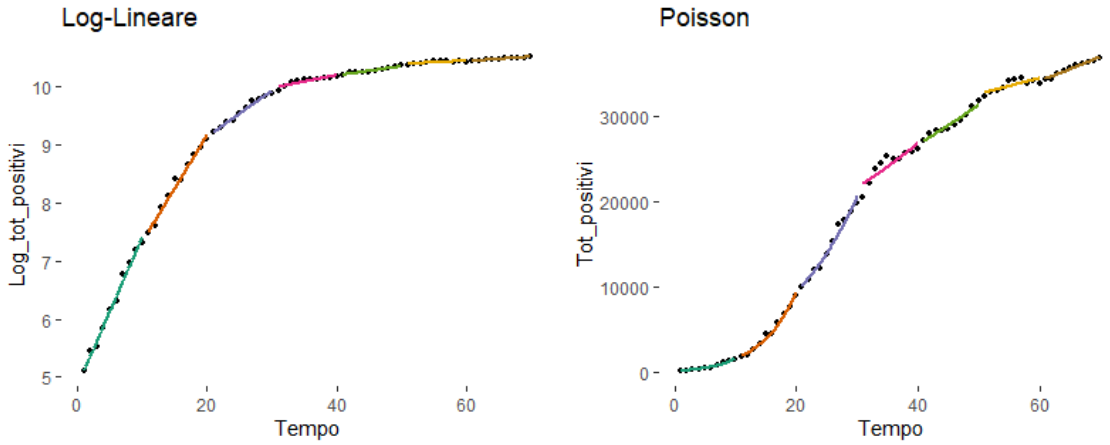


Figura 3.3: Differenze adattamento modelli ai dati

A questo punto siccome il parametro $R_0$ è strettamente collegato al numero degli infetti e in questo caso non risulta di nessun interesse la costruzione di un modello per la previsione (lo scopo è stimare $R_t$) si è scelto come metodo per il confronto tra Poisson, regressione lineare e log-lineare l'errore quadratico medio (comunemente abbreviato con MSE). Più tale valore è basso maggiore è l'adattamento del modello ai dati. I tre candidati sono quindi i seguenti modelli:

- Regressione lineare:
$$I(t) = \beta_0 + \beta_1 t + \epsilon_t \tag{3.6}$$
- Regressione di Poisson:
$$\log(\mu_t) = \underset{n\times k}{X_t'} \ \underset{k\times 1}{\beta} \tag{3.7}$$
ove $X_t'$ matrice trasporta contenente n osservazioni per k modalità e $\beta$ sono i k coefficienti di regressione.
- Regressione log-lineare:
$$\log(I(t)) = \beta_0 + \beta_1 t + \epsilon_t \tag{3.8}$$

Dato che ogni modello è stato utilizzato su finestre temporali di 5 giorni la scelta è stata effettuata sulla media dei vari valori MSE. I risultati sono i seguenti:

| **Modello** | **MSE** |
|---|---|
| Regressione Lineare | 233996.2 |
| Regressione Log-Lineare | 0.001192147 |
| Regressione di Poisson | 0.001202085 |

Il modello log-lineare risulta, anche se già intravisto graficamente, il migliore e si procede quindi con questo modello.
Applicando il logaritmo nella 3.5 si ottiene questa equazione:

$$\log I(t) \approx \log I_0 + (R_0 - 1)\,\gamma\,t \tag{3.9}$$

Preferendo una regressione log-lineare, il modello che si viene a creare per la stima degli infetti è quello definito dall'equazione 3.8. Ove chiaramente i $\beta$ rappresentano i coefficienti di regressione, $\beta_0$ l'intercetta, $\beta_1$ è il coefficiente angolare che indica la pendenza della retta e il parametro $t$ è la variabile indipendente o regressore. $\epsilon_t$ rappresenta l'errore ed è definito da una variabile casuale che in questo caso assume la forma di una normale.
Se si confrontano le due relazioni, 3.8 e 3.9 si può notare che:

$$\begin{aligned} \widehat{\beta}_0 &= \log I_0 \\ \hat{\beta}_1 &= (\hat{R}_0 - 1)\,\gamma \end{aligned} \tag{3.10}$$

Sistemando la 3.10 tramite alcuni elementari passaggi algebrici si ottiene la seguente relazione:

$$\hat{R}_0 = 1 + \frac{\hat{\beta}_1}{\gamma} \tag{3.11}$$

Ora, data la relazione 3.4 si ottiene un ulteriore uguaglianza:

$$\hat{R}_0 = 1 + \frac{\hat{\beta}_1}{\gamma} = \frac{\hat{\beta}}{\gamma} \tag{3.12}$$

da cui si può esplicitare beta per ottenere:

$$\hat{\beta} = (1 + \frac{\hat{\beta}_1}{\gamma})\,\gamma = \gamma + \hat{\beta}_1 \tag{3.13}$$

Questa relazione è fondamentale. Si è riusciti a trovare un legame tra il modello lineare, per la stima degli infetti, e il parametro $\beta$. Risulta quindi ovvio che trovando tale valore si riesca a stimare anche il parametro $R_t$.

Si è fatta anche un ulteriore considerazione. Siccome il numero degli infetti è altamente correlato con il numero di tamponi effettuati giornalmente si è deciso di inserire come peso tale informazione all'interno del modello di regressione log-lineare. In particolar modo ogni unità statistica avrà peso pari a $\frac{1}{N^{\text{r}}_{tamponi}}$ in modo da dar peso maggiore se i tamponi effettuati sono pochi. Basti pensare che più tamponi vengono effettuati, più è facile trovare individui positivi e quindi è corretto fare tale osservazione.
Il risultato della stima di $R_t$ è espresso dal grafico 3.4.

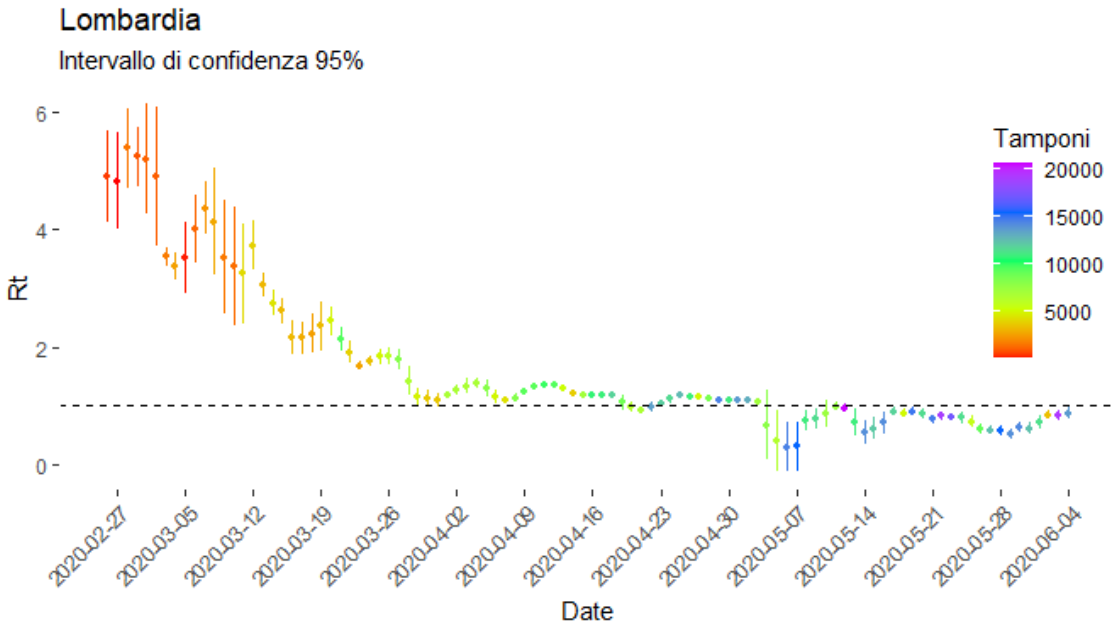


Figura 3.4: Andamento effettivo $R_t$ in Lombardia

Appare evidente un trend descrescente indice di azione governative, come i decreti, funzionanti. Si nota anche che il valore scende sotto la soglia di uno, espressa dalla linea tratteggiata di colore nero, in procinto della fase 2. Le linee prolungate indicano gli intervalli di confidenza, ovvero i valori assunti dal parametro con una certezza in questo caso del 95%. I primi giorni l'andamento è molto più variabile rispetto agli ultimi, questo è indice di un maggior imprecisione nella stima di $R_t$.

## 3.2 Modello $SEIR$

### 3.2.1 In termini teorici

Il modello SEIR è un derivato del modello SIR. S, I e R rappresentano ancora una volta il numero degli individui sensibili, infetti e guariti. Molte malattie hanno una fase latente durante la quale l'individuo è infetto ma non ancora infettivo. Questo ritardo tra l'acquisizione dell'infezione e lo stato infettivo può essere incorporato nel modello SIR aggiungendo una popolazione latente/esposta, E, e lasciando che gli individui infetti (ma non ancora infettivi) passino da S a E e da E a I. Tale lasso di tempo in una malattia altro non è che il periodo di incubazione.
Secondo i dati attualmente disponibili per quanto riguarda il Covid-19, le persone sintomatiche sono la causa più frequente di diffusione del virus. È ritenuto anche possibile che persone nelle fasi prodromiche[4] della malattia, e quindi con sintomi assenti o molto lievi, possano trasmettere il virus. Tale motivo che molto probabilmente rappresenta la causa della diffusione della malattia, non potrà non essere preso in considerazione.

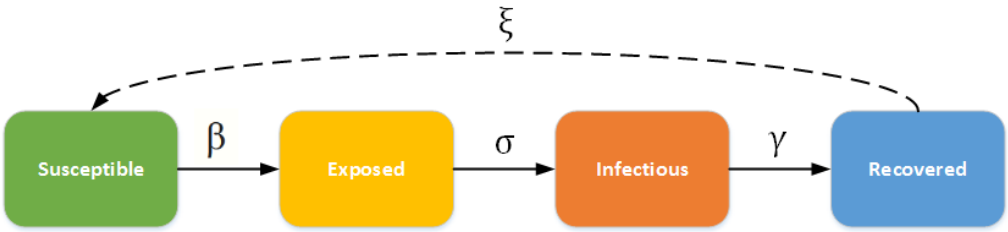


Figura 3.5: Modello SEIR

Il diagramma SEIR /SEIRS, immagine 3.5, mostra come gli individui si muovono attraverso ogni compartimento del modello. La linea tratteggiata mostra come il modello SEIR diventa un modello SEIRS (suscettibile - esposto - infettivo - recuperato - suscettibile), in cui le persone guarite possono diventare nuovamente sensibili (il recupero non conferisce immunità per tutta la vita). Il tasso di velocità con cui accade è dato come nel modello SIR dal valore di $\epsilon$. La malaria ad esempio è una malattia con lunghe durate di incubazione e in cui il recupero conferisce solo un'immunità temporanea. Nel caso del coronavirus non si sa ancora se dopo la malattia permane un'immunità protettiva indotta dagli anticorpi neutralizzanti e quale dose di anticorpi è necessaria perché ciò avvenga. Tutti i dati fino ad ora raccolti, e il

[4]manifestazione morbosa, senza carattere specifico, che precede l'insorgenza dei sintomi caratteristici di una malattia

confronto con un virus parente stretto, il Sars1[5] suggeriscono che almeno per qualche mese, se non anni, chi guarisce definitivamente non rischia di infettarsi di nuovo[14]. Per questo motivo tale parametro lo si assume pari a 0. Il significato di $\beta$ e $\gamma$ rimane identico a quello del caso SIR. Il nuovo parametro $\sigma$ è il tasso di infezione degli individui latenti ed è dato dal reciproco dei giorni di incubazione del virus.

Nella sua forma deterministica il modello SEIR può essere scritto tramite le seguenti equazioni differenziali:

- **Suscettibili:**

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{3.14}$$

- **Esposti:**

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \sigma E \tag{3.15}$$

- **Infetti:**

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{3.16}$$

- **Guariti:**

$$\frac{dR}{dt} = \gamma I \tag{3.17}$$

Anche in questo caso come nel modello SIR si presuppone che la popolazione sia chiusa, e come già spiegato questa approssimazione non porta a errori grossolani.
In questo modello poiché la latenza ritarda l'inizio del periodo infettivo dell'individuo, la diffusione secondaria da un individuo infetto avverrà in un secondo momento rispetto a un modello SIR, che non ha ritardi. Pertanto, un incubazione più lunga comporterà una crescita iniziale più lenta dell'epidemia. Il modello inoltre non includendo la mortalità ha come numero riproduttivo di base, $R_0 = \frac{\beta}{\gamma}$, che non cambia rispetto al modello visto in precedenza.

[5] responsabile dell'epidemia da coronavirus del 2003, simile al virus Mers, epidemia del 2014 nei Paesi arabi

### 3.2.2 Applicazione del modello in Lombardia

Applicando il modello SEIR nel territorio Lombardo si prenderà atto di come le cose sarebbero potute andare (qualsiasi modello non è indice di verità assoluta) se non fossero stati emessi dei decreti atti a limitare i contagi. Tralasciando quindi morti e nati possiamo definire i parametri iniziali del modello che sono:

- $R_0 = 4$ come in precedenza, per gli stessi motivi spiegati prima
- $\gamma$ sarà data ancora una volta da $\frac{1}{D}$ ove $D$ numero di giorni della durata media dell'infezione dalla comparsa dei sintomi, per il coronavirus pari a 10.
- $\beta$ è facilmente ancora calcolabile tramite relazione 3.4
- $\sigma$ è dato da $\frac{1}{G}$ ove $G$ giorni di incubazione del virus. Tale periodo per chi viene contagiato dal virus SARS-CoV-2 è mediamente di 5 giorni[15]. Dato che il modello implica che nel periodo di incubazione non si è infettivi, a tale numero bisogna sottrarre due giorni. Studi hanno dimostrato che mediamente un indivuo è infettivo negli ultimi due giorni prima della comparsa dei sintomi. Si avrà quindi un $D$ pari a $10 + 2 = 12$ e un $G = 3$
- *Parametri iniziali:* in questo caso si è scelto di iniziare il modello con i seguenti parametri uguali a quelli precedentemente usati:

  → $S = N - 1$. A tal proposito si ricorda che la popolazione è chiusa e quindi si presuppone che il primo infetto sia appartenente alla regione Lombardia (non si sa esattamente quanto sia veritiera questa informazione)

  → $I = 1$ rappresenta il paziente zero

  → $R = 0$

Il risultato che ne deriva è visibile dal seguente grafico:

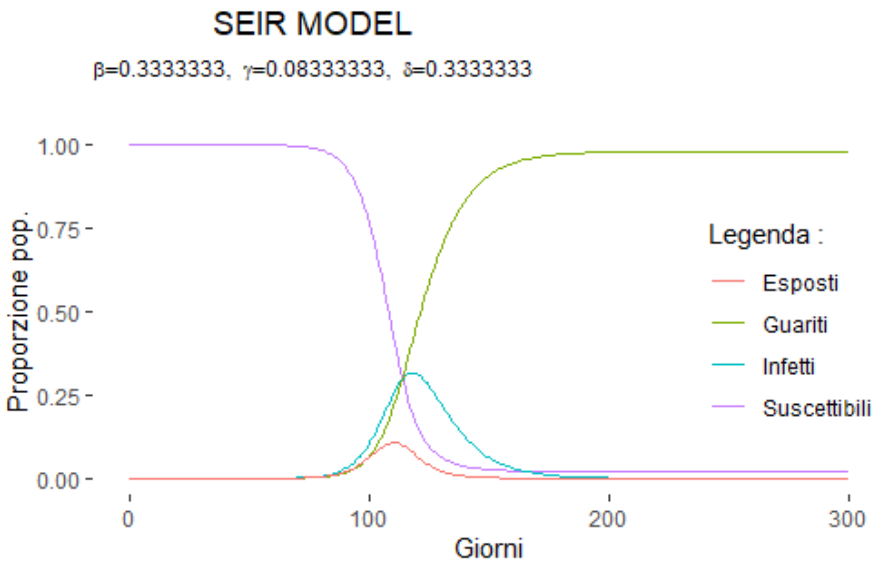


Figura 3.6: SEIR Lombardia

Essendo il modello utilizzato una rappresentazione rispetto al SIR più reale si possono fare interpretazioni più signicative:

- **Linea viola.** Indica la la velocità di diffusione della malattia. Si nota che appena iniziano a comparire un numero ragionevole di infetti la velocità di diffusione diventa molto rapida
- **Linea blu.** Si raggiunge il picco di infetti al giorno 110 circa per poi cadere lentamente nella seconda parte. Il motivo della discesa più lenta è il medesimo del modello SIR
- **Linea arancione.** Rappresenta la curva degli esposti
- **Linea rossa.** Rappresenta il numero di persone tolte dalla simulazione perchè sono guarite e hanno sviluppato immunita

La curva epidemica, rappresentata da questo modello ci indica che presumibilmente il picco si raggiunge al 110° giorno circa, ovvero ben 4 mesi dopo il primo contagiato. Tale risultato se confrontato con il picco dei valori reali ci mostra come in realtà la malattia sia arrivata sul territorio ben prima del 22 febbraio (data dei primi contagi). Sebben non si possa esprimere con certezza tale risulato risulta davvero impensabile che il picco possa essere stato raggiunto in soli due mesi, anche perchè le restrizioni

prese sono servite per appiattire la curva e quindi a prolungare il più possibile l'arrivo del punto di massimo assoluto. Se anche i parametri del modello vengono modificati, ad esempio abbassando l'$R_0$ ad una soglia più vicina a 3 (presumendo che non ci sia sottostima dai calcoli del capitolo 2), o aumentandolo in maniera significativa non si riesce a spiegare un picco in due mesi. E' davvero veritiero pensare che la malattia sia in circololo in Lombardia già dalla metà di dicembre. Alcuni medici dei più importanti ospedali Lombardi affermano di aver avuto polmoniti "strane" attorno appunto al 15/17 dicembre, il che porterebbe validità ai risultati espressi dal modello SEIR.

I risultati sono seppur molto interessanti, ancora deboli, in quanto se si tiene ad esempio in considerazione l'ipotesi di popolazione chiusa nella realtà non è del tutto realistica. Una popolazione chiusa studiata dal modello SEIR prevede che gli individui siano sempre messi in contatto gli uni con gli altri portando ad un contagio totale. Questa ipotesi non è del tutto sbagliata, ma chiaramente non aiuta a capire il vero possibile andamento dei contagi. Se ci si pensa, la popolazione in Lombardia è distribuita in maniera del tutto casuale e non si trova in un "contenitore" in cui tutti gli individui vengono a contatto sempre. In maniera più ampia, ad esempio un'analisi sulla diffusione mondiale, un modello SEIR sarebbe del tutto inadeguato. Non è detto che tutti gli individui debbano per forza riscontrare la malattia, come avviene in questo modello!
Un'ultima parte debole è che in questo caso non tutti gli individui che si infettano sono sintomatici, ma ci sono anche individui nella nostra popolazione che risultano asintomatici. L'osservazione non è banale in quanto all'interno del compartimento sono considerati come possibili infettori ma non rientrano negli infetti (I).

L'ultima osservazione interessa l'immunità di gregge che è un fenomeno per il quale, quando un'elevata quota della popolazione risulta immune dall'infezione, gli individui suscettibili sono indirettamente protetti. La proporzione della popolazione immune necessaria per fermare la trasmissione della malattia è in funzione di $R_0$ (non di $R_t$) e può essere ricavata dalla seguente formula:

$$1 - \frac{1}{R_0} \;=\; 1 - \frac{1}{4} \;=\; 75\% \tag{3.18}$$

Il passo finale del lavoro consiste quindi in una costruzione di un modello statistico molto più realistico che cerca di spiegare l'evoluzione della malttia in base agli spostamenti delle persone tramite mezzi pubblici all'interno della regione Lombarda. Non risolveranno chiaramente tutti i problemi descritti fino ad ora ma tale visualizzazione permetterà di trovare risultati più chiari.

# Capitolo 4

# La struttura dei mezzi pubblici in Lombardia

## 4.1 La matrice OD

Lo scopo di questo paragrafo è quello di introdurre i dati OD (origine-destinazione) che sono una componente importante per l'applicazione di molti modelli di pianificazione dei trasporti. Ci si servirà di tali informazioni per raggiungere l'obbiettivo ultimo che consiste nella costruzione di un modello stocastico che possa mettere in relazione gli spostamenti tramiti mezzi pubblici con la diffusione dell'epidemia. Si svolgerà anche un analisi generale che sarà assestante rispetto al lavoro fatto fino ad ora.

Come suggerisce il nome, i dati origine-destinazione (OD) rappresentano i movimenti attraverso lo spazio geografico, da una certa origine (O) a una destinazione (D). Il nome di matrice le viene attribuito per il fatto che ad ogni paese di origine ($j$) presente sulle righe viene attribuita una colonna ($i$) con una destinazione, e l'elemento in posizione $ji$ rappresenta il numero di spostamenti rilevati. Di solito però i dati sono, per facilità di utilizzo, in forma di semplice dataset. All'interno contengono quindi dettagli negli spostamenti tra due punti geografici o, più comunemente, zone[1]. La maggior parte dei dati OD fanno riferimenti a posizioni di inizio e fine con colonne, come in questo caso, identificate da degli *"ID"*, contenenti stringhe di caratteri che identificano i paesi di origine e destinazione. Alcune volte le stringhe vengono sostituite da latitudine e longitudine delle città di riferimento.

[1] spesso rappresentate da un centroide, alcune volte solo dal nome della città/paese

I dati OD in genere contengono più attributi non geografici. È sempre incluso il numero di viaggi che hanno luogo da una certa origine ad una destinazione in un determinato periodo di tempo (ad esempio una tipica giornata lavorativa). Ulteriori attributi possono includere la suddivisione in base alle modalità di trasporto utilizzate per i viaggi. Di solito, come in questo caso, viene catturata una sola modalità. Ad esempio se i viaggi vengono effettuati da una combinazione di modalità bici-treno-camminata viene conteggiata solamente la modalità di viaggio dominante. Ulteriori disaggregazioni dei conteggi complessivi possono comprendere conteggi di viaggio in periodi di tempo diversi (esempio suddivisione per orari).

### 4.1.1 La matrice OD Lombarda

La matrice origin-destination[16] è un dataset composto da 9 milioni di records e 45 variabili/colonne. Le informazioni presenti altro non sono che un aggiornamento della matrice OD riferita al 2014/2015. Tale dataset è articolato in tabelle per la descrizione sintetica degli spostamenti, con elementi disposti su più righe e su più colonne. Origini e destinazioni sono aggregate in zone di mobilità (sia interne al territorio della Lombardia, sia esterne). La matrice OD 2014 include:

- 8 modalità[2] (auto conducente, auto passeggero, TPL gomma[3], TPL ferro[4], moto, bici, piedi e altro[5])
- 5 motivi di spostamento (lavoro, studio, occasionale, affari e rientri a casa)

Per la costruzione di tale dataset nel documento associato all'elaborazione dei dati sono presenti alcune considerazioni:

- La matrice OD-2014 è il risultato della complessa interazione tra modellazioni trasportistiche, questionari on-line, interviste, analisi di indagini disponibili e della domanda esistente rilevata
- I dati si riferiscono ad un giorno feriale medio con periodo da febbraio a maggio
- Gli spostamenti a piedi, sono relativi ai tratti superiori a 10 minuti

[2]intese come modo prevalente dello spostamento. Qualora uno spostamento fosse costituito da più segmenti, la matrice OD-2014 lo interpreta come unico segmento, che ha come modalità quella prevalente tra i differenti segmenti

[3]bus urbani, bus extraurbani, pullman e filobus

[4]treno, metro e tram extraurbano

[5]compreso aereo

- Gli spostamenti descritti considerano la modalità prevalente (spostamenti più significativi)
- La zona di riferimento è il singolo comune
- La fascia di popolazione considerata è di età maggiore o uguale a 14 anni

Al fine di rappresentare al meglio la mobilità del territorio lombardo, l'area di studio è stata suddivisa in 1450 zone di mobilità, ove 1264 sono costituite da singoli comuni, 108 da aggregazione di alcuni comuni e 78 costituite dalla disaggregazione di grossi comuni. Nel dataset quindi sono presenti per ogni origine e destinazione dei valori trovati tramite complessi procedimenti che indicano il numero medio di persone che si muovono per causa e mezzo in un giorno lavorativo nel periodo indicato in precedenza.

La definizione della matrice regionale 2014 è basata su un modello di domanda di trasporto, ovvero uno strumento che, alimentato con dati rappresentativi della popolazione e del territorio regionale e calibrato con informazioni rilevate direttamente sul campo, ha lo scopo di rappresentare nel migliore dei modi la realtà. Per poter procedere alla determinazione delle matrice origine-destinazione, è stato utilizzato un approccio complesso a stadi successivi, precisamente 4, mediante opportuni modelli di:

- **Emissione ed attrazione degli spostamenti**. Sono stati utilizzati indici di mobilità. Per emissione si intendono gli spostamenti emessi da un certo paese e per attrazione gli spostamenti in entrata in una certa località. Il modello prevede poi un bilanciamento iterativo del numero di spostamenti attratti con quello degli spostamenti emessi (o viceversa, in funzione del motivo specifico). L'output di questa fase è rappresentato dal totale degli spostamenti emessi ed attratti da ciascuna zona e costituisce la base di partenza per il modello di distribuzione
- **Distribuzione degli spostamenti**. È stato utilizzato un modello gravitazionale che ha ripartito gli spostamenti in funzione di due fattori, attrattività delle zone di destinazione e impedenza temporale tra la zona di origine e quella di destinazione. Utilizzando il grafo di rete stradale, per ciascuna coppia origine-destinazione sono state determinate le matrici delle distanze (in km) e dei tempi (in minuti). Questa fase, inoltre, ha consentito di verificare la connessione di tutte le zone tramite i grafi

- **Ripartizione modale**. Per ogni spostamento, effettuato tra le zone origine e destinazione, per ogni motivo, è stata ricavata l'aliquota di spostamenti che utilizza un certo modo di trasporto. Nello specifico, è stato utilizzato un approccio "kite" considerando separatamente

  - Ripartizione modale degli spostamenti aventi origine in un singolo comune/aggregazione di comuni e destinazione in altro comune/aggregazione di comuni attraverso l'utilizzo del modello *logit/regressione logistica*
  - Ripartizione modale all'interno della diagonale della matrice (spostamenti con origine e destinazione all'interno dello stesso comune/aggregazione di comuni) e, nelle sottomatrici, spostamenti con origine e destinazione all'interno dello stesso comune (per i comuni disaggregati in più zone) tramite un approccio *deterministico*

- **Assegnazione alla rete tramite l'uso di grafi stradali e ferroviali**. Per assegnare la domanda alla rete, viene simulata l'interazione tra

  - domanda di trasporto, rappresentata dalla matrice origine-destinazione della modalità privata e da quella pubblica ferroviaria
  - offerta di trasporto, rappresentata dal grafo di rete stradale e dal grafo dei servizi ferroviari. Si determinano così i flussi di utenti e le prestazioni per ciascun elemento del sistema di offerta.

## 4.2 Analisi della matrice

Un'analisi sul dataset iniziale è utile per capire le abitudini di spostamento dei cittadini Lombardi. Le prime importanti osservazioni che si possono fare sono le seguenti:

- Gli spostamenti che quotidianamente interessano la Lombardia sono pari a 16,40 milioni e sono cresciuti, rispetto al 2002, del 4,6%
- Il Comune di Milano, come si poteva immaginare, si conferma la città con il maggior utilizzo dei mezzi pubblici

Risulta interessante comprendere, quale sia il mezzo di trasporto più utilizzato all'interno della regione regione. Il grafico a torta rappresenta la percentuale di impiego di ogni categoria di trasporto.

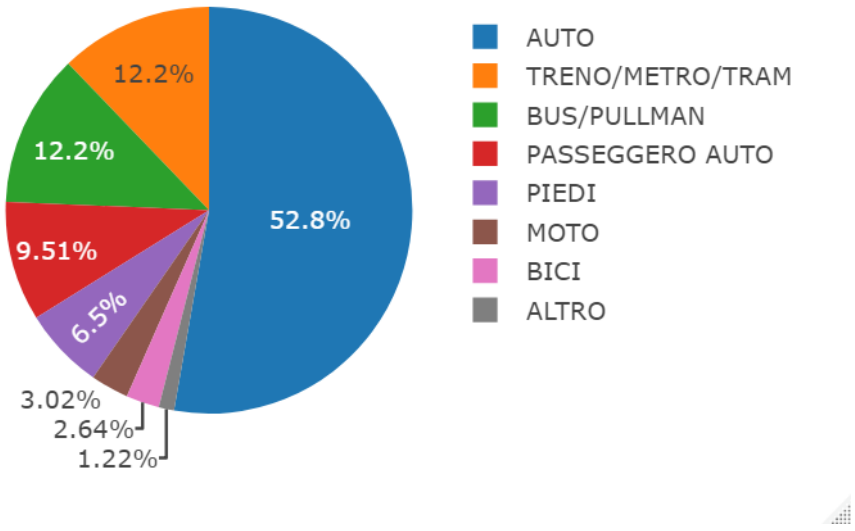


Figura 4.1: Mezzi di trasporto più utilizzati in Lombardia

Si evince come la macchina sia il mezzo di trasporto più utilizzato per più della metà degli spostamente giornalieri. Se si considerano anche gli individui che si spostamento con tale mezzo, senza essere conducenti, si spiegano addirittura il 62% dei movimenti. Treno/metro e tram hanno all'incirca gli stessi individui rispetto ai bus e pullman. Una percentuale minore è rappresentata da spostamenti a piedi in moto e in bici.

Una seconda analisi viene effettuata per capire quali siano i maggiori motivi da cui derivano gli spostamento giornalieri. Nel grafico sulle ascisse troviamo i motivi di spostamento che in base al colore, dal più scuro (meno intenso) al più chiaro indicano la percentuale rispetto al totale degli spostamenti.

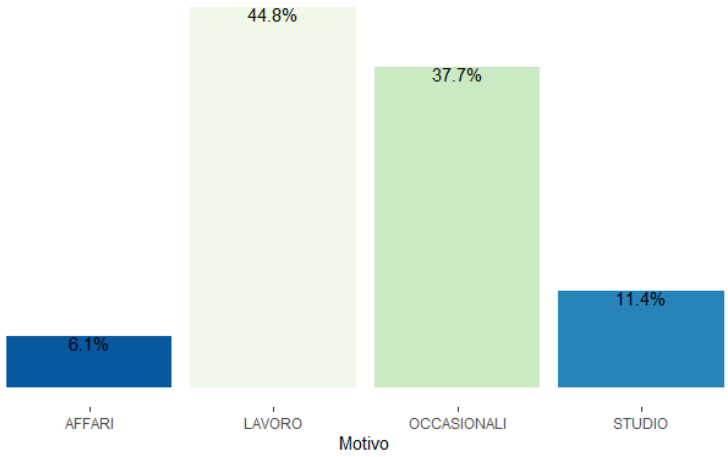


Figura 4.2: Motivi di spostamento in Lombardia

Il lavoro costituisce il motivo principale di spostamento, se unito agli spostamenti con affari si spiega il 50% del motivo di mobilità giornaliera. Una percentuale importante è comunque dovuta agli spostamenti occasionali. Nel grafico le somme delle percentuali non sono pari a 100 in quanto non sono considerati i rientri nelle varie abitazioni.

Un'ultima analisi è fatta per capire all'interno di ogni *motivo di spostamento* quale sia il mezzo più usato. In questo caso nell'asse delle ordinate sono presenti i vari motivi di movimento condizionati al mezzo preso e il colore ne indica l'intensità. Si riportano nel grafico soltanto percentuali, per ovvi motivi, maggiori o uguali a 1.

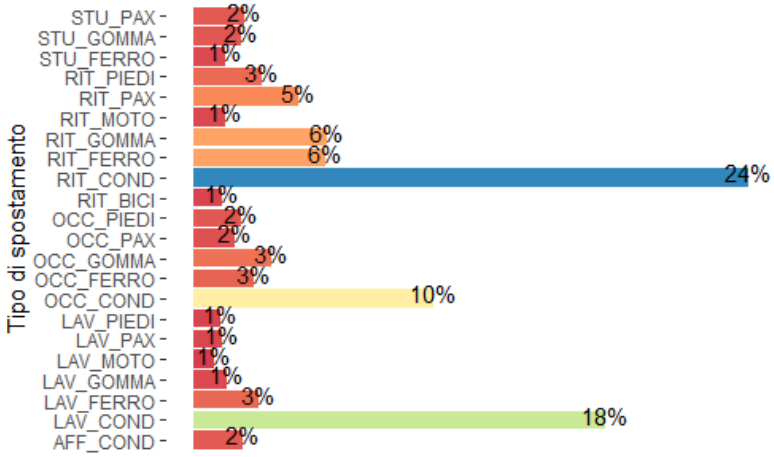


Figura 4.3: Mezzi di trasporto più utilizzati in Lombardia per motivo

Si evince che per ogni motivo di spostamento il mezzo più utilizzato per muoversi è sempre l'auto. I mezzi pubblici costituiscono percentuali al quanto minori, soprattutto per quanto riguarda i motivi di spostamento lavorativi, mentre assumono valori più alti per movimenti occasionali. Emerge anche che gli studenti preferiscono spostarsi maggiormente o come passeggeri in auto o tramite mezzi pubblici.

## 4.3 I collegamenti dei mezzi pubblici Lombardi

Per quanto riguarda la matrice OD-2019, utilizzata in tale lavoro, costituisce un aggiornamento della OD-2014. Sono stati ad esempio aggiornati i dati relativi a nuove tipologie di collegamenti.
Al fine di costruire la rete dei trasporti dei mezzi pubblici si sono considerate le sole modalità *TPL gomma e TPL ferro* indicanti tutti i trasporti pubblici relativi a tutti i possibili motivi di spostamento tranne quello degli affari (spostamenti quasi nulli).

Il risultato è stato il seguente:

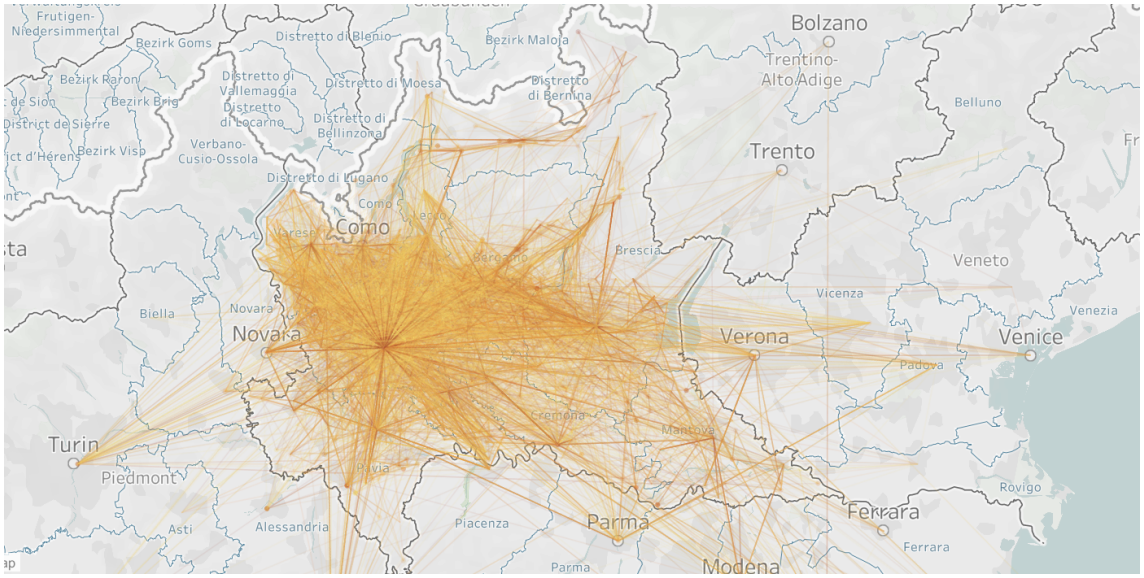


Figura 4.4: La rete dei trasporti dei mezzi pubblici Lombardi

Ove la colorazione più scura indica la presenza di maggiore quantità di persone in quella tratta. È evidente come la rete è al quanto densa e molto concentrata (come ci si poteva aspettare) attorno alla provincia di Milano. Si notano anche collegamenti con altre regioni, verso le quali le persone mediamente ogni giorno si spostamento dalla regione Lombardia.

Uno dei tanti problemi nella fase di pulizia è stato quello di gestire comuni come Milano e Monza che nel dataset vengono riportati in micro-aree (ad esempio Milano 1, Milano 2. . . ). La decisione è stata quella di aggregare tali informazioni per considerare i vari sottogruppi come unico ente. Le suddivisioni di spostamenti tramite differenti mezzi sono state raggruppate. In questo modo si ha un numero di persone medie che si sposta dal comune A al comune B con qualsiasi mezzo di trasporto. Si ricorda a tal proposito che lo scopo è capire quante persone si spostano da un punto ad un altro e non con che mezzo si muovono.

# Capitolo 5

# Modello stocastico $SEII_aR$

## 5.1 Introduzione al modello

ll mondo è in continua urbanizzazione, si stanno formando in questi ultimi decenni ammassi di aree urbane densamente popolate e aree rurali scarsamente popolate. Le persone oggi giorno si spostano continuamente nelle aree urbane e le città stanno crescendo di dimensioni. Trasporti e città affollate sono per i virus come l'ossigeno per il fuoco.
In questo ultimo capitolo si propone una combinazione tra la rete dei trasporti e un modello statistico atto a stimare la diffusione del coronavirus nel territorio Lombardo. Lo scopo è comprende come l'urbanizzazione e i mezzi di trasporto influenzano la trasmissione delle malattie, nello specifico per il Covid-19.

Alcune considerazioni possono essere già fatte. Nelle aree meno urbanizzate ci saranno meno mezzi di trasporto e quindi quello che ci si aspetta è che in questi luoghi l'andamento delle infezioni sia minore rispetto ad aree più urbanizzate. Si studierà quindi l'interazione tra urbanizzazione e diffusione delle malattie infettive.
I risultati hanno implicazioni dirette per le linee guida e le politiche di controllo delle malattie. Ad esempio è facile immaginare che l'applicazione delle restrizioni ai viaggi ha un impatto maggiore sul numero finale dei malati nei paesi più urbanizzati rispetto ai paesi meno urbanizzati. In particolare come si potrà osservare mediante il modello che si costruirà, le restrizioni ai viaggi sono più efficaci nel ritardare l'epidemia rispetto alla quarantena se applicate subito.

L'urbanizzazione comporta il raggruppamento di persone all'interno di una precisa area geografica. Questa informazione è da tenere in considerazione in quanto più persone si trovano in uno stesso scomparto più in teoria veloce ne sarà la diffusione

della malattia. Detto ciò lo scopo non è quello di descrivere singoli focolai in popolazioni specifiche o di trovare il "miglior modello" o la strategia per un paese specifico, ma piuttosto studiare il fenomeno da un punto di vista più generico per capire come il Covid-19 si sarebbe diffuso se la mobilità non fosse stata limitata.

Si utilizzerà un modello epidemiologico per descrivere l'epidemia di tipo $SEII_aR$ mentre per quanto riguarda la popolazione di ogni singolo comune nel territorio Lombardo, si è utilizzato il dataset di Regione-Lombardia per integrare alla matrice OD-2019 la densità demografica di ogni singolo paese.

Si simulerà la diffusione della malattia infettiva per tutta la regione a partire da un preciso comune.

## 5.2 La costruzione del modello

### 5.2.1 Le equazione stocastiche

Il modello di dinamica della malattia è un modello stocastico che può essere descritto come una rete in cui ogni nodo rappresenta una posizione e ogni margine tra le posizioni rappresenta le persone che viaggiano tra i paesi e che quindi possono diffondere ulteriormente la malattia. In ogni luogo esiste una serie separata di equazioni stocastiche che governano la dinamica della malattia locale, al quale si aggiunge un vettore di spostamenti.

Prima di iniziare con termini ed equazioni complesse è bene spiegare il funzionamento del modello da un punto di vista intuitivo tramite la figura 5.1.

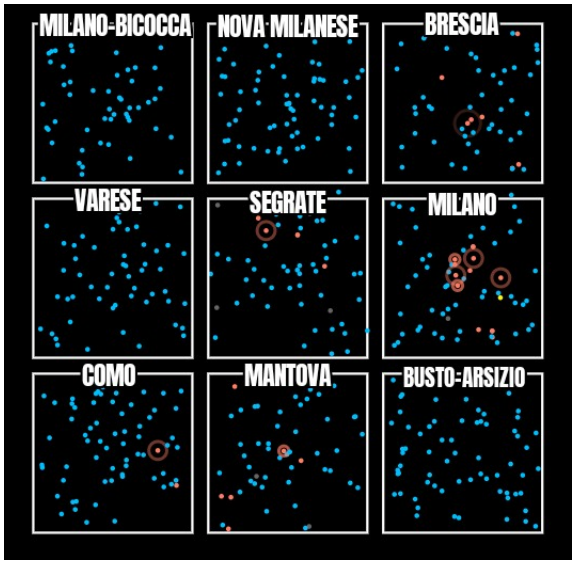


Figura 5.1: Funzionamento modello stocastico di tipo $SEII_aR$

Ogni paese può essere rappresentato tramite dei contenitori che in questo caso hanno nomi scelti in maniera del tutto casuale. Le palline al loro interno rappresentano la popolazione di riferimento. La colorazione rossa dei pallini rappresenta l'individuo infetto mentre rimane azzurra se l'individuo è sensibile. All'interno di ogni contenitore le persone si spostano in modo del tutto casuale, e l'andamento dell'infezione è descritta da un semplice modello, il cui funzionamento rimane identico a quello del capitolo precedente. In questo caso Milano-Bicocca non ha infetti all'interno del suo scomparto e quindi è comprensibile intuire come la malattia non si diffonderà. Avverrà invece il contrario nella città ad esempio di Milano. Eliminando l'ipotesi in cui gli individui sono statici nel loro comune, si permette alle persone di muoversi in diversi scomparti. Si immagini che una persona di Milano che ha contratto il coronavirus (pallina rossa) si sposti per motivi di studio nella città di Milano-Bicocca, allora in tale scomparto comparirà un individuo infetto che potrà contagiare il resto della popolazione.

La dinamica dell'infezione in ogni paese, viene descritta da un modello $SEII_aR$ stocastico. Siano $S^i(t)$ $E^i(t)$ $I^i(t)$ $I^i_a(t)$ il numero di individui sensibili, individui esposti, individui infettivi sintomatici e individui infettivi asintomatici al momento $t$ nella posizione $i$.

Le equazioni stocastiche di questo modello sono rispettivamente:

- **Suscettibili:**

$$S^i(t+\Delta t) = S^i(t) - Binom(S^i(t), \frac{\beta \Delta t I^i(t)}{N_i} + P_{ta}\frac{\beta \Delta t I_a^i(t)}{N_i}) \tag{5.1}$$

- **Esposti:**

$$E^i(t+\Delta t) = E^i(t) - Binom(S^i(t), \frac{\beta \Delta t I^i(t)}{N_i} + P_{ta}\frac{\beta \Delta t I_a^i(t)}{N_i} - \\ - Multinom(E^i(t), P_a\lambda\Delta t, (1-P_a)\lambda\Delta t) \tag{5.2}$$

- **Infettti sintomatici:**

$$I^i(t+\Delta t) = I^i(t) + Binom(E^i(t), (1-P_a)\lambda\Delta t) - Binom(I^i(t), \gamma\delta t) \tag{5.3}$$

- **Infettti asintomatici:**

$$I_a^i(t+\Delta t) = I_a^i(t) + Binom(E^i(t), P_a\lambda\Delta t) - Binom(I_a^i(t), \gamma\delta t) \tag{5.4}$$

Ove $\beta$ e $\gamma$ parametri già definiti in precedenza negli altri modelli. $N_i$ rappresenta la numerosità della popolazione nella posizione $i$, $Binom(n,p)$ è la distribuzione binomiale con $n$ prove e probabilità di successo pari a $p$ e $Multinom(n, p_1, p_2)$ è la distribuzione multinomiale con $n$ prove e probabilità di successo pari a $p_1$ e $p_2$. $P_{ta}$ non è altro che la probabilità di trasmissione dell'infettivo asintomatico ridotta rispetto a quello sintomatico mentre $P_a$ è la probabilità di essere un individuo asintomatico. L'equazione per il numero di persone guarite $R^i(t)$ è superflua in quanto in ogni momento $S^i(t) + E^i(t) + I^i(t) + I_a^i(t) + R^i(t) = N_i$ e quindi si può semplicemente ricavarla.
Con questo tipo di modello cambia l'equazione del basic reproduction number che risulta essere:

$$R_0 = \frac{\beta}{\gamma}\left(P_{ta}P_a + (1-P_a)\right) \tag{5.5}$$

Si ricorda che la distribuzione binomiale è una distribuzione di probabilità discreta che descrive il numero di successi in un processo di Bernoulli. Serve a capire il numero di "successi" in n "prove", in questo caso quante persone si infettano ("successo") dati $n$ contatti tra individui. Si fa notare come l'introduzione di questa variabile casuale rappresenti una grossa differenza rispetto ai modelli precedentemente considerati. La distribuzione multinomiale invece è una distribuzione di probabilità discreta che generalizza la distribuzione binomiale in più variabili. In altri termini, laddove la

distribuzione binomiale descrive il numero di successi in un processo di Bernulli, per il quale ogni singola prova può fornire due soli risultati, la distribuzione multinomiale descrive il caso più generale in cui ogni prova possa fornire un numero finito di risultati, ognuno con la propria probabilità[17]. In questo caso serve per attribuire una probabilità di contagio ad un potenziale individuo infetto (che ancora non abbia manifestato sintomi, classe E, esposti), in questo modo:

- $p_0$ = probabilità di non essere infettato
- $p_1$ = probabilità di essere infettato e avere sintomi
- $p_2$ = probabilità di essere infettato senza avere sintomi

### 5.2.2 Mobilità e dinamica globale delle infezioni

Nel modello, la dinamica della malattia nelle unità di analisi è formulata attraverso gli individui in viaggio. Ogni individuo ha una casa e un luogo di lavoro definiti che identificano rispettivamente l'origine e la destinazione. Durante il giorno, le persone si mescolano nel luogo di lavoro, mentre di notte si mescolano nel luogo di residenza. Si sottolinea l'importanza di tenere traccia degli individui pendolari, vale a dire che bisogna assicurarsi che gli stessi individui siano gli stessi pendolari ogni giorno.

Oltre al pendolarismo, si considerano i viaggi effettuati senza tale dinamica. Il viaggio senza pendolarismo è implementato nel modello consentendo a tutti gli individui che non si spostano tra paesi di viaggiare in una posizione casuale, con una certa probabilità fissa. Il modello distingue tra giorno e notte. Durante il giorno le persone possono infettare/essere infettata nel luogo in cui lavorano, mentre di notte possono infettare/essere infettate nel luogo in cui vivono. Sono gli stessi pendolari che viaggiano avanti e indietro ogni giorno. All'inizio di una giornata, tutti i pendolari vengono inviati al loro posto di lavoro, dove si mescolano per 12 ore. In questo modo i lavoratori possono infettare le persone nella citta di destinazione. Una volta finite le 12 ore lavorative i pendolari vengono quindi inviati alle rispettive posizioni di casa, dove si mescolano per le restanti 12 ore. In questo modo nelle 12 ore successive possono infettare le persone della città di orgine che a loro volta possono quindi infettare la popolazione di quel paese!

## 5.3 La scelta dei parametri

La parte fondamentale prima di eseguire il modello è la scelta dei parametri. Le scelte effettuate sono le seguenti:

- $P_a$ che rappresenta la probabilità che un individuo ha di essere asintomatico la si ipotizza pari al 66%[18]
- $P_{at}$ che è la probabilità di trasmissione in meno rispetto ad un indivio sintomatico di un asintomatico la si pone pari al 50% circa [19]
- $\gamma$ sarà ancora una volta il reciproco della durata media dei giorni dell'infezione che in questo caso è pari a 12 e non 10 per osservazioni già effettuate nel paragrafo 3.2.2
- $R_0$ lo si pone ancora una volta pari a 4
- $\beta$ da equazione 5.5 questa volta sarà differente rispetto ai modelli precedenti.

$$\beta = \frac{\gamma R_0}{(P_{ta}P_a + (1 - P_a))} \tag{5.6}$$

- $\sigma$ cioè il reciproco del periodo di incubazione sarà dato ancora una volta dal reciproco di 3, e non 5
- *Parametri iniziali:* In questo caso la scelta poteva ricadere su qualsiasi comune presente all'interno della regione Lombardia. Si è scelto per differenti ragioni di ipotizzare il primo caso infetto all'interno del comunque di Milano. Uno di questi motivi sono che considerando solo la rete dei trasporti ipotizzare un caso in un comune piccolo farà espandere il contagio in maniera molto lenta il che è molto distante dalla realtà. Per questo si è scelto di considerare quindi questi numeri:

  $\rightarrow S = N_{Milano} - 1$

  $\rightarrow I_{Milano} = 1$ questo rappresenta il paziente zero

  $\rightarrow I_a = 0$

  $\rightarrow R_{Milano} = 0$

## 5.4 Il modello applicato alla rete Lombarda

Una volta inizializzati i parametri iniziali, si può applicare il modello $SEII_aR$ alla rete dei trasporti Lombardi, immagine 4.1. Il risultato dei contagi è espresso dal seguente grafico che mette in relazione l'avanzare dei giorni con l'incidenza. Per rendere più agevole la visualizzazione dei risultati si è scelto di raggruppare tutti i comuni per la provincia di appartenenza. Si avrà cosi un'idea generale di come è avvenuta la diffusione. Provando con differenti comuni i risultati cambiano di poco, quello che varia di solito è il tempo con cui si raggiungo i vari picchi. Per quanto riguarda l'asse delle $y$ rappresenta l'incidenza giornaliera mentre per quanto riguarda l'asse delle $x$ rappresenta lo scorrere del tempo,espresso in questo caso in giorni.

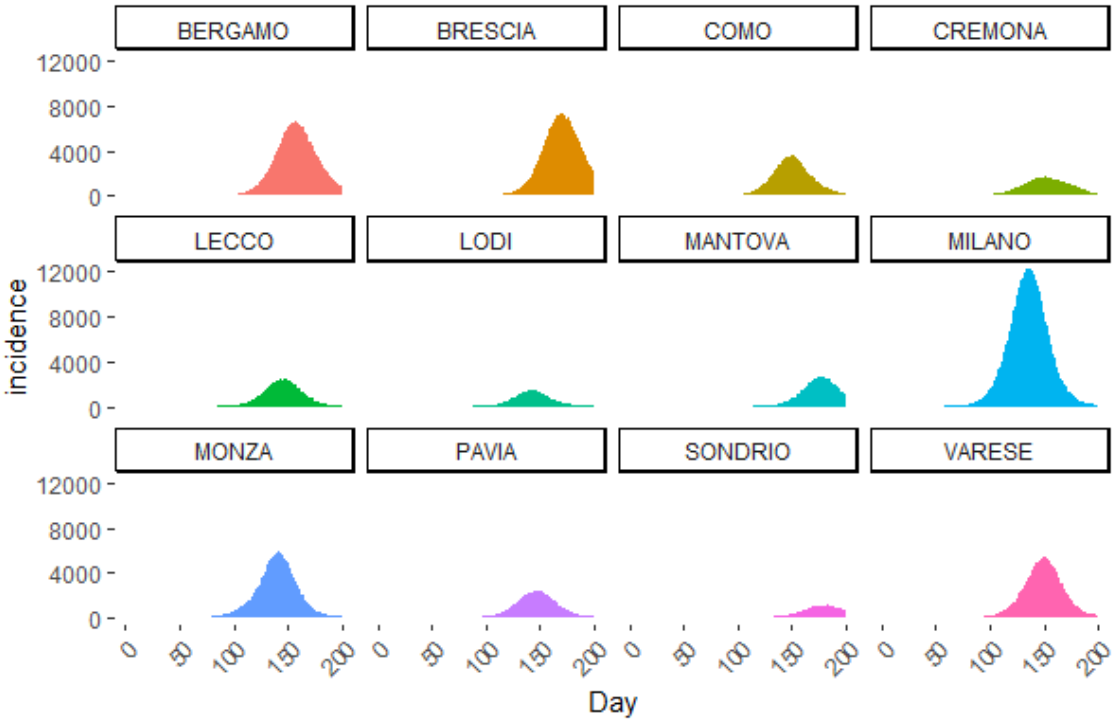


Figura 5.2: Andamento incidenza tramite modello stocastico

Si ha quindi un immagine di come l'epidemia si sarebbe diffusa se non fossero stati applicati decreti per il contenimento dei movimenti degli individui. La prima evidente osservazione da fare è che la provincia in cui si sarebbe diffusa più velocemente sarebbe stata Milano che è anche la provincia con la più alta densità di popolazione. Il risultato è in linea con le aspettative, i flussi dei trasporti sono maggiormente sviluppati in tale provincia e quindi ogni giorno ci saranno sempre più malati. Si intuisce quindi che la diffusione avviene in base all'intensità degli spostamenti delle

persone. La provincia di Sondrio, che sembra aver una bassa rete di trasporti ( si guardi immagine 4.4), è effettivamente quella con la più lenta diffusione. In province invece come Bergamo e Brescia, altamente collegata con Milano, l'andamento dell'infezione è simile anche se arriva con qualche giorno di ritardo. All'inizio dell'epidemia resta evidente come questa si sia sviluppata in maniera lenta, causa forse della sua non individuazione.

Per quanto riguarda la durata del virus si evince ancora una volta come i giorni per l'arrivo del massimo assoluto siano distanti dai 2 mesi avvenuti realmente. Bisogna comunque prendere in considerazione che il picco rilevato nei dati, come già spiegato, non è il picco reale, in quanto non si conosce con esattezza il numero di infetti. Per assurdo si potrebbe aver toccato il massimo anche qualche settimana più tardi senza averne evidenza empirica. Le politiche di intervento del governo sono state fatte in modo tale da appiattire la curva epidemica, ovvero ritardare il picco il più possibile per evitare sovraffolamenti all'inteno del sistema sanitario. In questo modo risulta sempre più impossibile che il coronavirus sia presente soltanto da febbraio nel nostro paese. Sarebbe possibile solo se le stime di $R_0$ fossero completamente sbagliate (servirebbe un valore molto più alto per spiegare due mesi), il che è altamente improbabile. È quindi evidente che si convive con l'infezione almeno dal mese di gennaio, molto probabilmente l'arrivo si spiega intorno alla metà di dicembre ma il modello non esclude neanche il mese di novembre. Purtroppo risulta difficile assegnare un periodo preciso, il massimo che si riesce ad ottenere è quanto detto.

È evidente come limitare l'utilizzo dei trasporti costituisca una grande importanza per la riduzione dei contagi. E' chiaro che tali misure prima vengono applicate maggiore è il loro impatto. Una volta che i contagiati hanno raggiunto e infettato un individuo in un determinato paese, la riduzione dei viaggi non ne costituisce un limite di alcun tipo per la diffusione della malattia in quel paese. Un vincolo degli spostamenti limitati al paese con i primi casi, costituisce il mezzo più efficace per sradicare le infezioni. Nel caso non si riuscisse a procedere in questo modo la quarantena ne risulta il mezzo necessario per abbassare i contagi. Basti pensare di nuovo all'immagine 5.1. Se in quel momento temporale si bloccassero tutti i trasporti in entrata e uscita dalle città di Milano, Brescia, Segrate e Como, non si avrebbero infetti nelle restanti città. Non ci sarebbero infatti vie di accesso per il virus, e quindi quei paesi potrebbe benissimo continuare le attività e gli spostamenti quotidiani.

# Conclusioni

In questa tesi sono state trattate molteplici tematiche. Sono stati introdotti diversi tipi di indicatori per lo studio del Covid-19, ognuno con i suoi pregi e difetti. È stata calcolata la curva epidemica di questa specifica malattia per l'Italia e Hubei, fondamentale per la stima del parametro $R_0$, necessario poi per la costruzione dei vari modelli, dal più semplice SIR al più complicato $SEII_aR$. Per quest'ultimo sono stati utilizzati i dati relativi agli spostamenti quotidiani delle persone nel territorio Lombardo per capire come si sarebbe evoluto il contagio senza l'introduzione dei decreti.

L'analisi della curva epidemica ha evidenziato come la velocità di diffusione del virus in Italia sia stata maggiore rispetto ad Hubei.
L'analisi dei tassi di mortalità all'interno del territorio italiano ha portato in evidenza che il Covid-19 ha una letalità maggiore di circa 3.5 volte rispetto all'influenza stagionale. I valori trovati sono stati precisamente di 0,056% per il Covid-19 e 0,016% per l'influenza. Questo numero sebben non risulta eccessivamente elevato, ha indotto a misure restrittive non tanto per la sua pericolosità, bensì per evitare un sovraffollamento delle strutture ospedaliere. Analizzando inoltre le tipologie di persone che sono decedute con la presenza del coronavirus, risulta che il 96% dei casi avevano almeno una malattia pregressa.

La stima del parametro $R_0$ tramite tasso di crescita esponenziale, il metodo più utilizzato, ha dato come risultato in Italia $\hat{R}_0 = 3.60$ con intervallo di confidenza $[3.53, 3.67]$. La stima tramite massima verosomiglianza, un altro metodo che viene usato, ha portato invece a valori leggermente diversi. La stima puntuale è risultata $\hat{R}_0 = 3.10$ mentre l'IC $[3.04, 3.17]$. I valori con entrambi i metodi sono risultati entrambe le volte più alti rispetto alla stima effettuata sui dati di Hubei. Il motivo potrebbe essere trovato dalle azione tardive e meno severe all'interno del nostro territorio. Ad oggi non esiste comunque un unico stimatore (i due usati in questo lavoro sono quelli maggiormente utilizzati, ma ne esisteno altri) e le molteplici assunzioni per l'utilizzo dei vari metodi quasi mai vengono rispettate nella realtà portando ad imprecisioni. Il risultato ottenuto, che indica quante persone un individuo infetta mediamente, appare più basso di altre malattie come ad esempio Vaiolo $\hat{R}_0 = [3.5,6.0]$

e Parotite $\hat{R}_0 = [10,12]$. Il problema associato al Covid-19 deriva infatti, non tanto dal valore di $R_0$ ma dallo studio dell'intervallo seriale che risultando molto più piccolo della Sars ne rappresenta una pericolosità maggiore (sebben i valori di $R_0$ sono pressoché identici). Si è scoperto infatti che le infezioni si trasmettono molto velocemente, un individuo può infettarne un altro, a cui possono comparire i sintomi prima della loro manifestazione nella persona trasmittrice.
Tramite il valore di $R_0$ si è trovato inoltre che lo sviluppo dell'immunità di gregge richiede il 75% della popolazione infetta.

Un passaggio ulteriore è stato quello di riuscire, integrando regressione lineare e modello SIR, a trovare l'andamento del *basic reproduction number* giornaliero dopo i decreti, $R_t$, e a valutarne il suo effettivo trend decrescente prima dell'ingresso nella fase 2. Si è confermato quanto detto dalla protezione civile, ovvero della sua vicinanza a valori come 0.6/0.7 in prossimità del 4 maggio. L'indice sembra anche risalire negli ultimi giorni e se si vuole capire cosa succederà in futuro, bisogna parlare di previsione e intervalli di previsione, un lavoro che non è stato affrontato in questa tesi.

I risultati dei modelli SIR e SEIR hanno prodotto le medesimi conclusioni. All'inizio l'epidemia si è innescata in maniera molto lenta (contagi oscurati dal periodo di incubazione e alta presenza di infetti asintomatici) il che può spiegare le manovre tardive effettuate in Italia. La velocità di diffusione cresce in maniera rapida appena buona parte della popolazione diventa infetta. Tutti i modelli inoltre si allontanano dall'ipotesi di insorgenza della malattia nel periodo di febbraio, anticipandola almeno di 1 mese. Si presume dai risultati ottenuti che la malattia sia arrivata nel territorio Lombardo tra la metà di dicembre e la metà di gennaio ma non si esclude un suo approdo nel mese di novembre, essendo il valore di $\hat{R}_0$ molto variabile.

Dall'analisi del modello stocastico $SEII_aR$ ne deriva che gli spostamenti tramite mezzi pubblici delle persone facilitano la diffusione del virus. I più colpiti sono i paesi con alta concentrazione di trasporti e popolazione, non a caso l'incidenza più alta è ricaduta sulla provincia di Milano. Si è scoperto anche che meno collegamenti ci sono tra un paese ed un altro più la diffusione avviene lentamente. La densità di popolazione gioca un ruolo chiave nell'andamento degli infetti, in quando ne è direttamente proporzionale. La riduzione e la chiusura delle reti di trasporto sono un passo fondamentale per fermare e ridurre i contagi se tali misure vengono prese immediatamente. Il motivo è che per far partire un'epidemia in un determinato paese basta una sola persona infetta e lasciando la circolazione libera si ha maggior probabilità che gli individui infetti entrino in molteplici città. Questo modello, più complesso dei due precedenti ha confermato ancora una volta l'impossibilità del raggiungimento del picco d'infetti in un periodo di soli 2 mesi confermando i risultati di SIR e SEIR.

# Appendice

LIBRARY

```
#Theme set#####
library(ggplot2)
theme_set(theme_classic()+
            theme(axis.line = element_line(colour =
               "white")))

#Library#####
library(dvmisc)
library(plotrix)
library(incidence)
library(ggpubr)
library(gridExtra)
library(deSolve)
library(tidyverse)
library(gridExtra)
library(spread)
library(fhidata)
library(viridis)
library(gganimate)
library(tidyverse)
library(plotrix)
library(ggridges)
library(tidyverse)
library(ggplot2)
library(ggthemes)
library(dygraphs)
library(xts)
library(tm)
library(deSolve)
library(wesanderson)
```

# CAPITOLO 1

```
1
2  #WORLD SITUATION 24/03/2020######
3
4  confermati=read.csv("https://raw.githubusercontent.com/CSSEGI
5  SandData/COVID-19/master/csse_covid_19_data/csse_covid
6  _19_time_series/time_series_covid19_confirmed_global.csv")
7
8  ggplot(data = confermati,aes(x = Long, y =
      Lat,fill=X3.24.20)) +
9    borders("legacy_world",  colour =
       "grey60",fill="antiquewhite",alpha=0.5) +
10   scale_fill_distiller(type = 'seq', palette =
       "Spectral",direction=1)+
11   theme_map()+
12   geom_point(colour = "black", alpha =
       0.40,shape=21,aes(size=X3.24.20)) +
13   labs(title = "24/03/20
       Infected",fill="Infetti",size="Ampiezza")+
14   theme(panel.background = element_rect(fill = "aliceblue"))
15
16 #ITALY DYNAMIC SITUATION REGION#####
17
18 dat_csv<-read.csv("https://raw.githubusercontent.com
19 /pcm-dpc/COVID-19/master/dati-regioni/dpc-covid19-ita
20 -regioni.csv",header=T)
21 colnames(dat_csv)
22 dat_csv$data=as.Date(dat_csv$data)
23 italy_map <- map_data("italy")
24 print(unique(italy_map$region))
25
26 set.seed(1492)
27 choro_dat <- data_frame(region=unique(italy_map$    region),
28                         value=sample(100, length(region)))
29
30 italy_proj <- "+proj=aea +lat_1=38.15040684902542
31 +lat_2=44.925490198742295 +lon_0=12.7880859375"
32
33 ggplot(data = dat_csv,aes(x = long, y =
      lat,fill=totale_casi,size=totale_casi))+
34   geom_map(data=italy_map, map=italy_map,
35            aes(long, lat, map_id=region),
36            fill="antiquewhite",, size=0.4, color="grey")+
```

```
37   scale_fill_distiller(type = 'seq', palette =
        "Spectral",direction=1)+
38   geom_point(colour = "black", alpha = 0.90,shape=21) +
39   labs(title = "Date: {frame_time}", size =
        "Infetti",subtitle = "INFETTI ITALIA LIVELLO
        REGIONALE") +
40   theme(panel.background = element_rect(fill = "aliceblue"))+
41   transition_time(data)
42
43 #italy dynamic situation province
44
45 dat_csv=read.csv("https://raw.githubusercontent.com/pcm-dpc
46 /COVID-19/master/dati-province/dpc-covid19-ita-province.csv")
47 dat_csv=dat_csv %>%
48   filter(lat!=0 & long !=0)
49 myDateTimeStr1 <- as.vector(dat_csv$data)
50 myDateTimeStr1= gsub("T"," ",myDateTimeStr1)
51 myPOSIXct1 <- as.POSIXct(myDateTimeStr1, format="%Y-%m-%d")
52 dat_csv$data=myPOSIXct1
53 myPoSIXct1 =removeWords(as.character(dat_csv$data),"CET")
54 dat_csv$data=as.Date(myPoSIXct1)
55
56 ggplot(data = dat_csv,aes(x = long, y =
      lat,fill=totale_casi,size=totale_casi))+
57   geom_map(data=italy_map, map=italy_map,
58            aes(long, lat, map_id=region),
59            fill="antiquewhite",, size=0.4, color="grey")+
60   scale_fill_distiller(type = 'seq', palette =
        "Spectral",direction=1)+
61   geom_point(colour = "black", alpha = 0.90,shape=21) +
62   labs(title = "Date: {frame_time}", size =
        "Infetti",subtitle = "INFETTI ITALIA LIVELLO
        PROVINCIALE") +
63   theme(panel.background = element_rect(fill = "aliceblue"))+
64   transition_time(data)
65
66 #EPIDEMIC CURVE ITALY VS HUBEI #####
67
68 cases=read.csv("https://raw.githubusercontent.com/CSSEGISand
69 Data/COVID-19/master/csse_covid_19_data/csse_covid_19_time
70 _series/time_series_covid19_confirmed_global.csv")
71 death=read.csv("https://raw.githubusercontent.com/CSSEGISand
72 Data/COVID-19/master/csse_covid_19_data/csse_covid_19_time
73 _series/time_series_covid19_deaths_global.csv")
74 recovered=read.csv("https://raw.githubusercontent.com/CSSEGIS
```

```
andData/COVID-19/master/csse_covid_19_data/csse_covid_19_time
_series/time_series_covid19_recovered_global.csv")

region_confirmed=cases %>%
  gather(X1.22.20:X4.29.20,key="date",value="n") %>%
  filter(Province.State=="Hubei" | Country.Region=="Italy" )
     %>%
  select(Province.State,date,n)

region_death=death %>%
  gather(X1.22.20:X4.29.20,key="date",value="n_morti") %>%
  filter(Province.State=="Hubei" | Country.Region=="Italy" )
     %>%
  select(Province.State,date,n_morti)

region_recovered=recovered %>%
  gather(X1.22.20:X4.29.20,key="date",value="n_guariti") %>%
  filter(Province.State=="Hubei" | Country.Region=="Italy" )
     %>%
  select(Province.State,date,n_guariti)

region=inner_join(region_confirmed,region_death)
region=inner_join(region,region_recovered) %>%
  mutate(casi=n-n_morti-n_guariti) %>%
  select(Province.State,date,casi)

region$date=str_replace(region$date, "X", "0")
region$date=as.Date(region$date,format = c("%m.%d.%y"))
val=which(region$Province.State==c(""))
region$Province.State=as.character(region$Province.State)
region$Province.State[val]=c("Italia")

ggplot(region,aes(x=date,y=casi,group=Province.State,
fill=Province.State))+
  geom_density_line(stat = "identity", size=.7,
     alpha=0.75,colour="Blue4") +
  scale_fill_manual(values=c("Green4","Yellow3"))+
  labs(title="Curva epidemiologica",subtitle = "Casi nel
     tempo",y="Casi",x="Date",fill="Stati")+
  scale_x_date(breaks="7 days")+
  theme(axis.text.x = element_text(angle=45,hjust=1))+
  guides(colour = guide_legend(override.aes = list(size=1)))+
  theme(legend.justification=c(1,0),
     legend.position=c(0.6,0.65),
  legend.text=element_text(size=10))
```

```
114
115 #CINA-->CREATION HUBEI INCIDENCE######
116
117 data=read.csv("https://raw.githubusercontent.com/CSSEGISand
118 Data/COVID-19/master/csse_covid_19_data/csse_covid_19_time_
119 series/time_series_covid19_confirmed_global.csv")
120 colnames(data)
121
122 infected_a=data %>%
123   filter(Province.State=="Hubei") %>%
124   select(-c("Lat","Long","Country.Region","Province.State"))
125 infected_a=as.numeric(infected_a[1,])
126
127 #Incidence
128 incidence_1=vector()
129 for(i in 2:length(infected_a)){
130   incidence_1[i-1]=infected_a[i]-infected_a[i-1]
131 }
132 incidence_cina=c(incidence_1)
133 incidence_cina=incidence_cina[1:50]
134 time=length(incidence_cina)
135
136 #Incidence output
137
138 date=seq.Date(from = as.Date("2020-01-23"),length.out = time
    , by = "day")
139 length(date)
140 inc=data.frame(date=date,incidence=incidence_cina)
141 ince<-xts(inc,order.by = date)
142 ince=data.frame(date=ince$date,incidence=incidence_cina)
143 ince$incidence=as.numeric(ince$incidence)
144 ince$date=as.Date(ince$date)
145
146 picchi=ince\%>\%
147   summarize(massimo=max(incidence),
148   date=date[which((incidence)==max(incidence))])
149
150 g1=ggplot(ince,aes(x=date,y=incidence))+
151   geom_bar(stat="identity",fill="springgreen3",
152   col="white")+geom_line()+
153   geom_point(picchi,mapping=aes(x=date,y=massimo),
154   col="Red",size=3)+
155   theme(axis.text.x = element_text(angle=45,hjust=1))+
156   theme(legend.position = "None")+
157   labs(title = "Incidenza Cina-Hubei",y="Incidenza")+
```

```
158   scale_x_date(breaks="7 days")+
159   theme(axis.text.x = element_text(angle=45,hjust=1))
160
161 #Incidence with lag 3 days
162
163 input = 1:length(ince$incidence)
164 multiple_of_3 = (input \%\% 3) == 0
165 valori=input[multiple_of_3]
166
167 somma=function (x,len){
168   aggregate(x,by=list(rep(1:(length(x)/len),each=len)
        ),FUN=sum)
169 }
170
171 mobile=data.frame(incidence=somma(incidence_cina
172 [-c(49,50)],3),date=ince$date[valori])
173
174 picchi=mobile\%>\%
175   summarize(massimo=max(incidence.x),
176   date=date[which((incidence.x)==max(incidence.x))])
177
178 g2=ggplot(mobile,aes(x=date,y=incidence.x))+
179   geom_bar(stat="identity",fill="springgreen3",col="white")+
180   geom_line()+
181   geom_point(picchi,mapping=aes(x=date,y=massimo),
182   col="Red",size=3)+
183   theme(axis.text.x = element_text(angle=45,hjust=1))+
184   theme(legend.position = "None")+
185   labs(title = "Incidenza Cina-Hubei",subtitle = "Finestra
        mobile 3 giorni",y="Incidenza")+
186   scale_x_date(breaks="7 days")+
187   theme(axis.text.x = element_text(angle=45,hjust=1))
188
189 ggarrange(g1, g2,ncol=2, nrow=1)
190
191 #ITALY INCIDENCE#####
192
193 data=read.csv("https://raw.githubusercontent.com/CSSEGISand
194 Data/COVID-19/master/csse_covid_19_data/csse_covid_19_time_
195 series/time_series_covid19_confirmed_global.csv")
196 colnames(data)
197 infected=data \%>\%
198   filter(Country.Region=="Italy") \%>\%
199   select(-c("Lat","Long","Country.Region","Province.State"))
200 infected=as.numeric(infected[1,-c(1:29)])
```

```
infected

incidence_italia=vector()
for(i in 2:length(infected)){
  incidence_italia[i-1]=infected[i]-infected[i-1]}
length(incidence_italia)
incidence_italia

date=seq.Date(from = as.Date("2020-02-21"), length.out =
   length(incidence_italia), by = "day")
length(date)

inc=data.frame(date=date,incidence=incidence_italia)
ince<-xts(inc,order.by = date)

ince=data.frame(date=ince$date,incidence=incidence_italia)
ince$incidence=as.numeric(ince$incidence)
ince$date=as.Date(ince$date)

tamponi=vector()
for(i in 2:length(dat_csv$tamponi)){
  tamponi[i-1]=dat_csv$tamponi[i]-dat_csv$tamponi[i-1]}
tamponi

ince=cbind(ince[-c(1:4),],tamponi=tamponi)

picchi=ince\%>\%
  summarize(massimo=max(incidence),
  date=date[which((incidence)==max(incidence))])

g1=ggplot(ince,aes(x=date,y=incidence))+
  geom_bar(mapping=aes(fill=tamponi),stat="identity",
  col="white")+geom_line()+
  theme(axis.text.x = element_text(angle=45,hjust=1))+
  labs(title = "Incidenza
     Italia",fill="Tamponi",y="Incidenza")+
  scale_x_date(breaks="7 days")+
  geom_point(picchi,mapping=aes(x=date,y=massimo),
  col="Red",size=3)+
  theme(axis.text.x = element_text(angle=45,hjust=1))+
  scale_fill_viridis(discrete = F,option = "D")+
  theme(legend.justification=c(1,0),legend.position=c(0.95,0.4),
  legend.text=element_text(size=10))


```

```
#Incidence with lag 3 days

input = 1:length(ince$incidence)
multiple_of_3 = (input \%\\% 3) == 0
valori=input[multiple_of_3]

mobile=data.frame(incidence=somma(incidence_italia
[-c(76,77,78,79,80)],3),date=ince$date[valori],
tamponi=somma(tamponi[-c(76,77,78,79,80)],3))

picchi=mobile\%>\%
  summarize(massimo=max(incidence.x),
  date=date[which((incidence.x)==max(incidence.x))])

g2=ggplot(mobile,aes(x=date,y=incidence.x))+
  geom_bar(mapping=aes(fill=tamponi.x),stat="identity",
  col="white")+geom_line()+
  geom_point(picchi,mapping=aes(x=date,y=massimo),
  col="Red",size=3)+
  theme(axis.text.x = element_text(angle=45,hjust=1))+
  labs(title = "Incidenza Italia",subtitle = "Finestra
     mobile 3 giorni",fill="Tamponi",y="Incidenza")+
  scale_x_date(breaks="7 days")+
  theme(axis.text.x = element_text(angle=45,hjust=1))+
  scale_fill_viridis(discrete = F,option = "D")+
  theme(legend.justification=c(1,0),
     legend.position=c(0.95,0.4),
  legend.text=element_text(size=10))

ggarrange(g1, g2,ncol=2, nrow=1)

#Death analysis

#MORTI######

dat_csv<-read.csv("https://raw.githubusercontent.com/pcm-dpc
/COVID-19/master/dati-regioni/dpc-
covid19-ita-regioni.csv",header=T)
myDateTimeStr1 <- as.vector(dat_csv$data)
myDateTimeStr1= gsub("T"," ",myDateTimeStr1)
myPOSIXct1 <- as.POSIXct(myDateTimeStr1,
   format="\%Y-\%m-\%d")
dat_csv$data=myPOSIXct1
myPoSIXct1 =removeWords(as.character(dat_csv$data),"CET")
dat_csv$data=as.Date(myPoSIXct1)
```

```
286
287 dat_csv\%>\%
288   filter(data=="2020-05-29")\%>\%
289   summarize(morti=sum(deceduti),
290   guariti=sum(dimessi_guariti))\%>\%
291   mutate(tasso=(morti/guariti))
292
293 dat_csv\%>\%
294   summarize(positivi=sum(nuovi_positivi))
295
296 dat_csv\%>\%
297   filter(data=="2020-03-10")\%>\%
298   summarize(morti=sum(deceduti),
299   guariti=sum(dimessi_guariti))\%>\%
300   mutate(tasso=(morti/guariti))
301
302
303 dat_csv\%>\%
304   filter(data=="2020-05-10")\%>\%
305   summarize(morti=sum(deceduti),
306   guariti=sum(dimessi_guariti))\%>\%
307   mutate(tasso=(morti/guariti))
```

## CAPITOLO 2

```
#R0 ESTIMATION-CHINA#####

library(R0)
china=region%>%
  filter(Province.State=="Hubei")

GT = generation.time ("lognormal",c(4.7,2.6))
order(infected_a,decreasing = T)
est.R0.EG(epid = infected_a, GT, begin =1, end =15,reg.met =
   "poisson")
est.R0.ML(infected_a,GT,begin = 1,end=15)

#R0 ESTIMATION-CHINA#####

GT = generation.time ("lognormal",c(4.7,2.6))
order(incidence_italia,decreasing = T)
#same time china's period
est.R0.EG(epid = infected , GT, begin =1, end =15,reg.met =
   "poisson")
est.R0.ML(infected,GT,begin =1,end=15)
```

## CAPITOLO 3

```
#PARAMETERS CHOICE#####

R0=4
Incubazione=3 ##5 medi - 2 medi dove sono contagioso
Periodo_infetto=10  ##10 medi
gamma=1/Periodo_infetto #recovered rate
sigma=1/Incubazione
beta=gamma*R0
beta

#SIR MODEL LOMBARDY#####

sir <- function(time, state, parameters) {
    with(as.list(c(state, parameters)), {
    dS <- -beta * S * I
    dI <-  beta * S * I - gamma * I
    dR <-                 gamma * I
     return(list(c(dS, dI, dR)))
  })
}
```

```
21
22 N_Lombardia=10.04*10^6
23
24 #Input
25
26 init<-c(S =(N_Lombardia-20-1)/N_Lombardia,I
      =20/N_Lombardia,R =0/N_Lombardia)
27 parameters <- c(beta = beta, gamma = gamma)
28 times <- seq(0, 200, by = 1)
29 out <- ode(y = init, times = times, func = sir, parms =
      parameters)
30 out <- as.data.frame(out)
31 out$I=out$I
32 out$R=out$R
33 out$time <- NULL
34 head(out, 10)
35
36 out
37 title <- bquote("SIR MODEL")
38 subtit <- bquote(list(beta==.(parameters[1]),
39 ~gamma==.(parameters[2])))
40
41 sir<-ggplot(out,aes(x=0:200))+
42   ggtitle(bquote(atop(.(title),atop(bold(.(subtit))))))+
43   geom_line(aes(y=S,colour="Suscettibili"))+
44   geom_line(aes(y=I,colour="Infetti"))+
45   geom_line(aes(y=R,colour="Guariti"))+
46   ylab(label="Proporzione pop.")+
47   xlab(label="Giorni")+
48   labs(color="Legenda : ")+
49   theme(legend.justification=c(1,0),
        legend.position=c(1,0.3))
50 sir
51
52 #R0 VARIATION IN LOMBARDY#####
53
54 detach("package:R0", unload=TRUE)
55 library(tidyverse)
56
57 dat_csv<-read.csv("https://raw.githubusercontent.com/pcm-dpc
58 /COVID-19/master/dati-regioni/dpc-covid19-ita-regioni.csv",
59 header=T)
60 colnames(dat_csv)
61 dat_csv=dat_csv %>%
62   filter(denominazione_regione=="Lombardia")
```

```
63
64 days<-dim(dat_csv)[1]
65 colnames(dat_csv)
66 dat_csv$t<-1:days
67 days
68 myDateTimeStr1 <- dat_csv$data
69 myDateTimeStr1= gsub("T"," ",myDateTimeStr1)
70 myPOSIXct1 <- as.POSIXct(myDateTimeStr1, format="%Y-%m-%d
      %H:%M:%S")
71 days_dy<-as.Date(myPOSIXct1)
72
73 #need to update manually
74 dat_csv_dy<-xts(dat_csv[,-c(1:6,18:20)], order.by = days_dy,
      frequency = 7)
75 length(dat_csv_dy$ricoverati_con_sintomi)
76 gruppi=c(rep(1,10),rep(2,10),rep(3,10),rep(4,10),rep(5,10)
77 ,rep(6,10),rep(7,10))
78 dat_csv_dy=cbind(dat_csv_dy,log=log(as.numeric
79 (dat_csv_dy$totale_positivi)),as.numeric(gruppi))
80
81 p1=ggplot(dat_csv_dy)+
82   geom_point(aes(x=t,y=log),size=1)+
83   scale_color_brewer(palette="Dark2")+
84   geom_line(aes(x=t,y=log,group=gruppi,col=as.factor(gruppi)),
85   stat="smooth",method="lm",se = F,alpha=1,size=1)+
86   theme(legend.position = "none")+
87   labs(y="Log_tot_positivi",x="Tempo",title = "Log-Lineare")
88
89 p2=ggplot(dat_csv_dy)+
90   geom_point(aes(x=t,y=totale_positivi),size=1)+
91   scale_color_brewer(palette="Dark2")+
92   geom_smooth(aes(x=t,y=totale_positivi,group=gruppi,
93   col=as.factor(gruppi)),method="glm",method.args =
        list(family = "poisson"),se = F,size=1)+
94   theme(legend.position = "none")+
95   labs(y="Tot_positivi",x="Tempo",title="Poisson")
96
97 grid.arrange(p1,p2,ncol=2)
98
99 duration<-15
100 beta_vec<-NULL
101 sd_vec<-NULL
102 mse_POIS=NULL
103 mse_LM=NULL
104 mse_LOGLM=NULL
```

```
105
106 tamponi=vector()
107 for(i in 2:length(dat_csv$tamponi)){
108   tamponi[i]=dat_csv$tamponi[i]-dat_csv$tamponi[i-1]}
109 tamponi[1]=dat_csv$tamponi[1]
110 tamponi[3]=492
111 dat_csv$tamponi=tamponi
112
113 for (i in 3:(days-2)){
114   fit <- glm((totale_positivi)~t,weights =
        1/tamponi,family="poisson",data=dat_csv[(i-2):(i+2),])
115   mse_POIS=c(mse_POIS,get_mse(fit))
116   fit <- glm((totale_positivi)~t,weights =
        1/tamponi,family="gaussian",data=dat_csv[(i-2):(i+2),])
117   mse_LM=c(mse_LM,get_mse(fit))
118   fit <- glm(log(totale_positivi)~t,weights =
        1/tamponi,family="gaussian",data=dat_csv[(i-2):(i+2),])
119   mse_LOGLM=c(mse_POIS,get_mse(fit))
120   beta_vec<-c(beta_vec,coef(fit)[2])
121   sd_vec<-c(sd_vec,coef(summary(fit))[2,2])
122 }
123 mean(mse_LM)
124 mean(mse_LOGLM) #the best
125 mean(mse_POIS)
126
127 label<-as.Date(substr(dat_csv$data,1,10))[3:(days-2)]
128 mean  <- 1+(beta_vec*duration)
129 lower <- 1+((beta_vec-1.96*sd_vec)*duration)
130 upper <- 1+((beta_vec+1.96*sd_vec)*duration)
131 df <- data.frame(label, mean, lower,
      upper,tamponi=tamponi[-c(1,2,length(mean),length(mean)-1)])
132 fp <- ggplot(data=df, aes(x=label, y=mean, ymin=lower,
      ymax=upper,col=tamponi)) +
133   geom_pointrange(size=0.2) +
134   scale_color_gradientn(colours = rainbow(5))+
135   geom_hline(yintercept=1,
        size=0.5,color="black",linetype="dashed") +
136   xlab("Date") + ylab("Rt")+
137   labs(title = "Lombardia",subtitle = "Intervallo di
        confidenza 95%",col="Tamponi" )+
138   scale_x_date(breaks="7 days")+
139   theme(axis.text.x = element_text(angle=45,hjust=1))+
140   theme(legend.justification=c(1,0),
        legend.position=c(1,0.3))
141 fp
```

```
142
143 #MODELLO SEIR LOMBARDY#####
144
145 seir_model = function (current_timepoint, state_values,
      parameters){
146   S = state_values [1]        # susceptibles
147   E = state_values [2]        # exposed
148   I = state_values [3]        # infectious
149   R = state_values [4]        # recovered
150   with (
151     as.list (parameters),{
152       dS = (-beta * S * I)
153       dE = (beta * S * I) - (delta * E)
154       dI = (delta * E) - (gamma * I)
155       dR = (gamma * I)
156       results = c (dS, dE, dI, dR)
157       list (results)
158     }
159   )
160 }
161
162 parameter_list = c (beta = beta, gamma = gamma, delta=sigma)
163
164 W = N_Lombardia-1
165 Y = 0
166 Z = 0
167 X = 1
168 N = W + X + Y + Z
169
170 initial_values = c (S = W/N, E = X/N, I = Y/N, R = Z/N)
171 timepoints = seq (0, 300, by=1)
172 output = lsoda (initial_values, timepoints, seir_model,
      parameter_list)
173
174 title <- bquote("SEIR MODEL")
175 subtit <- bquote(list(beta==.(parameter_list[1]),
176 ~gamma==.(parameter_list[2]),~delta==.(parameter_list[3])))
177 output=as.data.frame(output)
178
179 seir<-ggplot(output,aes(x=time))+
180   ggtitle(bquote(atop(.(title),atop(bold(.(subtit))))))+
181   geom_line(aes(y=S,colour="Suscettibili"))+
182   geom_line(aes(y=I,colour="Infetti"))+
183   geom_line(aes(y=R,colour="Guariti"))+
184   geom_line(aes(y=E,colour="Esposti"))+
```

```
  ylab(label="Proporzione pop.")+
  xlab(label="Giorni")+
  labs(color="Legenda : ")+
theme(legend.justification=c(1,0), legend.position=c(1,0.15))

seir
```

## CAPITOLO 4

```
# OD MATRIX-GENERAL ANALYSIS#############

matrice_od <- read.csv("C:/Users/matte/Desktop/Lavori
   tesi/File_Excel/matrice_od.csv")
od=matrice_od

val=od%>%
  filter(ZONA_ORIG!=ZONA_DEST) %>%
  summarise_each(funs(sum),-c(ZONA_ORIG,PROV_ORIG,PROV_DEST,
  ZONA_DEST,FASCIA_ORARIA))%>%
  gather(LAV_COND:RIT_ALTRO,key="Tipo",value="N")
for(i in 1:length(val$N)){
  val$N[i]=round(val$N[i],0)
}

val%>%
  filter(N>100000)%>%
  mutate(pct = round(prop.table(N),3))%>%
  ggplot(aes(x=N,y=Tipo,fill=N))+
  geom_bar(stat="Identity")+
  scale_fill_distiller(type = 'seq', palette =
     "Spectral",direction=1)+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),legend.position
           ="NULL")+
  geom_text(aes(label=scales::percent(pct)), angle = 360,
            position=position_dodge(width=0.2), vjust=0.2)+
  labs(y="Tipo di spostamento")

val$Tipo
gruppi=c(rep("LAVORO",8),rep("STUDIO",8),rep("OCCASIONALI",8),
rep("AFFARI",8),rep("RIENTRO",8))
new_val=cbind(val,gruppi)

```

```
new_val%>%
  filter(gruppi!="RIENTRO")%>%
  group_by(gruppi)%>%
  summarize(N=sum(N))%>%
  mutate(pct = prop.table(N))%>%
  ggplot(aes(x=gruppi,y=N,fill=N))+
  geom_bar(stat="Identity")+
  scale_fill_distiller(palette =4)+
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),legend.position
           ="NULL")+
  geom_text(aes(label=scales::percent(pct)),
  position=position_dodge(width=0.5), vjust=0.9)+
  labs(title ="Motivo di spostamento",x="Motivo")

val$Tipo
gruppi=gsub("[^;]*_(.*)", "\\1", val$Tipo)
new_val_2=cbind(val,gruppi)

new_val_2=new_val_2%>%
  group_by(gruppi)%>%
  summarize(N=sum(N))

new_val_2$gruppi=as.character(new_val_2$gruppi)
new_val_2$gruppi[3]=c("AUTO")
new_val_2$gruppi[4]=c("TRENO/METRO/TRAM")
new_val_2$gruppi[5]=c("BUS/PULLMAN")
new_val_2$gruppi[7]=c("PASSEGGERO AUTO")

fig <- plot_ly(new_val_2, labels = ~gruppi, values = ~N,
   type = 'pie')
fig <- fig %>% layout(xaxis = list(showgrid = T, zeroline =
   FALSE, showticklabels = FALSE),
                      yaxis = list(showgrid = T, zeroline =
                         FALSE, showticklabels = FALSE))

fig


#CLEAN OD_MATRIX TO BUILDING PUBLIC TRANSPORT MATRIX######

matrix_od=od %>%
  filter(ZONA_ORIG!=ZONA_DEST) %>%
  select(ZONA_ORIG,PROV_ORIG,ZONA_DEST,PROV_DEST,LAV_FERRO,
```

```
74    LAV_GOMMA,OCC_FERRO,OCC_GOMMA,RIT_GOMMA,
75    RIT_FERRO,STU_GOMMA,STU_FERRO) %>%
76    mutate(all=LAV_FERRO+LAV_GOMMA+OCC_FERRO+OCC_GOMMA+
77    RIT_GOMMA+RIT_FERRO+STU_GOMMA+STU_FERRO) %>%
78    select(ZONA_ORIG,PROV_ORIG,ZONA_DEST,PROV_DEST,all) %>%
79    filter(all>0)
80
81  #ripetitions eliminated
82
83  new_file=aggregate(matrix_od$all,
       by=list(matrix_od$ZONA_ORIG,matrix_od$ZONA_DEST), FUN=sum)
84
85  new_file=matrix_od%>%
86    group_by(ZONA_ORIG,ZONA_DEST,PROV_ORIG,PROV_DEST) %>%
87    summarize(all=sum(all))%>%
88    rename(from=ZONA_ORIG,to=ZONA_DEST,n=all)
89
90  #eliminated number in strings
91  new_file$from=as.character(new_file$from)
92  new_file$to=as.character(new_file$to)
93
94  split=function(val){
95    x = unlist(strsplit(val, split = '\\s+'))
96    val=paste0(x[Reduce('|', lapply(c('[A-Za-z]', '67'),
         grepl, x))], collapse = ' ')
97    return(val)
98  }
99
100 for(i in 1:length(new_file$from)){
101   new_file$from[i]=split(new_file$from[i])
102   new_file$to[i]=split(new_file$to[i])
103 }
104
105 new_file$from=as.factor(new_file$from)
106 new_file$to=as.factor(new_file$to)
107 costruzione_province=new_file
108
109 new_file$from=as.character(new_file$from)
110 new_file$to=as.character(new_file$to)
111 new_file=new_file[,c(1:2,5)]
112 new_file=new_file %>%
113   filter(from!=to)
114
115 new_file$from=as.factor(new_file$from)
116 new_file$to=as.factor(new_file$to)
```

```
117
118 new_file=aggregate(new_file$n,
       by=list(new_file$from,new_file$to), FUN=sum)
119 colnames(new_file)=c("from","to","n")
120
121
122
123 #write.csv(new_file,file = "OD.csv")
```

# CAPITOLO 5

```
#DATAFRAME WITH CITY AND POPULATION######

com<-read.csv("C:/Users/matte/Desktop/Lavori_tesi/File_excel
/11125_comuni Popolazione-residente-totale-per-classe-di-eta-
Totali-al-11---Comun.csv", sep=";")
popolazione=as_tibble(com)
colnames(popolazione)

popolazione=popolazione %>%
  filter(Anno==2019 & Livello.territoriale=="Comune") %>%
  select(Territorio,Totale) %>%
  transmute(Territorio=toupper(Territorio),Totale=Totale)

citta=levels(new_file$from)
pop=vector()
v=0
for(i in citta){
  v=v+1
  if(i %in% popolazione$Territorio){
    a=which(popolazione$Territorio==i)
    a=a[length(a)]
    pop[v]=popolazione$Totale[a]}
  else {
    pop[v]="Na"
  }
}
length(citta)
length(pop)

regioni=c("LOMBARDIA","LAZIO","CAMPANIA","SICILIA","VENETO",
"EMILIA-ROMAGNA","PIEMONTE","PUGLIA","TOSCANA","CALABRIA",
"SARDEGNA","LIGURIA","MARCHE","ABRUZZO","FRIULI VENEZIA
   GIULIA",
"TRENTINO ALTO ADIGE","UMBRIA","BASILICATA","MOLISE","VALLE
   D'AOSTA","ASTI","BIELLA","BOLOGNA","BOLZANO","CANTU",
"CASSANO D'ADDA","CASTIGLIONE D'ADDA","CUNEO","FERRARA",
"FORLICESENA","MODENA","MUGGIO'","NOVARA","PADOVA","PARMA",
"PEREGO","PIACENZA","RAVENNA","REVERE","RIMINI","TORINO",
"TRENTO","TREVISO","VENEZIA","VICENZA")
regioni_pop=c(10018806,5898124,5839084,5056641,5056641,
4448841,4392526,4063888,3742437,1965128,1653135,1565307,
1538055,1322247,1217872,1062860,888908,570365,310449,126883,
76164,44616,388367,106951,40007,18911,4665,56124,132009,
```

```
117946,184727,23579,104284,936887,194417,1757,102355,159057,
2508,148908,886837,117417,885447,261905,112198)

for(i in 1:length(regioni)){
  val=which(citta==regioni[i])
  pop[val]=regioni_pop[i]
}

for(i in 1:length(citta)){
  val=which(pop=="Na")
  pop[val]=1500
}

#STARTED VALUES#####

E=rep(0,length(pop))
R=rep(0,length(pop))
I=rep(0,length(pop))
Ia=rep(0,length(pop))
citta_completo=data.frame(location_code=citta,
S=as.numeric(pop),
E=E,I=I,Ia=Ia,R=R)

#MODEL#####

#STARTED VALUES FROM MILANO

which(citta_completo$location_code=="MILANO")
citta_completo[830,]
citta_completo[830,]$I=1
citta_completo[830,]$R=0
citta_completo[830,]$E=0
citta_completo[830,]
citta_completo
which(new_file$from=="TOSCANA" & new_file$to=="AFRICA")
new_file=new_file[-423,]

set.seed(123)
d <- commuter(
  seiiar=citta_completo,
  commuters=new_file,
  r0=4.0,
  latent_period = 3,
  infectious_period = 12,
  asymptomatic_prob=2/3,
```

```
  asymptomatic_relative_infectiousness=0.10,
  days_simulation=200,
  N=1
)
a=which(d$location_code=="MILANO")
d[a]

#BUILD A PROVINCE/CITY DATAFRAME#####

location_code=levels(citta_completo$location_code)
provincia=vector()
v=0

for(i in location_code){
  v=v+1
  a=which(costruzione_province$from==i)
  a=a[1]
  provincia[v]=as.character(costruzione_province$PROV_ORIG[a])
}

provincia=replace(provincia, provincia=="BG", "BERGAMO")
provincia=replace(provincia, provincia=="BS", "BRESCIA")
provincia=replace(provincia, provincia=="CO", "COMO")
provincia=replace(provincia, provincia=="CR", "CREMONA")
provincia=replace(provincia, provincia=="LC", "LECCO")
provincia=replace(provincia, provincia=="LO", "LODI")
provincia=replace(provincia, provincia=="MN", "MANTOVA")
provincia=replace(provincia, provincia=="MI", "MILANO")
provincia=replace(provincia, provincia=="MB", "MONZA")
provincia=replace(provincia, provincia=="PV", "PAVIA")
provincia=replace(provincia, provincia=="SO", "SONDRIO")
provincia=replace(provincia, provincia=="VA", "VARESE")

province=data.frame(location_code=location_code,
provincia=provincia)

##MODEL,BUILD WITH PROVINCE#########################

d <- merge(d,province,
   by.x="location_code",by.y="location_code")
county <- d[,.(
  S=sum(S),
  E=sum(E),
  I=sum(I),
  Ia=sum(Ia),
```

```
  R=sum(R),
  incidence=sum(incidence),
  pop=sum(pop)
),
keyby=.(provincia,location_code,week,day,is_6pm)]

#select only lombardy's province
lombardia=which(county$provincia=="BERGAMO" |
   county$provincia=="BRESCIA" | county$provincia=="COMO" |
   county$provincia=="CREMONA" | county$provincia=="LECCO" |
   county$provincia=="LODI" | county$provincia=="MANTOVA" |
   county$provincia=="MILANO" | county$provincia=="MONZA" |
   county$provincia=="CREMONA" | county$provincia=="PAVIA" |
   county$provincia=="SONDRIO" | county$provincia=="VARESE")

county_2=county[lombardia,]
colnames(county_2)

#write.csv2(county_2,file = "infetti lombardia.csv")

#model output##########

pop_province=county_2%>%
  group_by(provincia)%>%
  filter(day==1)%>%
  summarize(pop=sum(S))


somme=county_2%>%
  group_by(provincia,day)%>%
  summarize(S=sum(S),I=sum(I),R=sum(R),E=sum(E),Ia=sum(Ia))

popprov=left_join(somme,pop_province)

somme=popprov%>%
  mutate(Ss=S/pop,Ii=I/pop,Rr=R/pop,Er=E/pop,Iaa=Ia/sum(Ia))

ggplot(somme,aes(x=day))+
  geom_line(aes(y=Ss,colour="Suscettibili"))+
  geom_line(aes(y=Ii,colour="Infetti"))+
  geom_line(aes(y=Rr,colour="Guariti"))+
  geom_line(aes(y=Er,colour="Esposti"))+
  geom_line(aes(y=Iaa,colour="Asintomatici"))+
  facet_wrap(~provincia)+
  ylab(label="Proporzione pop.")+
```

```
170   xlab(label="Giorni")+
171   labs(color="Legenda : ")+
172   theme(legend.justification=c(2,0),
         legend.position=c(1,0.3))
173
174
175 colnames(county_2)
176 p <- ggplot(county_2, aes(x=day, y=incidence))+
177   geom_col()+
178   facet_wrap(~provincia)+
179   labs(x="Giorni",y="Curva incidence")+
180   theme(legend.position = "None")+
181   theme(axis.text.x = element_text(angle=45,hjust=1))
182 p
183 #Save plot
184 ggsave(p,device = "png",width = 10,height = 7,filename =
     "Andamento incidence.png")
```

# Bibliografia


[1] Treccani. Le parole del coronavirus. URL: http://www.treccani.it/magazine/parolevalgono/Le_parole_del_Coronavirus/index.html.

[2] Farr William. Causes of death in england and wales. *Second Annual Report of the Registrar General of Births, Deaths and Marriages in England*, 2:69–98, 1927.

[3] Sylvia Richardson. Coronavirus statistics: what can we trust and what should we ignore? *The Guardian*, 2020. URL: https://www.theguardian.com/world/2020/apr/12/coronavirus-statistics-what-can-we-trust-and-what-should-we-ignore.

[4] Regione Lomardia Sistema Socio Sanitario ASST Garda. Esecuzione tamponi covid-19 per accertamento guarigione, 2020. URL: https://www.asst-garda.it/notizie/tamponi-covid-19-di-controllo/.

[5] Istituto Superiore di Sanità (ISS). Caratteristiche dei pazienti deceduti positivi all'infezione da sars-cov-2 in italia. *Science*, 2020. URL: https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia.

[6] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[7] StatGroup-19. Meme ed epidemiologia, 2020. URL: https://statgroup-19.blogspot.com/.

[8] Andrei R. Akhmetzhanova Hiroshi Nishiuraa, Natalie M. Lintona. Serial interval of novel coronavirus (Covid-19) infections. *International Journal of Infectious Diseases*, 93:284–286, 2020.

[9] Anderson RM. *Infectious diseases of humans: dynamics and control*. Oxford, may r.m. oxford university press edition, 1991.

[10] Pagano M. White LF. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Stat med.*, 27(16):2999–3016, 2008.

[11] Donnelly P. Ball F. Strong approximations for epidemic models-stochastic processes and their applications, 1995.

[12] Istituto Superiore di Sanità (ISS). Che cos'è r0 e perché è così importante. *ISS*, 2020. URL: https://www.iss.it/primo-piano/-/asset_publisher/o4oGR9qmvUz9/content/id/5268851.

[13] Annals of Internal Medicine. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application, 2020. URL: https://www.acpjournals.org/doi/10.7326/M20-0504.

[14] Bai-Zhong Liu Quan-Xin Long. Antibody responses to sars-cov-2 in patients with covid-19. *Nature Medicine*, 26:845–848, 2020.

[15] Johns Hopkins University. New study on covid-19. *Bloomberg School of public health*, 2020. URL: https://www.jhsph.edu/news/news-releases/2020/new-study-on-COVID-19-estimates-5-days-for-incubation-period.html.

[16] Regione Lombardia-Open Data. La matrice di dati Origin Destination, 2019. URL: https://www.dati.lombardia.it/Mobilit-e-trasporti/Matrice-OD2016-Passeggeri/tezw-ewgk.

[17] L'enciclopedia libera Wikipedia. Distribuzione multinomiale, 2019. URL: https://it.wikipedia.org/wiki/Distribuzione_multinomiale.

[18] Sergio Romagnani. Professore ordinario di Immunologia clinica dell'Università di Firenze, 2020. URL: https://www.repubblica.it/salute/medicina-e-ricerca/2020/03/16/news/coronavirus_studio_il_50-75_dei_casi_a_vo_sono_asintomatici_e_molto_contagiosi-251474302/.

[19] Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368:489–493, 2020.