Analisi Esplorativa ECDC COVID-19 Dataset

Tesina di *Data Science for Health Systems*

Andrea Baldinelli
Università degli Studi di Perugia
Mat: 334568
andrea.baldinelli@studenti.unipg.it

Michele Baglioni
Università degli Studi di Perugia
Mat: 330085
michele.baglioni1@studenti.unipg.it

Eraldo Marku
Università degli Studi di Perugia
Mat: 334567
eraldo.marku@studenti.unipg.it

***Sommario*—Questo documento contiene i risultati dell'analisi esplorativa effettuata sui dati pubblicati dall' *European Centre for Disease Prevention and Control*. Analisi che si è soffermata sull'andamento dei dati per Italia e Svezia, prima e dopo l'inizio della campagna vaccinale.**

***Index Terms*—EDA, COVID-19, Italia, Svezia.**

I. Introduzione

La pandemia di COVID-19 ha impattato sulle vite di tutti, portando a mettere in discussione anche le abitudini più comuni. D'altro canto è emerso anche quanto siano fondamentali i dati, per cercare di comprendere il virus al fine di disporre misure appropriate per contenere la malattia.

In questo studio, si esplorano i dati che sono stati raccolti in questi due anni di pandemia, andando a contrapporre i dati di due nazioni che hanno adottato misure diverse per contrastare la malattia, l'Italia e la Svezia, in relazione anche all'avvento dei vaccini.

II. Dati

I dati utilizzati in questo studio provengono dal ***European Centre for Disease Prevention and Control***: un'agenzia dell'Unione Europea il cui obiettivo è rafforzare le difese dell'Unione contro le malattie infettive. Tra le tante cose di cui si occupa, c'è quello di raccogliere e raggruppare dati, sulle malattie, provenienti dalle singole nazioni.

Nel dettaglio, il tutto è stato effettuato lavorando con i seguenti dataset:

- [1] **ECDC European Centre for Disease Prevention and Control, Covid-19 Data vaccination in the EU/EEA**: raggruppa le informazioni, raccolte dal *The European Surveillance System, TESSy*, sulle vaccinazioni e i vaccini per ogni nazione dell'Unione, con cadenza settimanale. Riporta le informazioni su: il numero di dosi consegnate; il numero di dosi donate ad altri paesi; il numero di prime e seconde dosi di vaccino somministrate, più quelle addizionali. In più riporta informazioni sulle vaccinazioni somministrate per fasce d'età, per regione e per tipo di vaccino.
- [2] **ECDC European Centre for Disease Prevention and Control, Covid-19 Data on country response measures to COVID-19**: raggruppa le informazioni raccolte da fonti pubbliche, in alcuni casi interne, di ciascuno stato sulle restrizioni che ciascun paese ha applicato, indicando la data di inizio e fine di ciascuna restrizione.
- [3] **ECDC European Centre for Disease Prevention and Control, Data on the daily number of new reported COVID-19 cases and deaths by EU/EEA country**: raccoglie informazioni giornaliere, provenienti da bollettini rilasciati dal governo di ciascuno stato, su nuovi casi e nuovi morti.
- [4] **ECDC European Centre for Disease Prevention and Control, Data on testing for COVID-19 by week and country**: riporta il numero di tamponi effettuati, il numero di nuovi positivi, il tasso di test effettuati ed il tasso di positività; il tutto su base settimanale, con la suddivisione dei dati a livello nazionale e regionale. La fonte di queste informazioni non è univoca: ove possibile la fonte è sempre TESSy, in alternativa si compensa con fonti personali di ciascuno stato (come siti nazionali o *repository publici*.
- [5] **ECDC European Centre for Disease Prevention and Control, Data on hospital and ICU admission rates and current occupancy for COVID-19**: raccoglie informazioni su l'occupazione delle terapie intensive ed il numero di ospedalizzati. Il tipo di informazione può variare a seconda del paese dove sono stati raccolti i dati, in quanto per alcune nazioni viene riportato: l'occupazione giornaliera delle terapie intensive e dei posti letto oltre ai nuovi ingressi settimanali in area medica e in terapia intensiva per centomila abitanti; mentre altre nazioni l'informazione riportata è solo parziale. Un esempio è l'Italia, in quanto riporta solo il numero settimanale di nuovi ingressi in area medica.

III. Costruzione del Dataset

Durante una prima ispezione dei dataset, si è subito notato come la politica di campionamento delle informazioni non fosse uniforme fra tutti. Ad esempio, all'interno del dataset contenente le misure restrittive applicate da ogni nazione, non abbiamo una vera e propria cadenza regolare; infatti le misure vengono inserite solo quando effettivamente vengono attuate. Un altro esempio ancora è il dataset contenenti i morti da COVID-19, esso riporta le informazioni con cadenza giornaliera, mentre tutti gli altri con cadenza settimanale. Pertanto è stato necessario andare a normalizzare la politica di campiona-

mento fra le varie sorgenti di informazione, scegliendo quella settimanale. In tutti i dataset è stata effettuata la conversione delle date in standard **ISO 8601** e, dove necessario, sono stati raggruppati i dati giornalieri in modo da ottenere il valore settimanale dell'informazione.

Un'altra evidente incongruenza di formato è stata generata dal processo di inserimento dati nei campi relativi al nome della nazione poichè, in alcuni dataset veniva riportato il nome completo e in altri invece diversi standard ISO appositi. Si è deciso di convertire tutti i valori nel formato definito dallo standard **ISO 3166-1 alpha-2**, che riporta i nomi come codici univoci a due caratteri.

Alcuni dataset riportano dati sia a livello nazionale, sia filtrati a livello regionale. In questo studio è stato eseguito un filtraggio che escludesse tutti i dati che venivano presentati come *regionali* lasciando così solo le informazioni a livello nazionale.

Per quanto riguarda il valore delle dosi di vaccino somministrate settimanalmente, ci siamo basati sul numero assoluto di dosi effettuate su tutta la popolazione, senza discriminare sul tipo di vaccino inoculato o sul numero di prime, seconde o terze dosi effettuate.

Dopo aver effettuato queste operazioni, mantenendo solo le *features* utili all'analisi, è stato eseguito un *merge* dei dataset in base alla coppia di valori costituita dalla data e dal codice della nazione.

Da notare che ci sono due dataset che forniscono i casi di Covid-19. É stato scelto di mantenere come sorgente il dataset [4], in quanto si è verificato tramite il test adattamento di Kolmogorov-Smirnov che i campioni appartengono alla stessa distribuzione, con una statistica di test $D = 0.008$ ed un $p - value = 0.9985$. Un'ulteriore evidenza contro la sorgente rimossa è data dalla presenza di alcuni valori nulli nel campione.

L'ultimo passo è stato quello di filtrare questo dataset parziale rispetto alle due nazioni scelte in esame: **Italia**, indicata dal codice **IT**, e **Svezia**, indicata dal codice **SE**.

Il risultato di questa serie di operazioni effettuate è un nuovo dataset così composto:

- **year_week**: anno e settimana dell'anno in formato ISO 8601
- **country_code**: codice della nazione in formato ISO 3166-1 alpha-2
- **cases**: numero di casi positivi settimanali. Sorgente dei dati è il dataset [3]
- **positivity_rate**: tasso di positività settimanale. Informazione proveniente dal dataset [4]
- **deaths**: numero di morti settimanali. Informazione proveniente dal dataset [4]
- **doses**: numero di dosi inoculate per settimana. Proveniente dal dataset [1]
- **hospitalizations**: nuovi ingressi settimanali in area medica e/o in terapia intensiva per centomila abitanti
- **active_restrictions**: lista di restrizioni applicate per la settimana in esame. Tale lista è stata ricavata andando a fare un operazione di join fra questo nuovo dataset e il dataset [2]

Questo è il dataset che si è andati ad analizzare nella seguente fase.

IV. Analisi Esplorativa

L'analisi esplorativa del dataset è composta da 3 parti distinte: in un primo momento si è effettuata un'analisi delle distribuzioni dei valori nel campione e test di normalità; successivamente sono state confrontate Italia e Svezia durante la prima ondata della pandemia, ed infine il confronto è stato effettuato dopo l'avvento dei vaccini.

Al fine di facilitare la comprensione dell'analisi esplorativa e il confronto diretto, il dataset è stato riorganizzato nel seguente modo: invece di lavorare con un'osservazione settimanale che può avere come country_code sia *IT* che *SE*, si è preferito definire osservazioni in modo da allineare le informazioni di Italia e Svezia della stessa settimana.

A. Analisi preliminare

Come primo passo, si è andati ad osservare la media e la deviazione standard campionaria delle variabili numeriche presenti, per Italia e per Svezia. I risultati sono riportati nella Tabella I.

Nazione	Feature	Media	Incertezza tipo
Italia	Casi	162215.3	256988.9
	Morti	1725.253	1450.057
	Tasso di positività	6.770878	5.796835
	Ospedalizzazioni	10.13138	7.670398
Svezia	Casi	26768.04	46139.96
	Morti	189.9341	199.1179
	Tasso di positività	12.08984	10.47827
	Terapie intensive	0.9022031	0.7572622

Tabella I: Tabella dei momenti.

Successivamente, nonostante i dati appartengano ad un fenomeno con andamento esponenziale, abbiamo condotto dei test di adattamento per controllare la normalità dei dati. Per prima cosa abbiamo utilizzato il metodo dei momenti: è stata creata una distribuzione di probabilità gaussiana avente come parametri il valor medio e la deviazione standard campionaria delle rispettive features, ed è stata confrontata con l'istogramma generato dai dati stessi. I risultati sono riportati in Figura 1 e in Figura 2.

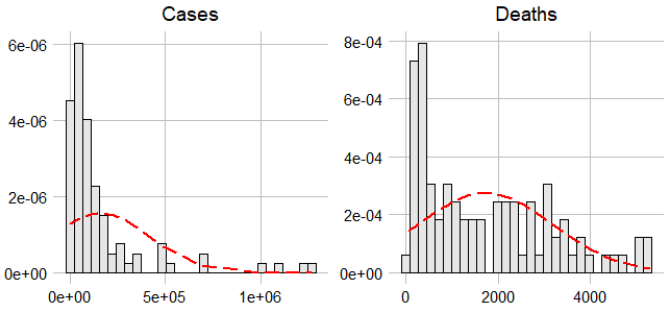

Figura 1: Metodo dei momenti applicato alla distribuzione delle osservazioni di morti e casi in Italia.

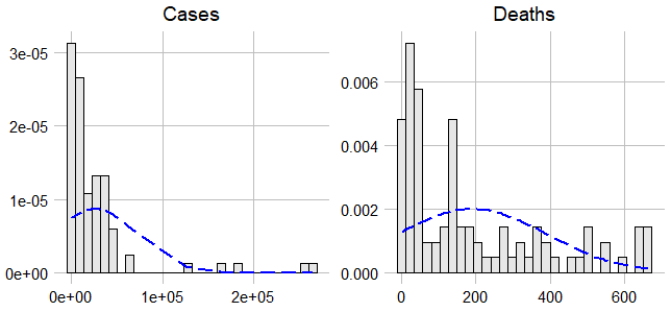

Figura 2: Metodo dei momenti applicato alla distribuzione delle osservazioni di morti e casi in Svezia.

Come controprova di un risultato già anticipatamente annunciato, è stato effettuato il test di Shapiro-Wilk, che ha confermato ancora una volta, come mostrato nella Tabella II, che i dati non sono stati generati da un processo normale.

Nazione	Feature	Statistica W	p-value
Italia	Casi	0.59586	1.899e-14
	Morti	0.89902	3.287e-06
	Tasso di positività	0.88276	6.603e-07
	Ospedalizzazioni	0.91803	2.595e-05
Svezia	Casi	0.50871	6.29e-16
	Morti	0.82839	6.834e-09
	Tasso di positività	0.7952	6.489e-10
	Terapie intensive	0.89066	1.416e-06

Tabella II: Test di Shapiro-Wilk.

B. Confronto: prima ondata

Per questo primo confronto, si è deciso di selezionare i dati che vanno dall'inizio del campionamento dei dati, l'undicesima settimana del 2020, fino alla ventiduesima settimana del 2020 (nella finestra che va da marzo a fine maggio 2020). É stato scelto questo intervallo in quanto l'impatto delle due nazioni con il COVID-19 non è stato lo stesso, né nel numero di contagi né nel tipo di misure restrittive applicate. Visto che in quel periodo i vaccini non erano ancora disponibili, le uniche contromisure attuabili erano le restrizioni sociali.

Come evidenzia la Figura 3, è subito evidente come la situazione delle due nazioni fosse molto diversa: innanzitutto l'Italia si è trovata ad affrontare un numero di casi settimanali molto più elevato di quelli in Svezia; infatti in Italia si raggiunge un picco massimo di circa 40000 casi settimanali, mentre la Svezia non supera mai i 1000 casi per settimana; anche a fronte di un numero di tamponi molto più elevato in Italia. É altresì eloquente il grafico a barre in Figura 3, che riporta il numero di restrizioni settimanali attive: nelle prime settimane in esame, l'Italia ha un numero molto alto di restrizioni, alcune molto severe (come ad esempio, la chiusura di tutte le attività non essenziali e le scuole); la Svezia invece si limitava a restrizioni mirate a limitare i contatti in ambito lavorativo e scolastico, *telelavoro* e *didattica a distanza*. Nelle settimane successive, la Svezia ha aumentato il numero di restrizioni, senza mai raggiungere misure estreme, mentre l'Italia, vista la differente situazione, ha preferito mantere un approccio più cauto. Quest'approccio però ha permesso di portare nel giro di poche settimane ad uno rapido decremento dei contagi, riportando a fine maggio un numero di infezioni settimanali leggermente più basso di quello svedese.

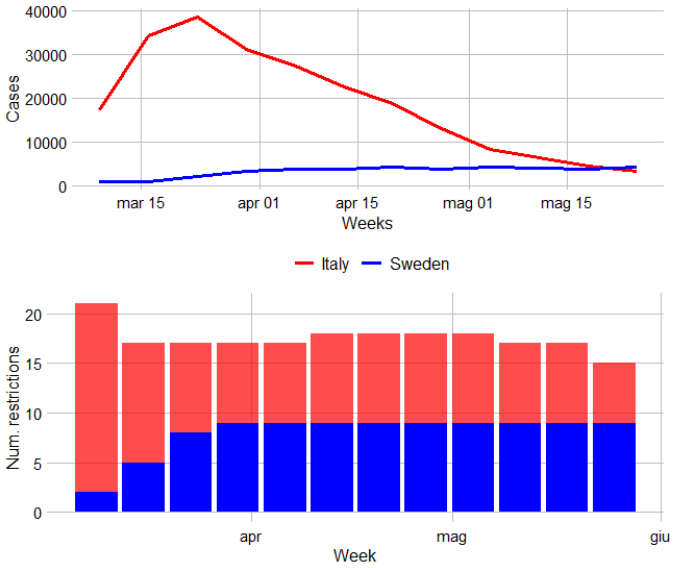

Figura 3: Confronto andamento contagi e numero restrizioni attive settimanali Italia - Svezia.

Quanto è stato descritto per l'andamento dei positivi settimanali, può essere traslato anche per il confronto sul numero di morti settimanali. Da Figura 4 si può notare come l'andamento sia pressoché identico, al netto dell'ordine di grandezza. Un'ulteriore evidenza al fatto che la prima ondata della pandemia di COVID-19 ha colpito molto più duramente.

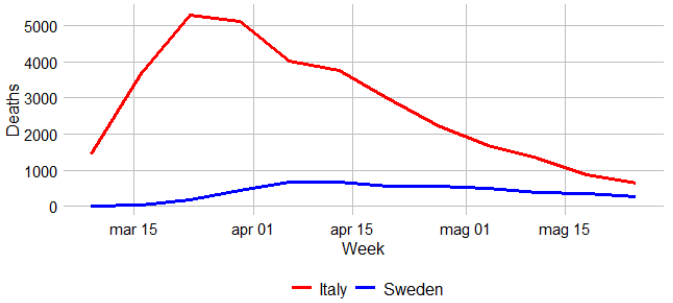

Figura 4: Confronto morti settimanali Italia - Svezia.

Sarebbe stato utile, per approfondire l'analisi, confrontare la pressione che COVID-19 ha esercitato sulle strutture ospedaliere; purtroppo si è posto dinnanzi un problema intrinseco: all'interno del dataset [5] Italia e Svezia non hanno riportato i dati con la stessa metrica: infatti mentre per la Svezia è riportato il numero di nuovi ingressi settimanali in terapia intensiva (metrica normalizzata per centomila abitanti), per l'Italia è presente solo il numero di nuovi ospedalizzati (in senso lato) settimanali, per centomila abitanti. Per questo,

nonostante in Figura 5 sono stati riportati i grafici affiancati, non è possibile fare un'analisi comparativa dei due andamenti, in quanto descrivono due fenomeni non equiparabili.

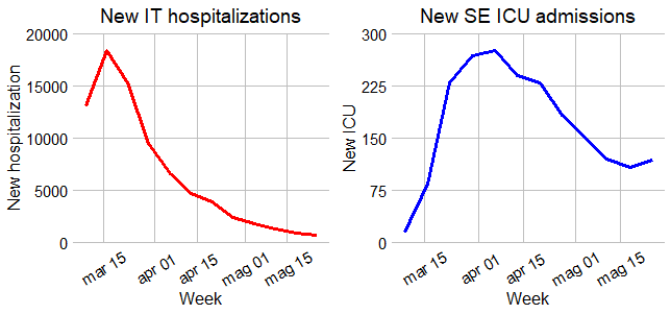

Figura 5: Sinistra: Nuove ospedalizzazioni Italia. Destra: Nuovi ingressi terapia intensiva Svezia. Dati settimanali. I grafici riportano i dati denormalizzati rispetto ai centomila abitanti.

Quello che si può analizzare è che l'andamento delle ospedalizzazioni in Italia segue lo stesso andamento mostrato dai casi e dai morti settimanali, infatti mostra un picco molto elevato attorno al 15 marzo, circa 20000 nuovi ingressi, ma subito dopo presenta una rapida decrescita che riporta, a fine della finestra di osservazione, la variabile ad un valore nettamente inferiore. Ulteriore conferma che le severe restrizioni sociali imposte hanno limitato la circolazione del virus.

In Svezia invece, l'andamento dei nuovi ingressi in terapia intensiva raggiungono un picco massimo qualche settimana dopo l'Italia e mostra una decrescita molto più lenta. Questo curva è ragionevole poiché mentre le ospedalizzazioni sono comuni, gli ingressi in terapia intensiva richiedono sintomi più significativi, di conseguenza sono generalmente più rare. Per quanto riguarda le restrizioni sociali è difficile notare un impatto significativo dato che la Svezia ha sempre avuto dati meno critici in termini assoluti di quelli italiani.

C. Confronto: impatto vaccini

In questa sezione, in cui sono riportati i risultati dell'analisi dei dati di Italia e Svezia con l'avvento dei vaccini, ci si è soffermati su una finestra temporale che va da autunno 2020, periodo in cui c'è stata una recrudescenza dei casi, fino a aprile 2022. Così facendo, si va ad effettuare una valutazione delle variabili immediatamente prima dell'inizio delle campagne vaccinali e, consecutivamente, all'avanzamento e proseguimento di queste ultime.

A primo impatto osserviamo, dalle Figure 6 e 7, che la percentuale stimata di vaccinazioni con almeno tre dosi è inferiore in Svezia rispetto all'Italia. Tuttavia l'andamento dei positivi sembra molto simile, in quanto in entrambi i paesi vediamo che dall'inizio delle somministrazioni di vaccino fino alla fine del 2021 i casi sono diminuiti sensibilmente. Ciò che risalta all'inizio del 2022 è che i casi aumentano i modo repentino. Nonostante la percentuale di popolazione vaccinata elevata ed in crescita, la curva dei contagi raggiunge comunque un picco massimo tra la fine dell'anno 2021 e l'inizio del 2022, con la diffusione della nuova variante, *omicron*.

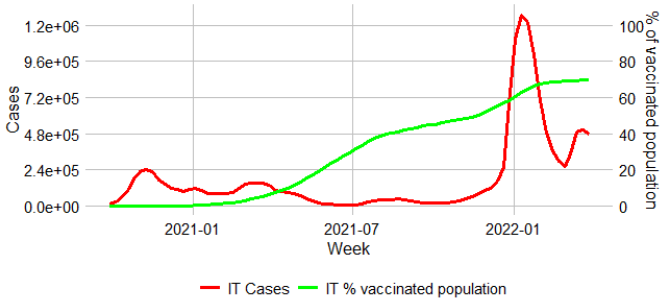

Figura 6: Andamento casi in Italia rapportato alla percentuale di popolazione vaccinata.

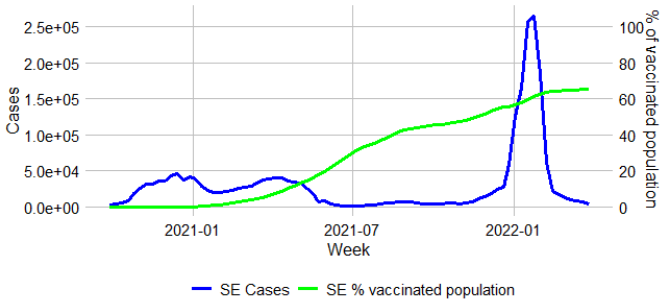

Figura 7: Andamento casi in Svezia rapportato alla percentuale di popolazione vaccinata.

É interessante però vedere i dati riguardanti il tasso di fatalità (1) [6] e il tasso di ospedalizzazione (2) [7] ottenuti in questo nuovo scenario.

$$tasso\_di\_fatalità = 100 * \frac{morti}{casi\_positivi} \quad (1)$$

$$tasso\_di\_ospedalizzazione = 100 * \frac{ospedalizzazioni}{casi\_positivi} \quad (2)$$

Si è deciso di fare ciò perchè permette di osservare, in modo diretto, il numero di decessi e di nuove ospedalizzazioni in funzione del numero di nuovi casi; senza dover tenerne conto separatamente.

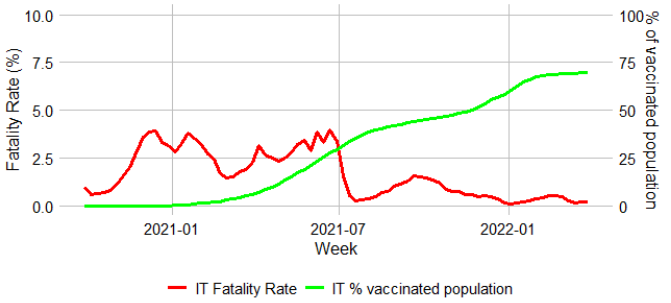

Figura 8: Andamento tasso di fatalità italiano rapportato alla percentuale di popolazione vaccinata.

Partendo dal tasso di fatalità in Italia, dalla Figura 8 si nota immediatamente che la diffusione dei vaccini ha diminuito notevolmente le morti. É importante inoltre osservare che all'inizio dell'anno 2022 i casi si trovavano ad un massimo

storico mentre il tasso di fatalità è rimasto costante ad un livello più basso rispetto al diffondersi dell'immunizzazione da vaccino.

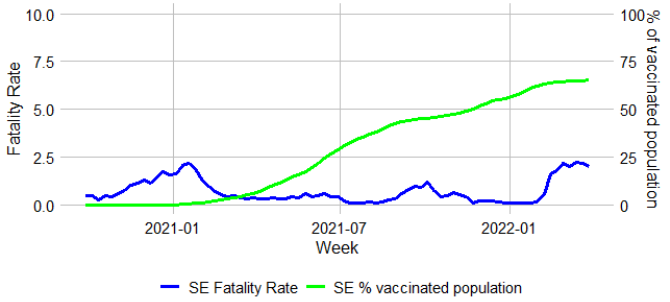

Figura 9: Andamento tasso di fatalità svedese rapportato alla percentuale di popolazione vaccinata.

Il grafico in Figura 9 invece mostra un andamento diverso in cui il tasso di fatalità sembra incrementare nella primavera 2022 anche con una elevata percentuale di popolazione vaccinata.

Per quanto riguarda le nuove ospedalizzazioni, in Italia si può vedere dalla Figura 10 che, dal momento in cui vengono iniziale le somministrazioni dei vaccini, i nuovi ingressi in area medica dovuti al COVID-19 si sono ridotti durante il periodo estivo e cosa più importante sono rimasti costanti a livelli inferiori all'1% durante l'inverno, periodo in cui c'è una recrudescenza della malattia, pertanto è più facile aspettarsi un incremento dei nuovi ricoveri.

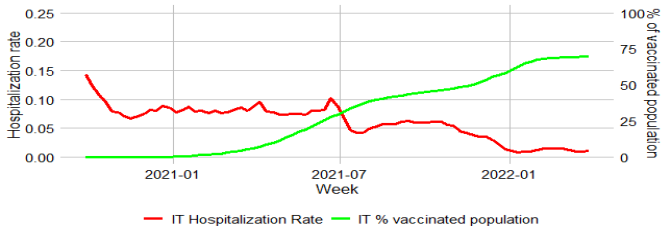

Figura 10: Andamento tasso di ospedalizzazioni in Italia rapportato alla percentuale di popolazione vaccinata.

In Svezia invece, come specificato precedentemente, la metrica non è relativa alle ospedalizzazioni ma rispetto alle nuovi ammissioni in terapia intensiva. Dalla Figura 11 osserviamo che il rapporto tra nuovi ingressi in terapia intensiva ed i casi è rimasto stabile e visivamente non si nota un comportamento inversamente proporzionale alla percentuale di popolazione vaccinata.

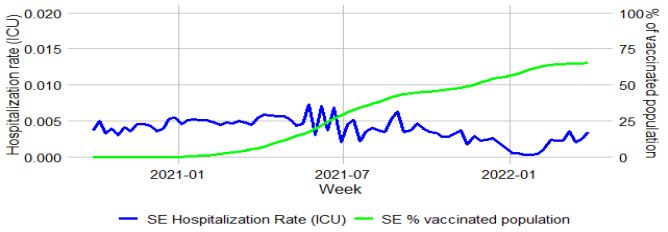

Figura 11: Andamento tasso di ospedalizzazioni (ICU) in Svezia rapportato alla percentuale di popolazione vaccinata.

V. Conclusioni

Le analisi effettuate mostrano l'andamento delle rispettive curve di contagi e decessi cercando di far notare come le contromisure adottate abbiano influenzato la pandemia. É difficile trarre forti correlazioni perché comunque i casi epidemiologici richiedono studi più approfonditi data la complessità del problema. Sarebbe comunque interessante effettuare studi simulativi per capire come sarebbe potuta evolvere la situazione se fossero state prese diverse misure preventive.

Riferimenti bibliografici

[1] ECDC European Centre for Disease Prevention and Control, Covid-19 Data vaccination in the EU/EEA, Maggio 2022, URL: https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea

[2] ECDC European Centre for Disease Prevention and Control, Data on country response measures to COVID-19, Maggio 2022, URL: https://www.ecdc.europa.eu/en/publications-data/download-data-response-measures-covid-19

[3] ECDC European Centre for Disease Prevention and Control, Data on the daily number of new reported COVID-19 cases and deaths by EU/EEA country, Maggio 2022, URL: https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country

[4] ECDC European Centre for Disease Prevention and Control, Data on testing for COVID-19 by week and country, Maggio 2022, URL: https://www.ecdc.europa.eu/en/publications-data/covid-19-testing

[5] ECDC European Centre for Disease Prevention and Control, Data on hospital and ICU admission rates and current occupancy for COVID-19, Maggio 2022, URL: https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-covid-19

[6] Hennekens, C., Buring, J., Mayrent, S. (1987). Epidemiology in Medicine. Little, Brown.

[7] Percentuale di pazienti affetti da Covid-19 ospedalizzati sul totale dei pazienti affetti da Covid-19, Maggio 2022, URL: https://www.agenas.gov.it/covid19/web/index.php?r=site%2Fgraph1