# Analisi della seconda ondata di Covid-19 in Italia

Marco Braga, m.braga@campus.unimib.it
Marco Maugeri, m.maugeri@campus.unimib.it
Andrea Maver, a.maver@campus.unimib.it
Anna Nava, a.nava38@campus.unimib.it
Oscar Zanotti, o.zanotti@campus.unimib.it

**Sinossi**
La diffusione del virus SARS-CoV-2 ha provocato una pandemia mondiale e ciò ha causato enormi disagi alla nostra vita quotidiana, sia direttamente che indirettamente. Il seguente articolo si pone l'obiettivo di analizzare l'andamento dell'epidemia di COVID-19 durante la seconda ondata in Italia. Il primo passo consiste nell'analizzare le serie storiche del numero giornaliero di pazienti ospedalizzati in Italia e nelle varie regioni. Il periodo di tempo analizzato parte dal 1 settembre 2020 al 30 aprile 2021. Per ogni serie storica si testano vari modelli, come il modello autoregressivo ARIMA, il modello exponential smoothing ETS, i modelli ibridi e il modello delle Reti Neurali NNAR. Sono state effettuate delle previsioni per un periodo temporale corrispondente a quarantacinque giorni e si è individuata, studiando vari indici, la tecnica migliore. Successivamente, utilizzando le Reti Neurali, si è svolta una what if analysis, cioè uno studio su come sarebbero evolute le serie se le misure restrittive adottate fossero state più rigide o maggiormente permissive. Infine, viene effettuata una clusterizzazione delle serie temporali regionali per individuare quali regioni dovessero attuare le medesime restrizioni e quali, invece, adottare misure differenti. Dai risultati si scopre che le tecniche di previsione migliori risultano i modelli ibridi, in quanto minimizzano gli errori. Dall'analisi what if emerge come la zona rossa porti a una netta diminuzione del numero di pazienti ospedalizzati nella regione di studio. Si ottiene, però, anche un risultato inaspettato: la differenza tra la zona gialla e la zona arancione non sempre è così netta come ci si potrebbe aspettare. Infine sono stati ottenuti cinque cluster suddividendo le regioni italiane in base all'andamento dell'epidemia.

**Parole chiave.**
Covid-19; Time Series; Forecasting; What if; Clustering.

**1. Introduzione**.
La malattia da coronavirus (COVID-19) è una nuova patologia causata dal virus SARS-CoV-2, nato a Wuhan, provincia di Hubei, in Cina, nel dicembre 2019. Nonostante inizialmente si trattasse solo di una serie di polmoniti di causa sconosciuta, successivamente la diffusione del virus ha causato rapidamente una crisi mondiale. Secondo l'ultimo dato riportato dalla Johns Hopkins University, divulgato dal sito di Rai News [17], quasi centosettantotto milioni di persone sono state infettate, con oltre tre milioni e ottocentomila morti in tutto il mondo. A causa di ciò, diversi Paesi hanno imposto lockdown nazionali, sono stati chiusi tutti i luoghi pubblici e sono state attuate varie politiche di restrizione delle attività per rallentare la diffusione della malattia. L'Italia è stato il primo Paese Europeo ad essere gravemente colpito dal COVID-19, ed è stato uno dei principali epicentri della pandemia per circa due mesi, ovvero da metà febbraio a metà aprile 2020, quando l'epidemia ha raggiunto il primo picco. Successivamente, la curva epidemica è progressivamente diminuita fino a metà agosto 2020, per poi però crescere nuovamente da settembre fino a maggio 2021. L'obiettivo principale di questo studio è fornire il modello migliore per effettuare previsioni a breve termine del numero di pazienti ricoverati in ospedale a causa del virus SARS-CoV-2. Le tendenze relative ai ricoveri ospedalieri consentono infatti di avere un quadro chiaro dello stress e della pressione complessivi sul sistema sanitario nazionale. Inizialmente si sono acquisiti i dati della Protezione Civile Italiana riguardo l'epidemia di COVID-19. Per effettuare le previsioni si utilizzano modelli quali ARIMA, ETS, Reti Neurali e algoritmi ibridi. L'obiettivo è capire quale modello si adatti meglio alle serie temporali e, per perseguirlo, vengono effettuati diversi test e calcolati vari indici di errore. Inoltre, si vuole fornire un metodo per valutare le misure restrittive prese dal Governo Italiano attraverso una what if analysis. La domanda che guida tale

analisi è: quale sarebbe stata la variazione nel numero di pazienti ospedalizzati se fossero state adottate misure restrittive differenti? Infine, si realizza una cluster analysis sulle serie temporali relative alle diverse regioni, sfruttando differenti tipologie di agglomerazione e distanze e valutando i risultati utilizzando dei Cluster Validity Index (CVI). Per concludere, si interpretano i risultati e le loro implicazioni.

**2. Obiettivo/problema affrontato**.
Data la situazione di emergenza che ha drammaticamente caratterizzato l'ultimo anno, si è deciso di cercare di individuare una modalità per analizzare l'andamento dell'epidemia e valutare le misure restrittive adottate, basandosi sull'analisi del numero di totale ospedalizzati. L'obiettivo del seguente articolo è dunque duplice. Si vuole offrire un metodo di previsione per il numero di totale di ospedalizzati. Ciò potrebbe risultare utile agli istituti ospedalieri per poter agire preventivamente, sapendo come allocare le risorse all'interno dei reparti Covid ed evitare che si verifichino situazioni critiche. Per esempio, si può capire quanti operatori sanitari sarebbe necessario assumere e come implementare sufficienti strutture ospedaliere, attrezzature e letti in terapia ordinaria e intensiva. Il secondo obiettivo consiste nello stimare il numero di totale ospedalizzati se fossero state prese differenti misure restrittive. Questa analisi potrebbe rappresentare uno strumento per la valutazione dei provvedimenti attuati e un supporto per le future decisioni durante una terza ondata di Covid-19.

**3. Aspetti metodologici**. Per analizzare le serie storiche vengono utilizzati diversi modelli. Il primo modello è $ARIMA(p,d,q)$: è un modello non stazionario definito (Bell, Holan e McElroy, 2016), dalla seguente equazione: $\Delta^d Y_t = \theta_0 + \phi_1 \Delta^d Y_{t-1} + \ldots + \phi_p \Delta^d Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-} + \ldots + \theta_q \epsilon_{t-q}$, dove $Y = (Y_1, \ldots, Y_t)$rappresenta la serie storica, $\Delta = (1 - B)$ è il different operator, *B* è l'operatore di backshift e i parametri AR e MA sono rispettivamente $\phi_1, \ldots, \phi_p$ e $\theta_1, \ldots, \theta_q$.

Il modello ETS si basa (Hyndman et al., 2008), sulla scomposizione della serie temporale. Nel caso del modello additivo ETS(A,A,A), l'equazione rappresentativa è la seguente: $Y_t = l_{t-1} + b_{t-1} + s_{t-m} + \epsilon_t$, $l_t = l_{t-1} + b_{t-1} + \alpha\epsilon_t$, $b_t = b_{t-1} + \beta\epsilon_t$, $s_t = s_{t-m} + \gamma\epsilon_t$ , dove $\epsilon_t$ è un rumore bianco, $l_t$ rappresenta il trend al tempo $t$, $b_t$viene definito come la slope, cioè la derivata del livello, mentre $s_t$è la componente stagionale; infine $\alpha, \beta, \gamma > 0$ sono parametri del metodo. Sarà utilizzato anche il modello moltiplicativo ETS(M,M,M) le cui equazioni sono(Hyndman et al.,2008): $Y_t = l_{t-1} * b_{t-1} * s_{t-m}(1 + \epsilon_t)$, $l_t = l_{t-1} * b_{t-1}(1 + \alpha\epsilon_t)$, $b_t = b_{t-1}(1 + \beta\epsilon_t)$, $s_t = s_{t-m}(1 + \gamma\epsilon_t)$.

Il modello NNAR assume la forma $y_t = f(y_{t-1}) + \epsilon_t$, dove $y_{t-1} = (y_{t-1}, \ldots, y_{t-n})$ rappresenta il vettore dei lag utilizzati, $f$ la funzione delle reti neurali e $n$ è il numero di neuroni nascosti. I modelli ibridi sono una semplice combinazione dei modelli descritti sopra. Per i metodi ibridi, vengono normalizzati i dati in modo da rendere l'ipotesi di normalità degli errori più plausibile. Il metodo utilizzato è la trasformazione Box-Cox (Li et al., 2021). Viene utilizzata per variabili positive ed è definita come: sia $y = Y_t$, cioè il valore di una serie temporale $Y$ al tempo $t$, allora la trasformazione è $g_\lambda(y) = \frac{y^\lambda - 1}{\lambda}$se $\lambda$ è diverso da zero, altrimenti $g_0(y) = log(y)$. Il valore ottimale di $\lambda$ può essere calcolato applicando la funzione "BoxCox.lambda", contenuta nel pacchetto "forecastHybrid", dell'ambiente R. Effettuata la previsione con i vari modelli, i risultati ottenuti sono approssimati all'intero più vicino dato che il numero di pazienti risulta essere un numero naturale.

Siano $(Y_1, \ldots, Y_n)$ i valori predetti da un modello e siano $(X_1, \ldots, X_n)$ i valori osservati della serie temporale, allora gli indici utilizzati per confrontare i vari modelli sono:

-l'errore assoluto medio, definito come $MAE = \frac{1}{n}\sum_i^n |Y_i - X_i|$;

-l'errore assoluto medio in percentuale, definito come $MAPE = \frac{100}{n}\sum_{i}^{n}\frac{|Y_i - X_i|}{X_i}$. Si può applicare in questo caso dato che, nei dati analizzati, non esistono giorni con nessun paziente ospedalizzato, cioè $X_i$ è sempre non nullo;

-la radice dell'errore quadratico medio, $RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(Y_i - X_i)^2}$.

Per effettuare una what if analysis è stato utilizzato il modello di regressione delle Reti Neurali. Una rete neurale è costituita da un'unità di computazione elementare denominata neurone; ognuno di essi ha tipicamente un insieme di neuroni di input e uno di output. Sono collegati in modo orientato grazie a delle sinapsi, a cui è associato un valore reale, detto peso, rappresentante la forza della connessione tra essi. Sono inoltre caratterizzati da due ulteriori elementi: una soglia, detta anche bias o threshold, e da una funzione di attivazione o di trasferimento. Ogni neurone è tale da ricevere segnali da altri simili, cioè accetta stimoli in ingresso dai propri input, e ne invia altri agli output. Possono essere presenti anche degli strati di neuroni nascosti, che non ammettono connessioni tra di loro. La funzione utilizzata è stata "nnet", contenuta nell'omonimo pacchetto dell'ambiente R. Uno degli input dell'algoritmo di regressione sono i valori della serie temporale del numero di ospedalizzati ritardati di un giorno (il lag a un giorno). Ciò crea però un problema durante un'analisi what if: dato il primo item del test set, è noto il valore del lag e si può quindi effettuare la previsione a un giorno utilizzando un valore certo per l'attributo, ma per le previsioni a due giorni, non è disponibile il dato del lag. Si è allora deciso di effettuare un ciclo con previsioni a un giorno e si è poi inserita la previsione nella serie temporale del lag, al posto del valore reale del numero di ospedalizzati che non si conosce.

La cluster analysis è uno studio che permette di raggruppare le istanze all'interno di gruppi il più omogenei possibile al loro interno ma dissimili l'uno con l'altro. Il clustering gerarchico agglomerativo, in particolare, consiste nell'unire, tramite varie iterazioni, le due istanze più simili di un dataset, fino a quando non sarà stato creato un unico gruppo contenente tutte le osservazioni. Per valutare la vicinanza tra due osservazioni vengono utilizzate misure di dissimilarità. Il clustering gerarchico può essere riassunto nel dendrogramma, cioè un grafico ad albero che mostra come i dati sono stati raggruppati. Una scelta fondamentale per questa tipologia di analisi è la definizione del numero di cluster ottimale per massimizzare la somiglianza tra i vari gruppi. Per calcolare la similarità tra le serie temporali sono state utilizzate delle distanze definite appositamente per questa tipologia di dati: la Dynamic Time Warping (DTW) e la Shape-based distance (SBD). La DTW viene definita come segue (Montero e Vilar, 2014): siano $X = (X_1,\ldots, X_T)$ e $Y = (Y_1,\ldots, Y_T)$ due serie temporali. Sia $M$ l'insieme di tutte le possibili sequenze di $m$ coppie che preservano l'ordine delle osservazioni nella forma $r = ((X_{a_1}, Y_{b_1}),\ldots, (X_{a_m}, Y_{b_m}))$, con $a_i, b_j \in \{1,\ldots T\}$, tali che $a_1 = b_1 = 1$, $a_m = b_m = T$, e $a_{i+1} = a_i$ o $a_{i+1} = a_i + 1$ e $b_{i+1} = b_i$ o $b_i + 1$, per $i \in \{1,\ldots, m-1\}$. Allora la distanza DTW viene definita come $d_{DTW}(X,Y) = min_{r\in M}(\sum_{i=1,\ldots,m}|X_{a_i} - Y_{b_i}|)$. La DTW consente di riconoscere forme simili, anche in presenza di shifting e/o scaling. Tuttavia ignora la struttura temporale dei valori, in quanto la vicinanza si basa sulle differenze $|X_{a_i} - Y_{b_i}|$, indipendentemente dal comportamento intorno a questi valori. La misura SBD (Espinosa, 2019) si basa sulla cross correlation con normalizzazione dei coefficienti (NCCc) tra due serie temporali; questo indice è sensibile rispetto alla scala, perciò Paparrizos e Gravano (2015) consigliano di normalizzare i dati. La sequenza di NCCc è ottenuta attraverso la convoluzione delle due serie, in questo modo diversi allineamenti possono essere considerati. La distanza può essere calcolata come segue: $SBD(X,Y) = 1 - \frac{max(NCCc(X,Y))}{||X||_2 * ||Y||_2}$ e varia tra 0 e 2, dove 0 indica la perfetta similarità. Si è deciso di applicare ai dati un cluster gerarchico con due differenti metodi di aggregazione:

-Ward: siano $A$ e $B$ due cluster, allora $\Delta(A,B) = \sum_{i\in A\cup B}||\vec{x}_i - \vec{m}_{A\cup B}||^2 - \sum_{i\in A}||\vec{x}_i - \vec{m}_A||^2 -$

$\sum_{i\in B} ||\vec{x}_i - \vec{m}_B||^2 - \frac{n_A n_B}{n_a+n_B}||\vec{m}_A - \vec{m}_B||^2$, dove $\vec{m}_j$ rappresenta il centro del cluster $j$e $n_j$è il numero di elementi del cluster $j$. Quindi la distanza di Ward rappresenta di quanto aumenta la distanza tra le osservazioni quando queste vengono unite in un unico cluster. Questo metodo assicura (Murtagh e Legendre, 2011) che la varianza tra i gruppi venga massimizzata, e quella all'interno di un unico cluster sia minimizzata.

-Average linkage: la distanza tra gruppi viene calcolata come la media delle distanze tra tutte le coppie di elementi, in cui uno fa parte di un cluster e uno dell'altro.

Per valutare la qualità delle varie clusterizzazioni sono stati utilizzati vari indici di validità interna, che consentono anche di individuare il numero ideale di cluster in cui suddividere i dati. Gli indici interni, a differenza di quelli esterni, non necessitano, a priori, di una partizione con cui confrontare i cluster ottenuti. Quelli utilizzati in questo studio vengono forniti come standard dal pacchetto 'dtwclust' e sono parte di un gruppo che performa bene in diverse situazioni e con tanti tipi di dati. Siano $C_1$,..., $C_k$kcluster, allora si definiscono i seguenti indici:

-Davies-Bouldin: Sia $\mu_j = \frac{1}{|C_j|}\sum_{x_i\in C_j} x_i$il centro di gravità relativo al cluster $C_j$, allora si definisce la misura di Davies-Bouldin come $DB(k) = \frac{1}{k}\sum_{j=1}^{k} max_{l=1,\dots,k;l\neq j}\frac{\rho(C_j)+\rho(C_l)}{d(C_j,C_l)}$, dove $\rho(C_j) = \frac{1}{|C_j|}\sum_{x\in C_j}||x-\mu_j||$ denota la distanza media tra le osservazioni del cluster $C_j$e il centro di gravità del cluster $C_j$, mentre $d(C_j, C_l)$indica la distanza tra i centri di gravità dei cluster. L'indice deve essere minimizzato.

-Silhouette: sia $x \in C_j$, allora $a(x) = \frac{1}{n_j-1}\sum_{y\in C_j} d(x,y)$ dove $n_j = |C_j|$ e quindi $a(x)$ corrisponde alla media delle distanze tra $x$ e gli altri elementi appartenenti al cluster $C_j$. Si definisce $b(x) = min_{l=1,\dots,k;l\neq j}\{\frac{1}{n_l}\sum_{y\in C_l} d(x,y)\}$ dove $n_l = |C_l|$e quindi $b(x)$è il minimo tra le medie delle distanze tra $x$ e gli elementi degli altri cluster. Allora l'indice di Silhouette associato all'elemento $x$ si definisce come $s(x) = \frac{b(x)-a(x)}{max(a(x),b(x))} \in [-1, 1]$. Se l'indice è vicino a $1$, allora l'elemento $x$ è molto vicino agli elementi del cluster $C_j$ a cui appartiene e molto distante dagli elementi degli altri cluster e quindi si ottiene un buon risultato. Per ottenere l'indice di Silhouette di tutta la clusterizzazione si effettua una media degli indici di Silhouette calcolati per ogni osservazione.

-Dunn: è utilizzato per misurare la compattezza di un cluster ed è definito come $D(k) = min_{j=1,\dots,k}\{min_{l=j+1,\dots,k}\frac{d(C_j,C_l)}{max_{v=1,\dots,k}diam(C_v)}\}$ dove $d(C_j, C_l)$ è la distanza Single linkage tra i due cluster, cioè la distanza tra i due elementi più vicini appartenenti a un diverso cluster, e $diam(C_v) = max_{x,y\in C_v}||x-y||$ è il diametro del cluster $C_v$. L'indice deve essere massimizzato in funzione del numero di cluster $k$. È costoso computazionalmente e sensibile agli outliers.

-Calinski-Harabasz: è definito come $I_{CH} = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k} d(u_i,U)}{\sum_{i=1}^{k}\sum_{x_j\in C_i} d(x_j,u_i)}$, dove $u_i$è il centro di gravità del cluster $i$, $U = \frac{1}{M}\sum_{j=1}^{M} x_j$,in cui $M$è il numero totale di osservazioni, $N$è il numero di features del dataset, $k$il numero di cluster (Łukasik et al., 2016). Maggiore è il rapporto $I_{CH}$, migliore è la partizione dati. Il numero ottimale di cluster è la soluzione con il più alto valore dell'indice.

-COP: l'indice viene definito come (Arbelaitz et al., 2021) il rapporto tra le distanze con i centroidi rispetto alle distanze tra le singolo osservazioni, cioè $COP = \frac{1}{N}\sum_{j=1}^{k}|C_j|\frac{\frac{1}{|C_j|}\sum_{x\in C_j} d(x,\mu_j)}{min_{x_i\notin C_j}\{max_{x_k\in C_j}\{d(x_i,x_k)\}\}}$.

**4. I dati**. I dati analizzati provengono dalla repository [15], in cui sono pubblicati i rilevamenti della Protezione Civile Italiana riguardo l'andamento dell'epidemia da SARS-CoV-2. Possono essere scaricati in formato .csv e sono organizzati come segue: in ogni riga è identificata una regione e il giorno dell'anno in cui sono stati misurati vari attributi, come il numero di positivi giornaliero, il numero di totale ospedalizzati, il numero di pazienti in terapia intensiva. Il numero di regioni rappresentate è ventuno, in quanto il Trentino Alto Adige è stato separato nelle due province autonome di Trento e Bolzano. Ogni regione è anche caratterizzata da due codici nuts, cioè la nomenclatura delle unità territoriali statistiche dell'Italia. L'elenco dei codici nuts per le rispettive regioni si può trovare in [23]. Il numero totale di righe corrisponde a $5082$. Per studiare l'andamento dell'epidemia si è deciso di analizzare le serie storiche riguardanti il numero totale di ospedalizzati, anziché considerare il numero di individui giornalieri risultati positivi al virus. Questa scelta è conseguenza delle numerose criticità legate al raccoglimento dei dati del numero di totale positivi. Infatti, il valore dipende dalla quantità di tamponi effettuati, dal giorno in considerazione, per esempio nei festivi il totale di positivi diminuisce a causa del minor numero di tamponi analizzati. Quindi il numero di individui risultati positivi è un indicatore soggetto a oscillazioni e molto rumore. Invece, il numero totale di individui ospedalizzati non subisce oscillazioni legate al giorno della settimana considerato, è meno rumoroso e si può quindi individuare più precisamente un modello sottostante. Inoltre è un indicatore di quanto i reparti Covid degli ospedali siano occupati e quindi quante risorse gli ospedali debbano ancora utilizzare per la lotta alla pandemia. Ci si concentra sulla cosiddetta seconda ondata, cioè il periodo a partire dal primo settembre $2020$. Questa scelta è legata a due motivazioni. La seconda ondata è stata caratterizzata da misure restrittive differenti tra le varie regioni, identificate dai colori rosso, arancione, giallo e bianco, in ordine decrescente di criticità della situazione pandemica e sanitaria. Poiché l'obiettivo è effettuare un'analisi what if, questa varietà di disposizioni, che possono cambiare ogni settimana, si presta molto a questa tipologia di lavoro rispetto al lockdown nazionale della prima ondata. Inoltre, i dati della prima ondata riguardano una situazione sensibilmente differente rispetto a quelli dell'ondata successiva. Quando la pandemia è scoppiata a marzo $2020$, la situazione nel Nord Italia era critica: non erano state designate linee guida chiare da seguire che potessero indicare come agire in modo uniforme e coeso, gli ospedali erano in difficoltà a gestire i ricoveri e il tasso di pazienti con bisogno di cure ospedaliere era molto più alto. Durante la seconda ondata, invece, grazie all'esperienza acquisita nei mesi precedenti, si è riusciti a gestire meglio l'emergenza e si è evitato di ripetere i medesimi errori. Non solo è diminuito il numero di malati gravi, ma è migliorata anche l'assistenza ai pazienti ospedalizzati (per approfondire, si veda [14]). Inoltre, da settembre $2020$in poi, i contagi hanno caratterizzato tutta la penisola, a differenza della prima ondata che ha colpito prevalentemente il centro-nord. La maggiore chiarezza nelle norme, l'introduzione dei colori e la più precisa raccolta dei dati sono quindi le cause che hanno portato alla decisione di analizzare la seconda ondata. I dati riguardanti le misure restrittive, cioè i colori sulle regioni, vengono forniti dalla Protezione Civile, ma in un formato di difficile elaborazione. Per questo, si è ricorsi alla repository [16], nella quale vengono resi disponibili in formato .csv. I dati iniziano il giorno $11$ novembre $2020$, durante il quale è entrata in vigore la prima ordinanza con la suddivisione a zone dell'Italia. Sono presenti tre attributi: la data in formato YYYY-MM-DD, il nome della regione e il colore, indicato con una stringa, che può essere "rosso", "arancione", "giallo" o "bianco".

**5. Analisi/Processo di trattamento dei dati**.

Il primo passo dell'analisi corrisponde a una fase di preprocessing dei dati: si crea, attraverso un'operazione di join, un unico dataset contenente, per ogni riga, la data, il nome della regione, il colore attribuito e il numero dei pazienti ospedalizzati in quel giorno. Per il periodo che va da settembre ai primi di novembre, quando ancora non erano state istituite le zone, si è considerato come colore la zona bianca. Inizialmente sono stati confrontati vari modelli per la predizione puntuale delle serie temporali. Gli algoritmi scelti sono stati ARIMA, ETS, NNAR e Modelli Ibridi. Questi metodi richiedono una serie temporale univariata, quindi si estrae dal dataset iniziale solo l'attributo riguardante il numero di individui ospedalizzati. Si

procede quindi con la suddivisione delle serie temporali in train e test set: il train set è composto di circa l'ottanta percento delle osservazioni, cioè dal primo settembre $2020$ al $15$ marzo $2021$, le restanti osservazioni fino al $30$ aprile $2021$compongono il test set. Si utilizza la funzione "auto.arima", definita in [18], contenuta nel package "forecast" nell'ambiente R, applicata ai train set. Questa funzione segue passaggi sequenziali per individuare il modello ARIMA che si adatta meglio ai dati e cerca quindi di individuare i valori migliori per $p$, cioè il numero di parametri del processo autoregressivo (AR), l'ordine $i$ della differenziazione e $q$, il numero di parametri per il processo a media mobile (MA). Successivamente segue la fase di validazione del modello: si osservano i grafici riguardanti la correlazione e la distribuzione degli errori, come in Figura 1. I grafici nella prima riga rappresentano l'autocorrelazione totale e quella parziale tra i residui. Le linee tratteggiate rappresentano l'intervallo di incertezza centrato nel valore $0$. I restanti due grafici danno indicazioni sulla distribuzione dei residui: il primo a sinistra è un istogramma, il secondo è il QQ-plot. Infine, si effettua la previsione e la si confronta con il test set, calcolando gli indici MAE, MAPE e RMSE definiti in precedenza. Per il modello ETS, si utilizza la funzione "ets", proveniente sempre dal package "forecast" di R e definita nella documentazione ufficiale in [19]. Il processo di validazione e previsione non cambia. Successivamente si testa il modello NNAR attraverso la funzione "nnerar" definita in [21]. Prima di applicare i modelli ibridi si effettua una normalizzazione utilizzando la funzione BoxCox contenuta nel pacchetto "forecastHybrid" dell'ambiente R. Tale funzione si applica ai dati di train e per individuare il valore ottimale di $\lambda$ si utilizza la già citata BoxCox.lambda. Infine si applicano i modelli ibridi: la funzione utilizzata è "hybridModel" del pacchetto "forecastHybrid". Si testano le varie modalità: "en" è la combinazione tra ETS e NNAR, "an" la combinazione tra ARIMA e NNAR, "aen" la combinazione di tutti e tre insieme, "aent" la combinazione tra i tre modelli già citati e l'algoritmo tbats implementato da R.

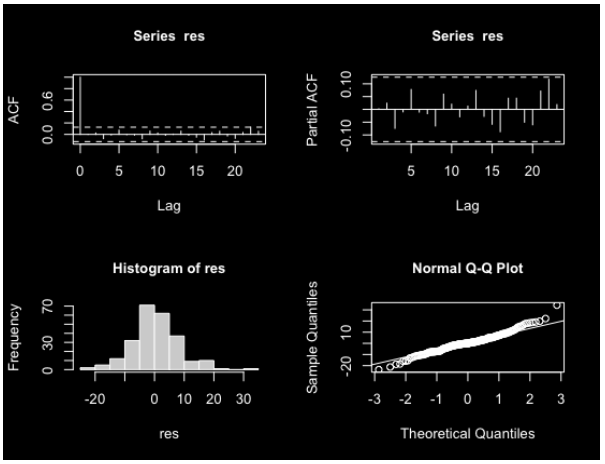

Figura 1. Esempio di residui del Modello ARIMA(p,d,q)

Successivamente, si svolge la what if analysis utilizzando la funzione "nnet", che computa una rete neurale con un singolo strato nascosto (si veda la documentazione R [20]). In particolare, tale funzione regola i valori dei parametri della rete neurale (pesi e bias) per ottimizzare le prestazioni. Si scelgono come variabili di input il numero di ospedalizzati traslato di un giorno e i lag sui colori della regione corrispondenti a $10$ giorni precedenti. Si nota infatti che il modello non riesce a capire in autonomia che l'effetto delle misure restrittive sulla serie temporale non si riscontra immediatamente ma giorni dopo l'entrata in vigore. In particolare, in media, i sintomi della malattia si manifestano dopo 5 o 6 giorni dal contagio, come riportato dall'Organizzazione Mondiale della Sanità ([22]), mentre il tempo mediano fra gli ospedalizzati che passa tra l'insorgenza dei sintomi e il ricovero corrisponde, secondo l'Istituto Superiore di Sanità ([13]), a $5$ giorni. Scegliamo quindi di utilizzare come regressore il lag a $10$ giorni. La Rete Neurale viene allenata sul periodo che inizia il primo settembre $2020$ e si conclude il 16 aprile $2021$ e come test set si considerano le ultime due settimane del mese. Si noti che, poiché "nnet" è soggetta a un processo stocastico, è stato fissato un seed per poter replicare il codice, il quale si può trovare allegato al report. Una volta ottenuto un modello soddisfacente, si può passare alla domanda centrale dell'analisi: come sarebbe cambiato l'andamento del numero di pazienti ospedalizzati se in passato si fossero applicate misure restrittive più rigide o, viceversa, più permissive? Si sceglie di studiare il caso in cui le restrizioni fossero state anticipate di una settimana rispetto alla reale entrata in vigore,

ovvero dal 30 ottobre invece che dal 6 novembre. Vengono imposti tali colori per tre settimane, fino al 21 novembre, mentre nel grafico vengono rappresentate le previsioni fino al 3 dicembre. Inizialmente, si osserva anche il comportamento del modello con attributo il reale colore attribuito alle regioni in modo da verificare la bontà della previsione. Infine si confrontano gli andamenti delle serie temporali rispetto al colore della zona imposto e al caso reale.

Per la cluster analysis, vengono utilizzate le librerie "tsclust" e "dtwclust", implementate nell'ambiente R, specifiche per la clusterizzazione delle serie storiche. Si considera la serie temporale univariata del totale degli ospedalizzati e si crea un dataset avente per ogni riga una serie temporale per ogni regione. Quindi si ottiene un dataset con 21 osservazioni nel quale le colonne rappresentano il giorno in cui è stata effettuata la misurazione. Per calcolare le distanze tramite SBD è richiesta la normalizzazione con z-score delle serie temporali. I pacchetti citati consentono di modificare vari parametri in funzione della tipologia di raggruppamento che si desidera: si può scegliere, ad esempio, un clustering gerarchico o partizionato, quale metodo di aggregazione applicare e quale misura utilizzare per il calcolo delle distanze fra punti. Le proprietà comuni a tutti i test eseguiti sono state l'opzione di clustering gerarchico e la "shape extraction" come funzione di calcolo del centroide. Viene inserito anche un intervallo di valori che indica il numero potenziale di cluster ottenibili, in modo da poter eseguire successivamente dei confronti. Per quanto riguarda i singoli test, invece, sono stati eseguiti con distanza DTW o SBD e con agglomerazione per media o con metodo di Ward. Ciò genera quattro diversi risultati da valutare. Agli output di questa funzione viene poi applicata una "sapply" che permette di calcolare i valori dei CVI interni descritti in precedenza. Gli indici sono poi inseriti e ordinati in tabelle, in modo tale da rendere più agevole la consultazione. Utilizzando infine dei plot si possono osservare i dendrogrammi e le serie che fanno parte di ogni cluster, per avere una facile analisi grafica dei risultati. I codici sono implementati in Espinosa (2019), ma adattati alle serie temporali riguardanti l'epidemia.

**Risultati**. La tabella 1 nell'appendice A riporta i risultati riguardo l'applicazione dei vari modelli al numero di pazienti ospedalizzati in tutta Italia. L'output della funzione "auto.arima" applicata alla serie storica è $ARIMA(0,2,5)$. Nella figura 2 si possono osservare i grafici riguardanti l'auto correlazione totale e parziale, l'istogramma e il QQ-plot relativi ai residui del modello $ARIMA(0,2,5)$ applicato alla serie temporale nazionale.

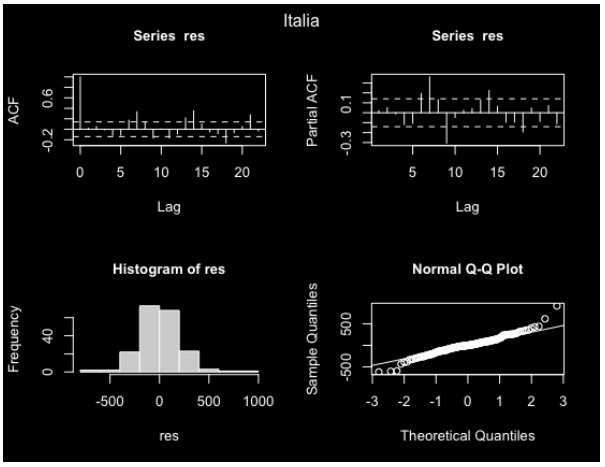


Figura 2. Residui per modello $ARIMA(0,2,5)$, applicato ai pazienti ospedalizzati in Italia

Si può notare come l'autocorrelazione parziale e totale sia contenuta all'interno dell'intervallo di confidenza centrato in zero ad eccezione di alcuni valori casuali. Inoltre l'istogramma è centrato in zero e il QQ-plot segue, ad eccezione delle code, l'andamento della normale. Quindi si può affermare che gli errori siano incorrelati e con distribuzione normale standard, cioè un rumore bianco. Questo è un test per la validità del modello ARIMA trovato. Si può trarre la medesima conclusione di validazione del modello ETS(A,A,N), che tra quelli exponential smoothing minimizza gli indici nella tabella 1, osservando il comportamento dei residui nella figura 13, che si può trovare nell'appendice B. I parametri del modello ETS(A,A,N), le cui componenti sono rappresentate nella figura 3, sono i seguenti: $\alpha = 0.999$, $\beta = 0.5958$. Nella prima immagine in alto a sinistra si può notare il trend della serie, che, nel momento in cui finisce il train set, è in leggera crescita. La seconda immagine in alto mostra la slope del modello e l'immagine in basso i residui. Essendo un modello in cui tutte le componenti presenti sono additive, la serie dei pazienti ospedalizzati sarà data dalla somma delle tre componenti rappresentate nella figura.

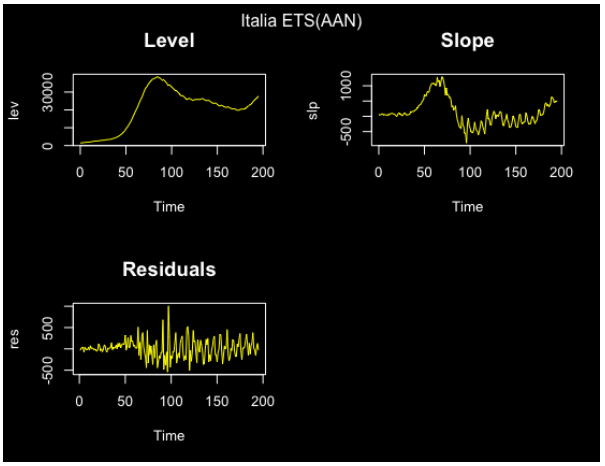


Figura 3. Componenti per il modello ETS(A,A,N) applicato ai pazienti ospedalizzati in Italia

Il migliore tra i modelli non ibridi è NNAR con 120 lag e 5neuroni, il cui indice MAPE è minore di 10 e quindi, come affermato da Lewis (1982), le previsioni vengono considerate molto accurate. Si può notare come i modelli ibridi applicati ai dati normalizzati abbiano un indice MAPE<10 e in generale sono più efficienti rispetto ai modelli non ibridi. Nella figura 4 si possono notare le previsioni ottenute sulla serie non normalizzata e nella figura 5 ottenute sui dati normalizzati. E' evidente come il modello NNAR e quelli ibridi riescano a capire meglio il rallentamento della crescita della curva.

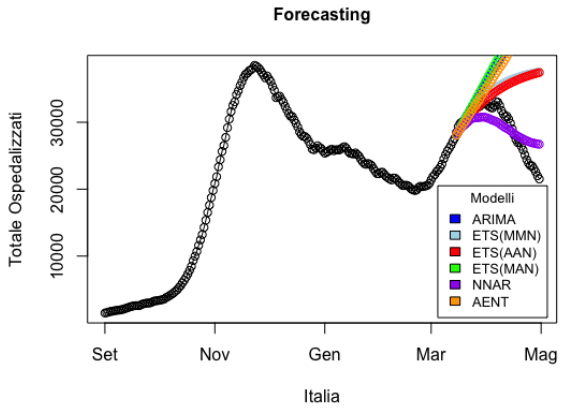


Figura 4. Serie storica del numero di ospedalizzati nazionale con dati non normalizzati, modelli applicati rappresentati nella legenda

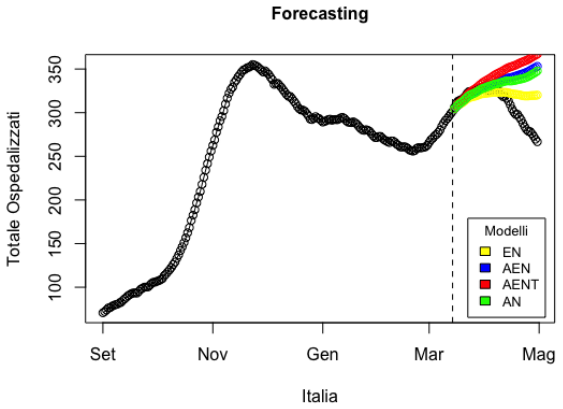


Figura 5. Serie storica del numero di ospedalizzati nazionale con dati normalizzati, modelli applicati indicati nella legenda

Vengono riportati i risultati per alcune delle regioni colpite più duramente dal covid: Lombardia nella tabella 2, Toscana nella tabella 3 e Campania nella tabella 4, in appendice.

Per quanto riguarda la Lombardia il modello ARIMA applicato è $ARIMA(1,1,1)$. Dalla figura 14 nell'appendice B si nota come i residui non siano correlati, infatti le barre dei grafici rappresentati l'autocorrelazione totale e parziale sono all'interno dell'intervallo di confidenza centrato in zero. Inoltre la distribuzione dei residui è normale, a meno delle code. Quindi, i residui del modello sono un rumore bianco e di conseguenza il modello è convalidato. Tra i modelli exponential smoothing l'errore minore è attribuito a ETS(M,M,N), cioè modello privo di stagionalità e con componenti moltiplicative. Anche questo modello viene validato dalla figura 15. Il miglior modello non ibrido è NNAR con 120 lag e 5 neuroni, il miglior modello ibrido è AN. I modelli ibridi con dati normalizzati hanno MAPE minore di 15, quindi, forniscono delle previsioni molto buone.

Studiando la serie temporale relativa ai dati della Toscana, si scopre che il modello ARIMA è il seguente: $ARIMA(4,2,2)$. Nonostante la distribuzione dei residui relativi ad ARIMA mostrata nella figura 16 non sia simmetrica, si può comunque concludere che i residui siano un rumore bianco. Osservando la tabella 3 nell'appendice A, si intuisce come il miglior modello exponential smoothing sia ETS(M,M,N),

la cui validazione può essere effettuata per mezzo della figura 17. I parametri del modello ETS(M,M,N) sono $\alpha = 0.9894$, $\beta = 0.2085$. Anche in questo caso i modelli ibridi offrono delle previsioni più precise rispetto all'indice MAPE, il migliore dei quali è EN.

I risultati migliori, rispetto all'indice MAPE, sono ottenuti con la regione Campania, e ciò si può anche notare dalla figure 18 e 19 dell'appendice B. Il modello arima ottenuto con "auto.arima" è $ARIMA(1,2,1)$. Come si può osservare dalla figura 20, gli errori hanno autocorrelazione totale e parziale nulla e la distribuzione nel QQ-plot coincide con quella di una normale. Ciò permette di validare il modello ARIMA(1,2,1). Osservando gli indici, il miglior modello exponential smoothing è ETS(M,M,N), che consente di ottenere un'ottima previsione. Anche in questo caso i modelli ibridi sono i migliori dato che hanno un MAPE minore di due.

Dai risultati descritti in precedenza emerge come il modello ARIMA non sia sufficiente per ottenere una buona previsione poiché non riesce a cogliere il reale andamento della serie. I modelli exponential smoothing lavorano bene con alcune serie, come nel caso della regione Campania, e male in altre, come la Lombardia. Per effettuare le previsioni, i modelli che minimizzano gli errori sono i modelli ibridi, in particolare EN o AN. Infatti introducono un livello di complessità maggiore e consentono di effettuare previsioni con un indice MAPE molto basso, generalmente minore di $10$. Quindi dovrebbero essere questi modelli ad essere utilizzati in un'ipotetica terza ondata per effettuare previsioni sul numero di pazienti ospedalizzati nazionale e regionale.

Come in precedenza, si è scelto di riportare i risultati della what if analysis ottenuti per le regioni Lombardia, Campania e Toscana. Inizialmente vengono calcolati gli indici MAPE relativi alla previsione svolta con i reali colori vigenti nelle regioni. Si è scelto di utilizzare il MAPE perché altri indici espressi in termini assoluti vengono influenzati dalla diverso numero di abitanti nelle varie regioni. Si può notare che il risultato peggiore si riscontra per la Lombardia, con un indice pari a $10.98$, mentre gli indici della Toscana e della Campania valgono rispettivamente $4.9$ e $4.0$. Si inizia l'analisi a partire dalla Lombardia: si osservi la figura 6 rappresentante la serie temporale del numero di ospedalizzati reale in verde e le previsioni per le quattro restrizioni stampate in bianco, giallo, arancione e rosso, esattamente come i colori corrispondenti. Tale convenzione è stata adottata per tutti i grafici.

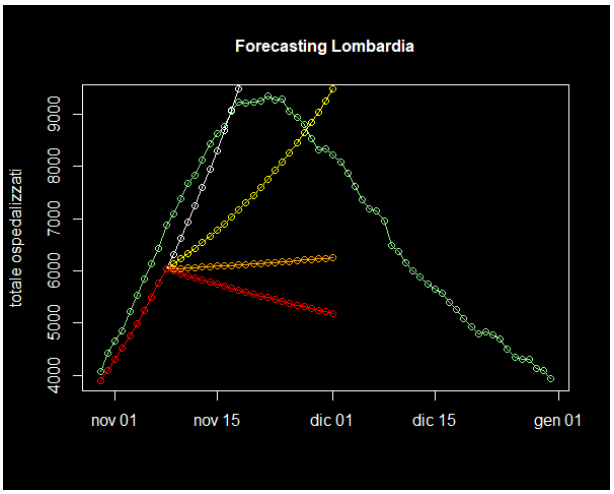


Figura 6. What if Analysis sul numero di pazienti ospedalizzati per la regione Lombardia

Coerentemente con il modello scelto, gli effetti delle zone si notano dopo $10$ giorni. L'unico colore che porta alla decrescita del numero di ospedalizzati è il rosso, ma anche l'arancione provoca una diminuzione notevole della pendenza della curva. Nel risultato per la Toscana, riportato nella figura 7, il rosso è l'unico a comportare una decrescita del numero di ospedalizzati, ma stavolta tale decrescita è notevolmente più veloce. Gli andamenti delle serie temporali relative all'arancione e al giallo sono piuttosto simili, anche se alla zona gialla consegue una crescita più rapida. La zona arancione sembra essere meno efficace in questo caso rispetto a quanto visto per la Lombardia. Una ragione possibile potrebbe essere la seguente: in Lombardia, per qualche settimana facente parte del training set, si è imposta la zona arancione rafforzata, che invece non è mai stata adottata in Toscana.

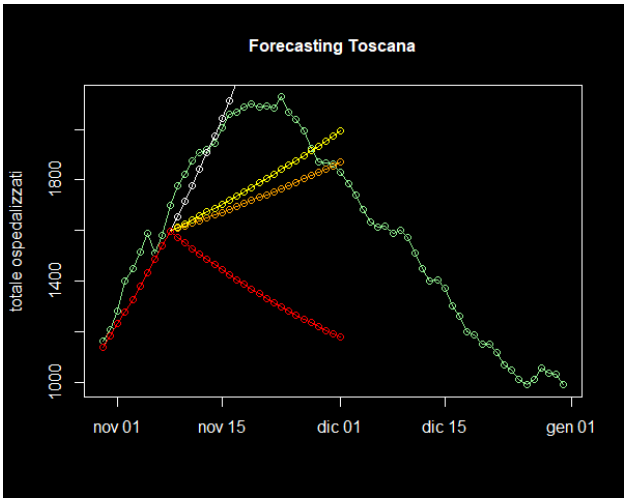


Figura 7. What if Analysis sul numero di pazienti ospedalizzati per la regione Toscana

Infine, si osservi la figura 8 per l'analisi effettuata sui dati relativi alla regione Campania. In questo particolare caso, tutti i colori, escluso ovviamente il bianco, causano una decrescita del numero di ospedalizzati. Soprattutto per quanto riguarda la zona gialla, tale previsione di una diminuzione nel numero degli ospedalizzati stride con quanto ci si aspetterebbe. Si potrebbe forse ricondurre tale comportamento alla chiusura delle scuole in zona gialla le scuole, imposta dal Governatore della regione. E' stata presentata un'analisi che potrebbe essere approfondita maggiormente, però sono stati ottenuti dei risultati interessanti: in ordine di velocità della decrescita della curva, si ha prima il rosso, che corrisponde infatti all'adozione di norme più severe, poi l'arancione ed infine il giallo.

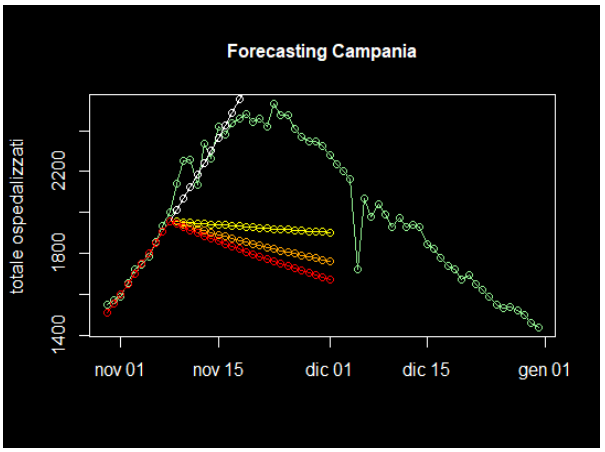


Figura 8. What if Analysis sul numero di pazienti ospedalizzati per la regione Campania

Per confrontare le serie temporali delle varie previsioni, vengono riportate nella tabella 5 le differenze medie tra il numero di ospedalizzati reali e la previsione svolta imponendo la zona rossa dal 30 ottobre, tra la zona rossa e la zona arancione e tra la zona arancione e la zona gialla. Si nota subito che tutti i valori sono negativi: coerentemente con quanto ci si aspetterebbe, il numero di ospedalizzati è mediamente minore con la zona rossa, nella quale vigono le regole più rigide, mentre è un più alto con la zona arancione ed ancora maggiore con la zona gialla. In valore assoluto, le differenze medie più grandi si riscontrano in Lombardia. Questo risultato non sorprende poiché la Lombardia è la regione più popolosa tra quelle analizzate.

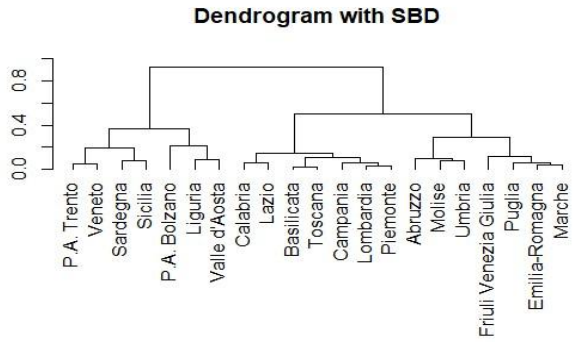


Figura 9. Dendrogramma relativo alla misura SBD

Per quanto riguarda la cluster analysis, si può iniziare la valutazione dei risultati confrontando quale metodo di aggregazione produca cluster migliori. Per entrambe le distanze, il metodo di Ward fornisce delle soluzioni più chiare: dall'osservazione dei dendrogrammi, è evidente come, utilizzando la distanza media, non sia facilmente identificabile l'altezza a cui tagliare l'albero. Per questo motivo verranno prese in considerazione le clusterizzazioni con la distanza Ward. A questo punto è necessario capire quale misura di dissimilarità possa performare meglio: dall'esame preliminare dei dendrogrammi in Figura 9 e 10 sembra che SBD generi dei gruppi più definiti, in quanto le distanze tra un aggregazione e l'altra sono maggiori. Il dendrogramma ottenuto con DTW fornisce risultati accettabili, ma le distanze sono molto compresse e ciò non fornisce una chiara indicazione della soluzione. In questo caso, infatti, il risultato più evidente corrisponderebbe a due

cluster, esito non efficace per la tipologia di studio affrontato. Per confermare che SBD offre soluzioni migliori, si possono confrontare i valori degli indici osservabili nelle Tabelle 6 e 7.

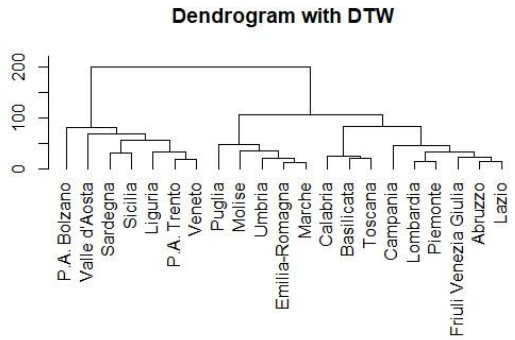


Figura 10: Dendrogramma con distanza DTW

Indici per Cluster con misura SBD

| Numero di Cluster\Indici | Sil | CH | DB | D | COP |
|---|---|---|---|---|---|
| 2 | 0,465 | 11,172 | 0,807 | 0,214 | 0,275 |
| 3 | 0,398 | 10,189 | 0,770 | 0,178 | 0,149 |
| 4 | 0,426 | 10,180 | 0,671 | 0,231 | 0,110 |
| 5 | 0,439 | 9,792 | 0,571 | 0,262 | 0,091 |
| 6 | 0,424 | 9,167 | 0,436 | 0,357 | 0,079 |
| 7 | 0,438 | 9,031 | 0,369 | 0,462 | 0,065 |
| 8 | 0,381 | 8,983 | 0,404 | 0,487 | 0,052 |
| 9 | 0,350 | 8,493 | 0,374 | 0,523 | 0,045 |

Tabella 6. Indici per Cluster con SBD

Indici per Cluster con misura DTW

| Numero di Cluster\Indici | Sil | CH | DB | D | COP |
|---|---|---|---|---|---|
| 2 | 0,349 | 13,357 | 0,963 | 0,149 | 0,370 |
| 3 | 0,293 | 10,920 | 0,812 | 0,144 | 0,290 |
| 4 | 0,346 | 9,532 | 0,982 | 0,207 | 0,212 |
| 5 | 0,337 | 8,151 | 0,699 | 0,265 | 0,167 |
| 6 | 0,309 | 7,381 | 0,568 | 0,265 | 0,135 |
| 7 | 0,282 | 6,952 | 0,632 | 0,308 | 0,119 |
| 8 | 0,329 | 6,805 | 0,603 | 0,447 | 0,101 |
| 9 | 0,294 | 6,286 | 0,523 | 0,525 | 0,089 |

Tabella 7. Indici per Cluster con DTW

È necessaria una doverosa premessa: i valori determinati dagli indici devono essere considerati come uno strumento di supporto alla scelta del numero ottimale di cluster, non come dei risultati assoluti. Come si può osservare dalle tabelle, in base all'indice che si considera, vengono spesso indicati come migliori dei risultati contrastanti, o completamente opposti. Bisogna quindi ponderare le varie possibilità, considerando il dendrogramma, i CVI e l'obiettivo dello studio. Dalle tabelle si evince che i valori degli indici Silhouette (Sil) e Calinski-Harabasz (CH), che vanno massimizzati, per SBD siano in generale leggermente migliori rispetto a quelli di DTW. Osservando gli indici che devono essere minimizzati invece, cioè Davies-Bouldin (DB) e COP, sono di poco inferiori. Per questo motivo, e per la maggiore chiarezza del dendrogramma, si decide di considerare la clusterizzazione con SBD e Ward come la preferibile tra le quattro possibilità studiate. L'ultimo passo consiste nell'individuare il numero di cluster ottimale da ottenere. I valori dei CVI non forniscono risultati definiti, ma si possono ricavare alcuni spunti interessanti. Infatti l'indice di Silhouette è massimizzato a due gruppi, risultato, come detto precedentemente, non utile. Il secondo valore che massimizza l'indice di Silhouette corrisponde a cinque gruppi. Si nota come gli indici associati a cinque cluster rappresentino un buon valore intermedio tra tutte le varie scelte. Un buon risultato potrebbe essere ottenuto, sempre studiando i vari indici, con sette cluster: dividendo così le ventuno osservazioni del dataset si otterrebbero però gruppi composti da uno o due regioni, che non sarebbero molto significativi per le nostre finalità. Per questo motivo, anche a seguito dell'analisi del dendrogramma, viene scelto cinque come numero ottimale di cluster da considerare.

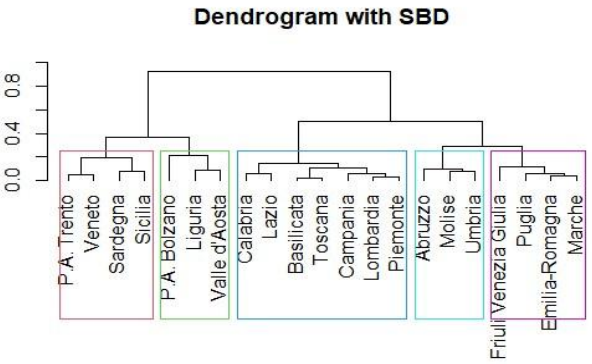


Figura 11: Suddivisione in cinque cluster

In Figura 11 si può osservare la suddivisione delle regioni in gruppi. La Figura 12 mostra invece l'andamento delle serie all'interno di ogni singolo cluster.

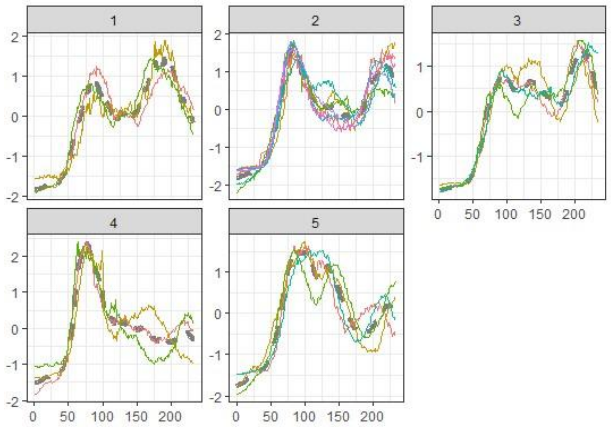


Figura 12: Serie storiche regionali suddivise in cluster

Si può osservare come nel cluster 2, quello più numeroso, gli andamenti siano tutti molto simili. Ciò indica che il raggruppamento è stato efficace. Questo vale anche per il gruppo 1 e, in parte, per per il gruppo 5. I due cluster rimanenti presentano invece situazioni più anomale. Questi sono composti da tre o quattro serie ciascuno e in entrambi è possibile osservare lo stesso fenomeno: presentano una serie con un andamento quasi identico a quello del centroide (indicato con la linea tratteggiata), mentre le altre risultano sopra o sotto il centroide, come se i loro valori si pareggiassero tra di loro. Vengono elencati successivamente i cluster con le regioni che vi appartengono. Cluster 1: Abruzzo, Molise, Umbria; cluster 2: Calabria, Lazio, Basilicata, Toscana, Campania, Lombardia, Piemonte; cluster 3: Friuli Venezia Giulia, Puglia, Emilia-Romagna, Marche; cluster 4: P.A. Bolzano, Liguria, Valle d'Aosta; cluster 5: P.A. Trento, Veneto, Sardegna, Sicilia. Si può osservare che, in particolare, molte delle regioni più colpite sono state raggruppate insieme, come Lombardia, Campania e Piemonte. Dello stesso gruppo fanno parte anche regioni meno popolose, come Calabria e Basilicata, ma che hanno avuto un andamento simile, dovuto al fatto che i dati sono stati normalizzati, quindi il fattore più importante è la forma della curva. Regioni con andamenti altalenanti come la Valle d'Aosta e la provincia di Bolzano sono anch'esse state raggruppate nello stesso cluster. Si può notare anche il cluster contenente Abruzzo, Umbria e Molise, regioni poco popolate, ma che hanno un picco molto alto nel numero di ospedalizzati nella prima parte della seconda ondata, che va poi a decrescere con l'entrata nel nuovo anno. È interessante notare come ogni gruppo presenti un andamento definito e diverso da quello degli altri, a dimostrare una suddivisione più che soddisfacente. Si può notare come le serie appartenenti al cluster uno abbiamo un primo picco più basso rispetto al secondo; invece, al contrario, per le serie del cluster due, il primo picco è più altro rispetto al secondo. Sia ha quindi un comportamento opposto: nelle regioni del primo cluster la situazione è peggiorata con il secondo picco rispetto agli ultimi mesi del 2020; mentre, nelle regioni appartenenti al secondo cluster, nonostante un nuovo aumento di casi, non è stata raggiunta la situazione critica degli ultimi mesi del 2020. La situazione migliore si ha nelle regioni appartenenti al quarto cluster, in cui, dopo il picco iniziale, si ha una discesa quasi continua fino alla fine del periodo analizzato.

**Conclusione e possibili sviluppi**

Dall'analisi predittiva possiamo concludere che, fra i modelli testati, i modelli ibridi siano i migliori rispetto all'indice MAPE. Tra di essi, in particolare, spiccano EN e AN, che combinano rispettivamente ETS con NNAR e ARIMA con NNAR. La sola autoregressione ARIMA risulta invece essere piuttosto deludente, mentre ETS non sempre riesce a fornire previsioni soddisfacenti. Un possibile sviluppo, per migliorare ulteriormente tale analisi, consisterebbe nello studiare anche modelli dotati di regressori esterni. Si potrebbe a quel punto valutare se le previsioni ottenute siano migliori delle precedenti, o, ancora, quali fra i regressori scelti siano i più influenti.

Durante l'analisi precedente, si è vista l'efficienza dei modelli ottenuti con le Reti Neurali e, per questo, si è scelto di utilizzarle per la what if analysis, seppur ricorrendo ad una nuova funzione ("nnet"). I risultati ottenuti rispettano piuttosto bene le aspettative iniziali: con la zona rossa si prevede una decrescita nel numero di ospedalizzati e nei grafici si può vedere come, supponendo di imporre colori con regole più rigide, la serie temporale prevista si abbassi. Per

le prime settimane considerate, la serie temporale sovrastante tutte le altre è quella dei casi reali, di poco differente rispetto alla serie prevista imponendo la zona bianca. Se si volessero applicare i modelli ottenuti per poter stimare gli effetti delle restrizioni introdotte, sarebbero però da considerare le diverse criticità riscontrate. Innanzitutto, si nota che il modello non capisce in autonomia che l'effetto delle restrizioni non si riscontra nell'immediato, ma solo dopo diversi giorni, motivo per cui si è reso necessario fornire come input i valori dei colori laggati, complicando la scelta della regressione ottimale da utilizzare. Inoltre, non sempre le stesse misure restrittive erano vigenti in tutto il territorio regionale: vari decreti, infatti, hanno imposto misure più restrittive solo in alcuni comuni e non in tutta la regione. Unitamente a ciò, si deve anche considerare che, non raramente, le regioni scegliessero in autonomia di rafforzare le restrizioni. Infine, le regole sono state continuamente aggiornate nel tempo anche a livello nazionale: si pensi, per esempio, all'istituzione dell'arancione rafforzato, il quale, chiaramente, non viene indicato nel dataset, che diventerebbe altrimenti troppo complicato. Tutto ciò rende più difficile per il modello comprendere i reali effetti dei colori. Nonostante queste criticità, le previsioni ottenute risultano essere plausibili e aprono ad eventuali studi successivi. Risulta interessante chiedersi perché in alcune regioni le misure adottate risultano essere più efficaci rispetto alle altre e confrontare con maggiore profondità gli effetti delle restrizioni, svolgendo anche modelli sulla serie temporale del numero di positivi giornalieri. Un altro dato interessante che non è stato considerato in questo report è il numero di vaccini effettuati. Questa informazione potrebbe aiutare il modello per capire l'andamento del numero di pazienti ospedalizzati, su cui il numero di individui vaccinati ha una grossa influenza.

La cluster analysis permette di individuare regioni con andamenti della pandemia simili, e quindi di valutare se nelle suddette regioni siano state applicate le stesse misure restrittive o se differenti misure hanno avuto lo stesso effetto. Nell'analisi svolta, si è scelto di utilizzare cinque gruppi, dopo aver confrontato gli indici di valutazione per numeri di cluster differenti. I risultati della clusterizzazione si possono considerare soddisfacenti e forniscono degli interessanti spunti di riflessione. Innanzitutto, si è osservato che diverse zone d'Italia sono state colpite con diverse intensità dal virus; si è notato inoltre che non tutte le regioni appartenenti allo stesso gruppo sono state sottoposte al medesimo tipo di restrizioni, a indicare che, nonostante tutto, misure diverse hanno portato allo stesso risultato. Le ragioni potrebbero essere molteplici: la poca efficacia di alcune restrizioni, il non rispetto delle limitazioni da parte della cittadinanza, particolari comuni o province con misure diverse rispetto al resto della regione o, ancora, le difficoltà di diverse strutture ospedaliere a rispondere tempestivamente all'improvvisa emergenza sanitaria.

**Riferimenti bibliografici**.


[1]Arbelaitz O., Gurrutxaga I., Muguerza J., Pérez J.M., Perona I. (2012), *An extensive comparative study of cluster validity indices*, in *Pattern Recognition*,Volume 46, Issue 1 (2013), pag. 243-256

[2] Bell W.R, Holan S.H., McElroy T.S (2012), *Economic Time Series, Modeling and Seasonality*, Taylor & Francis Group, LLC, pag. 88

[3]Espinosa A.S. (2019), *Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package*, disponibile al seguente url https://cran.r-project.org/web/packages/dtwclust/vignettes/dtwclust.pdf

[4]Gravano L., Paparrizos J. (2015), *k-Shape: Efficient and accurate Clustering of Time Series*

[5] Hyndman R.J., Koehler A.B, Keith Ord J., Snyder R.D. (2008), *Forecasting with Exponential Smoothing, The State Space Approach*, Springer, pag. 45

[6] Legendre P., Murtagh F. (2011), *Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm*

[7]Lewis, C. D. (1982), *Industrial and business forecasting methods: A Radical guide to exponential smoothing and curve fitting*. London; Boston: Butterworth Scientific.

[8]Li P. , Yu T. , Chen B. , Qin J(2021), *Maximum profile binomial likelihood estimation for the semiparametric Box–Cox power transformation model*

[9]Łukasik S.,Kowalski P.A., Charytanowicz M., Kulczycki P. (2016), *Clustering using Flower Pollination Algorithm and Calinski-Harabasz Index*

[10] Montero P., Vilar J.A. (2014), *TSclust: An R Package for Time Series Clustering*

[11]Saitta S., Raphael B., and Smith I.F.C (2007), *A Bounded Index for Cluster Validity*, Machine Learning and Data Mining in Pattern Recognition, Springer, Heidelberg, pp. 174-187

[12] Wierzchoń S.T., Kłopetek M.A., (2018), *Modern Algorithms of Cluster Analysis*, pag. 171, Spinger

**Riferimenti sitografici**

[13]https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia#5, consultato il 22 giugno 2021

[14]https://www.fondazioneveronesi.it/magazine/articoli/cardiologia/covid-19-come-vengono-curati-i-pazienti-in-terapia-intensiva, consultato il 22 giugno 2021

[15]https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni, consultato il giorno 31 maggio 2021

[16]https://github.com/imcatta/restrizioni_regionali_covid, consultato il giorno 30 aprile 2021

[17]https://www.rainews.it/dl/rainews/articoli/Coronavirus.-Nel-mondo-178-milioni-di-casi-5a4a73da-ab38-4487-97f9-7e86283544db.html, consultato il 22 giugno 2021

[18]https://www.rdocumentation.org/packages/forecast/versions/8.15/topics/auto.arima, consultato il 22 giugno 2021

[19]https://www.rdocumentation.org/packages/forecast/versions/8.15/topics/ets, consultato il 22 giugno 2021

[20]https://www.rdocumentation.org/packages/nnet/versions/7.3-16/topics/nnet, consultato il 22 giugno 2021

[21]https://www.rdocumentation.org/packages/forecast/versions/8.15/topics/nnetar, consultato il 23 giugno 2021

[22]https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19, consultato il 22 giugno 2021

[23]https://it.wikipedia.org/wiki/Nomenclatura_delle_unit%C3%A0_territoriali_per_le_statistiche_dell%27Italia, consultato il 22 giugno 2021

**Appendice A**

Tabella 1. Errori dei modelli applicati ai dati nazionali

Italia

| Modello\Indice | MSE | MAPE | MAE | Normalizzazione | Ibrido |
|---|---|---|---|---|---|
| ARIMA | 1.779564e+08 | 37.60467 | 9.746213e+03 | No | No |
| ETS(A,A,N) | 5.294962e+07 | 19.81272 | 5.053830e+03 | No | No |
| ETS(M,M,N) | 5.601613e+07 | 20.68876 | 5.314936e+03 | No | No |
| ETS(M,A,N) | 1.973295e+08 | 39.76860 | 1.032411e+04 | No | No |
| NNAR | 4.934438e+06 | 6.925607 | 1.910681e+03 | No | No |
| EN | 443.422383 | 5.058764 | 14.687223 | Sì | Sì |
| AEN | 1327.357602 | 8.449571 | 24.418366 | Sì | Sì |
| AENT | 2026.35733 | 10.87098 | 31.65195 | Sì | Sì |
| AN | 1107.798795 | 7.674336 | 22.153467 | Sì | Sì |
| AENT | 1.444680e+08 | 33.40066 | 8.601617e+03 | No | Sì |

Tabella 2. Errori dei modelli applicati ai dati della Lombardia

Lombardia

| Modello\Indice | MSE | MAPE | MAE | Normalizzazione | Ibrido |
|---|---|---|---|---|---|
| ARIMA | 8.189537e+06 | 40.45857 | 2.115298e+03 | No | No |
| ETS(A,A,N) | 9.181921e+06 | 42.90903 | 2.247043e+03 | No | No |
| ETS(M,M,N) | 5107980.2979 | 31.7541 | 1650.0851 | No | No |
| ETS(M,A,N) | 1.953144e+07 | 62.94184 | 3.315617e+03 | No | No |
| NNAR | 4.074130e+05 | 8.054349 | 5.545106e+02 | No | No |
| EN | 210.456342 | 9.765369 | 9.583541 | Sì | Sì |
| AEN | 384.82506 | 13.76839 | 13.63979 | Sì | Sì |
| AENT | 415.76635 | 14.57130 | 14.50077 | Sì | Sì |
| AN | 139.666754 | 8.009523 | 7.884503 | Sì | Sì |
| AENT | 1.295154e+07 | 51.02857 | 2.673979e+03 | No | Sì |

Tabella 3. Errori dei modelli applicati ai dati della Toscana

Toscana

| Modello\Indice | MSE | MAPE | MAE | Normalizzazione | Ibrido |
|---|---|---|---|---|---|
| ARIMA | 844793.89362 | 41.96607 | 749.89362 | No | No |
| ETS(A,A,N) | 812258.63830 | 41.01718 | 732.80851 | No | No |
| ETS(M,M,N) | 77301.27660 | 12.97891 | 229.61702 | No | No |
| ETS(M,A,N) | 800063.68085 | 40.68517 | 726.82979 | No | No |
| NNAR | 76831.29787 | 12.34873 | 225.38298 | No | No |
| EN | 2.701456 | 2.122266 | 1.356237 | Sì | Sì |
| AEN | 37.604842 | 7.918778 | 5.101270 | Sì | Sì |
| AENT | 87.629925 | 12.019776 | 7.758091 | Sì | Sì |
| AN | 12.789001 | 4.593136 | 2.944753 | Sì | Sì |
| AENT | 641600.02128 | 35.61601 | 634.82979 | No | Sì |

Tabella 4. Errori dei modelli applicati ai dati della Campania

Campania

| Modello\Indice | MSE | MAPE | MAE | Normalizzazione | Ibrido |
|---|---|---|---|---|---|
| ARIMA | 257610.04255 | 25.02478 | 420.80851 | No | No |
| ETS(A,A,N) | 73191.0851 | 13.8023 | 232.5745 | No | No |
| ETS(M,M,N) | 56836.44681 | 12.26552 | 206.78723 | No | No |
| ETS(M,A,N) | 312950.89362 | 27.71468 | 466.21277 | No | No |
| NNAR | 6108.765957 | 3.954793 | 67.148936 | No | No |
| EN | 0.01010539 | 0.81317004 | 0.07948882 | Sì | Sì |
| AEN | 0.08964451 | 2.47534198 | 0.24203097 | Sì | Sì |
| AENT | 0.0838973 | 2.4263445 | 0.2372692 | Sì | Sì |
| AN | 0.1055515 | 2.6618487 | 0.2602441 | Sì | Sì |
| AENT | 148328.70213 | 19.00863 | 319.63830 | No | Sì |

Tabella 5. Differenze medie nel numero di ospedalizzati per confronto tra misure restrittive per la what if analysis

Differenza tra colori

| Differenza\Regione | Lombardia | Toscana | Campania |
|---|---|---|---|
| Rosso-Reale | -2268.36689 | -453.84381 | -399.77303 |
| Rosso-Arancione | -397.54373 | -260.75948 | -34.77541 |
| Arancione-Giallo | -1003.19310 | -41.91648 | -54.44613 |

**Appendice B**

Figura 13: Residui relativi al modello ETS(A,A,N) per gli ospedalizzati nazionali

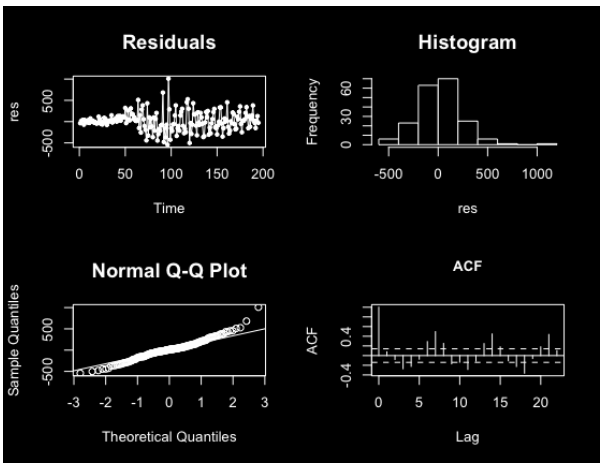


Figura 14: Residui relativi al modello ARIMA(1,1,1) applicato agli ospedalizzati della Lombardia

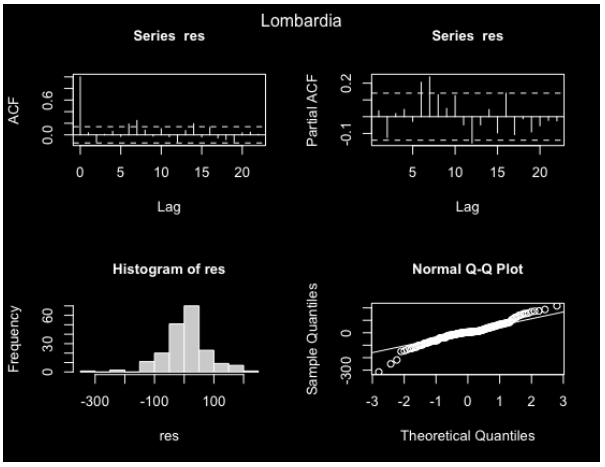


Figura 15: Residui relativi al modello ETS(M,M,N) applicato alla serie del numero di ospedalizzati in Lombardia

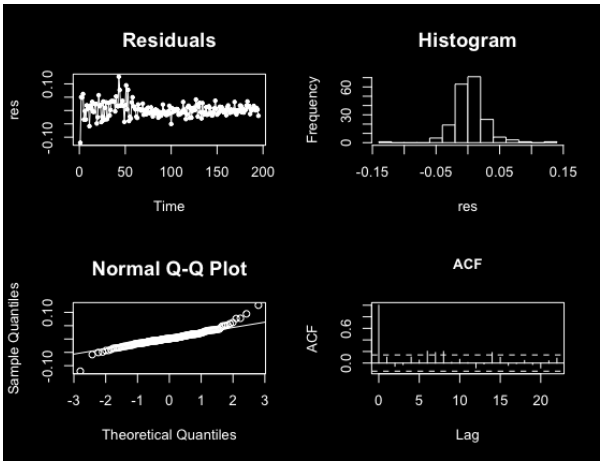


Figura 16: residui relativi al modello ARIMA(4,2,2) applicato ai dati della Toscana

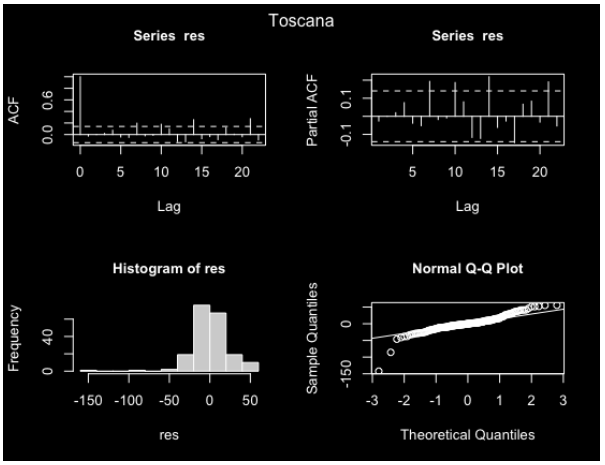

Figura 17: validazione per modello ETS(M,M,N) applicato alla Toscana

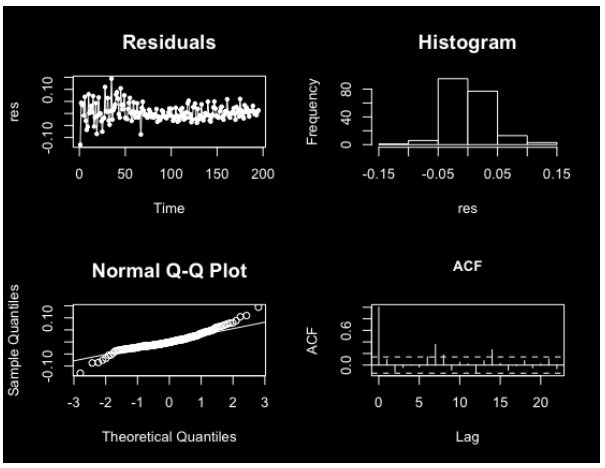


Figura 18: Previsione ospedalizzati in Campania senza normalizzazione dei dati

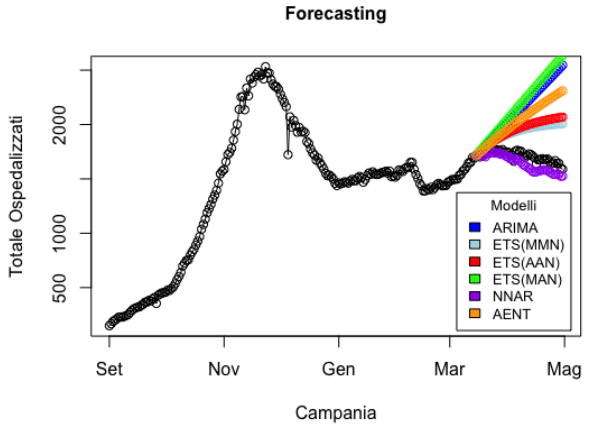


Figura 19: serie relativa agli ospedalizzati in Campania con dati normalizzati

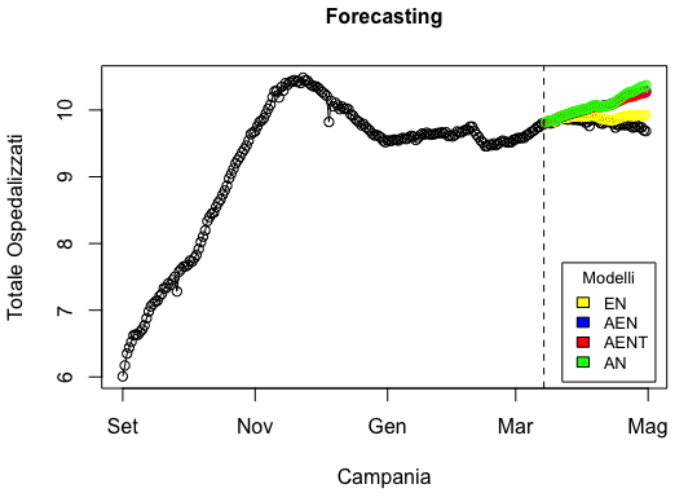


Figura 20: residui del modello ARIMA applicati ai dati relativi alla regione Campania