Città al riparo dal virus: è possibile?

GIOVANNI DE FEUDIS[1], CAMILLA DI MARTINO[2], IMAN RAS[3], ADELE ZANFINO[4], AND NATALIYA ZAYEVA[5]

[1] *820602 CdLM Data Science, Università degli Studi di Milano-Bicocca, g.defeudis1@campus.unimib.it*
[2] *873147 CdLM Data Science, Università degli Studi di Milano-Bicocca, c.dimartino2@campus.unimib.it*
[3] *812509 CdLM Data Science, Università degli Studi di Milano-Bicocca, i.ras@campus.unimib.it*
[4] *867496 CdLM Data Science, Università degli Studi di Milano-Bicocca, a.zanfino@campus.unimib.it*
[5] *867981 CdLM Data Science, Università degli Studi di Milano-Bicocca, n.zayeva@campus.unimib.it*

Compiled January 15, 2022

L'obiettivo di questo breve studio è analizzare la possibilità che la morfologia urbana possa essere cruciale nella diffusione di una malattia infettiva, ed in particolare il nostro studio si è basato sui dati della pandemia da COVID-19 in corso. Dopo aver raccolto i dati relativi ai contagi e aver pulito il dataset, abbiamo scelto alcune città rappresentative delle forme urbane da noi prese in esame: a scacchiera, lineari e radiali. Sono state effettuate, successivamente, delle analisi di cluster attraverso l'utilizzo del programma R ed in particolare della libreria *TSclust*. In seguito tramite il linguaggio di programmazione Python, abbiamo ricercato delle correlazioni utilizzando l'indice di Spearman, abbiamo testato la presenza di differenze all'interno delle categorie (Lineari, Radiali, Scacchiera) mediante test ANOVA e abbiamo svolto delle previsioni; inoltre, con opportuni modelli di forecasting come Prophet e ARIMA abbiamo studiato l'andamento dei contagi in città rappresentative per ogni cluster risultante.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

CONTENTS

1. INTRODUZIONE

La pandemia di COVID-19 attualmente in corso è ormai tristemente nota ad ognuno di noi. A partire dal rilevamento della sua diffusione in Europa, nel febbraio 2020, abbiamo tutti subito un grosso cambiamento nelle nostre più comuni abitudini; questo per via delle precauzioni adottate per contrastare la diffusione del virus, che sta tuttora mietendo migliaia di vite umane ogni giorno nel mondo. In questo clima di tensione, ricerca e precauzione, è lecito chiedersi se il luogo nel quale si vive abbia una certa rilevanza sull'espansione del virus: vivere in un centro popoloso rende indubbiamente più arduo

isolarsi dalle altre persone, rispetto a quanto potrebbe accadere in un piccolo paese. Le città, avendo una più alta densità abitativa, sono senza dubbio luogo di un maggior numero di contatti umani e di conseguenza di contagi. Le differenze nell'andamento della curva dei contagi cambiano notevolmente da città a città, ed è lecito chiedersi quali fattori, oltre alla densità abitativa, possano incidere su questo fattore. Abbiamo dunque pensato, rifacendoci ad uno studio già pubblicato (AbouKorin S.A., Haoying H., Mahran M,2020), di esaminare come la forma architettonica della pianta di alcune città italiane può incidere sulla diffusione del Coronavirus nelle stesse.

Le piante vengono divise in tre categorie:

1. **A scacchiera o a griglia**: questo schema prevede un sistema di strade che si intersecano ad angolo retto, formando isolati rettangolari quadrangolari.
2. **Radiali**: questa pianta è caratterizzata da strade divergenti a raggiera che si propagano a partire da un nucleo centrale, in genere.
3. **Lineari**: la città caratterizzata da questa pianta sorge lungo un asse generatore, sul quale si innestano le vie trasversali, e tende dunque ad avere una forma stretta e lunga.

Nel *paper* sono analizzate nove città tedesche, nove città inglesi e nove città italiane, nel nostro progetto invece sono state scelte ventidue città italiane da esaminare con lo stesso scopo: la ricerca di somiglianze nella diffusione dell'epidemia in città costruite seguendo lo stesso modello architettonico.
Diversamente dall'articolo di riferimento, sono state scelte città esclusivamente italiane per avere omogeneità riguardo alla fonte dei dati e al modo in cui questi ultimi sono stati raccolti e catalogati.
Il cuore del lavoro è diviso in cinque parti:

1. *Raccolta dati e preprocessing* : I dati, forniti dalla Protezione Civile Italiana, sono stati raccolti a partire da Febbraio 2020, e vengono descritti nel paragrafo 4.
2. *Scelta delle città*: i dati sono forniti in base alla provincia, e non alla città (ciò è naturale se si pensa alla difficoltà di definire il confine tra un centro urbano e quelli limitrofi), perciò abbiamo preso in considerazione solo città in cui numero di abitanti sia significativo rispetto a quello della propria provincia.
3. *Clustering di serie temporali*: le serie storiche analizzate sono state suddivise in gruppi tramite algoritmi di clustering, ottenendo un raggruppamento delle serie simili tra di loro. Il risultato è stato confrontato con il raggruppamento teorico delle stesse, ovvero la classificazione delle città nelle tre tipologie di assetto urbanistico.
4. *Ricerca di correlazione* : abbiamo testato un'eventuale differenza fra le medie dell'*indice Rt* per ogni categoria e allo stesso modo abbiamo testato se ci fossero differenze fra l'*infection rate cumulato* delle tre piante cittadine. Successivamente abbiamo testato la correlazione tra le medie dell'indice Rt e i diversi tipi di forma urbana; infine abbiamo calcolato la correlazione tra questi ultimi e l'infection rate cumulato.
5. *Forecasting di serie temporali*: scelte alcune città per ogni cluster, abbiamo cercato il modello che si adattasse meglio ai dati facendo uso dei modelli *Prophet* e *Arima*.

2. OBIETTIVO/PROBLEMA AFFRONTATO

Ispirandoci alla tesi proposta nel paper, il nostro obiettivo è di indagare se la forma di una città possa influenzare la diffusione di una malattia infettiva, come quella alla base dell'epidemia attuale. In particolare il criterio architettonico che sottende all'intersezione delle strade principali potrebbe influenzare il modo in cui le persone si incontrano, e di conseguenza il modo in cui può avvenire la trasmissione della malattia. Chiaramente esistono vari aspetti su cui riflettere al fine di tentare di contenere il numero di persone affette da COVID-19, come le qualità e quantità dei mezzi pubblici o l'igiene generale della città. Di certo la variabile riguardante la pianta architettonica stradale non è esaustiva nello spiegare le motivazioni dietro ai *trend* nell'andamento della pandemia.
Tuttavia, la protezione contro l'insorgenza di nuove malattie infettive è un aspetto rilevante da considerarsi nella costruzione di nuove città o nell'ampliamento di quelle esistenti, e in questo la forma della loro pianta architettonica potrebbe essere un punto di partenza per l'edificazione di nuovi spazi urbani più sicuri per l'uomo.
Infatti la crescente urbanizzazione nel mondo (Véron, 2008) sta portando, e porterà maggiormente in futuro, sempre più persone a vivere nelle città, ponendo queste ultime davanti alla sfida di doversi ampliare per ospitare più individui di quanti ne abbiano mai conosciuti nella storia. Di conseguenza si creeranno spazi ad altissima densità abitativa e ciò, come già accennato, è uno scenario particolarmente pericoloso se si pensa a qualunque tipo di malattia infettiva.
Se si considera inoltre che le malattie di questo tipo sono destinate ad aumentare in conseguenza del *climate change* e della conseguente alterazione di molti ecosistemi, è cruciale che il processo di urbanizzazione avvenga in modo sostenibile e a misura d'uomo.
In questo contesto la preferenza di un modello urbanistico a discapito di un altro nella costruzione di nuovi spazi urbani potrebbe essere davvero significativa, andando a mitigare gli effetti di una eventuale prossima epidemia.

3. ASPETTI METODOLOGICI

Lo studio svolto rivolge l'interesse alla diffusione del patogeno Sars-Cov-2 all'interno del territorio cittadino, rapportato all'assetto urbanistico delle singole città. Seguendo gli schemi di disposizione degli edifici, divisi secondo i tre modelli principali (lineare, radiale, griglia), abbiamo cercato di verificare una relazione tra i suddetti schemi e i dati sul contagio in nostro possesso.
A tal scopo abbiamo svolto tre analisi, ciascuna caratterizzata da un diverso scopo e da differenti modelli : il clustering della serie temporale di ogni città, lo studio della correlazione e delle differenze all'interno delle tre piante cittadine ed infine la ricerca del miglior modello di previsione nel tempo.

A. Clustering

La prima analisi è un clustering , ossia il tentativo di dividere gli oggetti di studio (in questo caso le città a nostra disposizione) in gruppi disgiunti, in modo che gli elementi appartenenti ad ogni gruppo siano il più possibile simili tra di loro e diversi da quelli appartenenti ai gruppi diversi.
Dato l'insieme di città $C = c_1, ..., c_n$, definiamo l'insieme dei cluster $A = A_1, ..., A_m$ tali che siano soddisfatte le seguenti condizioni:

- $A_i \neq \emptyset \quad \forall i = 1,..m$, ossia ogni cluster contiene almeno una città;
- $A_i \cap A_j = \emptyset \quad$ se $i \neq j$, i cluster sono disgiunti tra di loro;
- $\forall j = 1,...,n \quad \exists! i = 1,...,n \quad$ tale che $\quad c_j \in A_i$, ogni città è assegnata ad esattamente un cluster.

Per valutare la bontà della partizione dei cluster ottenuti si scelgono una o più metriche che definiscono quanto gli oggetti in esame sono vicini tra loro (cioè simili) o lontani, e in base ad ognuna di queste funzioni di distanza vengono creati i cluster, uno per ogni metrica utilizzata.
Esistono vari tipi di cluster, che differiscono in base alla loro forma o al modo in cui vengono costruiti; in questo caso abbiamo utilizzato un clustering gerarchico, in cui i cluster sono rappresentati tramite alberi e possono presentare dei sottocluster. In particolare l'insieme dei dati viene preliminarmente partizionato in cluster, e successivamente ogni cluster in cui sono presenti due o più gruppi omogenei tra loro viene diviso a sua volta in altrettanti cluster, ottenendo così una nuova partizione.
La particolarità del clustering gerarchico è di tenere traccia di queste successive divisioni costruendo un albero per ognuno dei cluster originali, in cui quest'ultimo è la radice e i suoi nodi figli sono i cluster in cui viene diviso. Successivamente, tagliando gli alberi ad una certa altezza fissata, cioè una distanza dai nodi terminali, si ottiene una partizione in cluster con le proprietà sopra elencate. A diverse altezze si ottengono naturalmente diverse partizioni.

A.1. Metriche utilizzate

È cruciale la scelta della distanza adottata secondo la quale vengono costruiti i cluster, perché esse determinano completamente il concetto di somiglianza o dissomiglianza tra le serie in esame. Le distanze utilizzate sono dunque le seguenti (nell'elenco seguente x e y denotano i vettori che contengono i dati della coppia di serie storiche in esame):

1. **Distanza euclidea**:

$$ED(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

dove N è la lunghezza delle due serie storiche.

2. **Complexity-Invariant Distance Measure For Time Series** (**CID**) : distanza basata sulla metrica euclidea corretta dalla stima della complessità della serie. È definita come:

$$CID(x,y) = ED(x,y)\dot{C}F(x,y)$$

dove $ED(x,y)$ è la distanza euclidea e $CF(x,y)$ è un fattore di correzione della complessità definito come:

$$CF(x,y) = \frac{\max(CE(x), CE(y))}{\min(CE(x), CE(y))}$$

dove $CE(x)$ è la stima della complessità (*complexity estimate*) della serie temporale x, definito come la radice della somma dei quadrati delle differenze successive della serie temporale x, cioè le differenze tra il valore x_{t_1} della serie x assunto al tempo t_1 e il valore x_{t_0}, assunto al tempo t_0, precedente a t_0.

$$CE(x) = \sqrt{\sum(\text{diff}(x))^2}$$

3. **Compression-based dissimilarity measure** (**CDM**): misura la dissimilarità tra due serie temporali in base alla dimensioni delle stesse serie compresse. È definita come:

$$d(x,y) = \frac{C(xy)}{C(x) + C(y)}$$

dove $C(x)$ e $C(y)$ sono le dimensioni, misurate in byte, delle rispettive serie temporali. $C(x,y)$ è invece la dimensione, sempre misurata in byte, delle serie x e y concatenate. La CDM non è una metrica, ma una misura di dissimilarità, che varia da $\frac{1}{2}$ (perfetta uguaglianza) a 1 (discrepanza massima).

La prima, la distanza euclidea, è una metrica standard che restituisce cluster basati sulla forma delle serie temporali. La seconda e la terza invece sono distanze basate sulla complessità, ovvero si basano sul concetto di *complessità di Kolmogorov* di un oggetto x, che intuitivamente è la minima quantità di informazione richiesta perché un algoritmo generi x. Poiché la complessità di Kolmogorov non è computabile, non è possibile calcolare distanze che la utilizzino direttamente, tuttavia esistono misure di dissimilarità basate su sue approssimazioni come d_{CMD} (Montero e Vilar, 2014).

B. Correlazioni e differenze tra le categorie di piante cittadine

All'interno della seconda analisi sono state prese in considerazione le eventuali correlazioni mediante l'utilizzo dell'indice di correlazione di Spearman, che ha permesso di mettere a confronto e testare l'esistenza della correlazione tra la media dell'indice Rt e successivamente l'infection rate cumulato con la tipologia di pianta cittadina.
L'indice di correlazione R per ranghi di Spearman è una statistica non parametrica che misura la relazione non monotona tra due variabili.

B.1. Coefficiente di Spearman

1. **Coefficiente di Spearman**:

$$\rho_s = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2}\sqrt{\sum_i (s_i - \bar{s})^2}}$$

Come per altri coefficienti di correlazione, quest'ultimo può variare tra -1 e +1, dove 0 indica l'assenza di correlazione mentre +1 e -1 indicano un'esatta relazione monotona.
La valutazione della correlazione ha come scopo quello di valutare se possa esistere una relazione tra la tipologia della pianta cittadina e l'indice Rt o l'infection rate cumulativo, quindi l'ipotesi nulla H0 indica, nel caso preso in esame, la non presenza di una relazione monotonica tra le variabili.
Come parametro del livello di significatività nei test è stato utilizzato $\alpha = 0.05$, cioè verranno rigettate le ipotesi nulle nel caso il p-value presenti valori più piccoli.

B.2. Test ANOVA

Il test ANOVA(Analysis of Variance) presuppone che i gruppi di dati provengano da una popolazione normale, che sia rispettata l'omoschedasticità e ci sia indipendenza tra i gruppi. Esso permette di confrontare due o più gruppi di dati osservando la variabilità interna a questi gruppi(Varianza within) con la variabilità tra i gruppi(Varianza Between). L'ipotesi Nulla prevede che i dati abbiano tutti la stessa origine, nel caso specifico l'ipotesi H0 prevede che non ci sia differenza tra le città all'interno delle tre categorie di pianta prendendo in considerazione la

media dell'indice Rt e dell'infection rate cumulato. La variabilità all'interno dei gruppi è considerata un errore casuale, mentre la variabilità tra i gruppi è attribuibile alle differenza tra i gruppi, quindi all'interno dell'ipotesi nulla si è interessati a verificare che le medie di tutti i gruppi siano uguali tra loro:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_c$$

Per verificare l'ipotesi la variabilità totale viene scomposta in due componenti, una attribuibile alle differenze tra i gruppi e una seconda componente che fa riferimento alle differenze riscontrate all'interno dei gruppi. Facendo quindi riferimento alla statistica test F per l'ANOVA a una via si ottiene:

$$F = \frac{\frac{SSA}{n-c}}{\frac{SSW}{c-1}} = \frac{MSA}{MSW}$$

Dove SSA è la variabilità tra gruppi, SSW è la variabilità all'interno dei gruppi e si trova che la statistica F ha distribuzione F con $(c-1)$ gradi di libertà al numeratore e $(n-c)$ gradi di libertà al denominatore. Ricordando che il valore del p-value è la probabilità di osservare un valore di F maggiore o uguale a quello osservato, nel caso l'ipotesi nulla sia vera. Per la valutazione della correlazione tra i gruppi è stato utilizzato il coefficiente di correlazione di Pearson, l'indipendenza è stata testata utilizzando il test Chi-quadro, per la normalità è stato utilizzato lo Shapiro-Wilk test e per testare l'omoschedasticità è stato utilizzato il test di Levene.

C. Forecasting

Per strutturare delle previsioni sono stati utilizzati i modelli Prophet e Arima. Il primo semplifica l'attività dell'analista sviluppando le previsioni su serie temporali in cui i valori futuri dipendono dai dati assunti in precedenza, così che il conoscitore di dominio possa avere a disposizione parametri intuitivi utili alla comprensione dell'utente finale tanto da portare il tuning a un livello di complessità approcciabile anche da utilizzatori meno esperti. Il secondo è un derivato del modello ARMA che diviene stabile attraverso le differenze di ordine d. Verificata la validità del modello, questo viene indirizzato alla previsione così da poter raggiungere l'obbiettivo d'analisi prefissato.
Veniamo ora ai due modelli da noi utilizzati.

C.1. ARIMA

L'abbreviazione sta per Autoregressive Integrated Moving Average. Questa classe di modelli ci consente di analizzare serie storiche stazionarie. La stazionarietà può essere forte o debole; nel primo caso il comportamento di una serie di valori risulta essere identico allo stesso insieme spostato nel tempo di un intervallo h, quindi la stazionarietà forte viene definita come l'invarianza nel tempo dell'intero processo dei dati. Il secondo tipo di stazionarietà implica che la media e la varianza siano indipendenti dal tempo, ovvero che la funzione del valor medio,μt, è costante e non dipende dal tempo t. La funzione di autocovarianza *y(s,t)*, dipende da s e t solo attraverso la distanza $|s\text{-}t|$.
L'acronimo include le parti più importanti del modello:

- AR (AutoRegressive): usa la relazione di dipendenza tra un'osservazione e le osservazioni n-lag;
- I (Integrated): rende la serie stazionaria attraverso la differenziazione della stessa;
- MA (Moving Average): usa la relazione di dipendenza tra un'osservazione e l'errore residuo attraverso un modello a media mobile applicato alle osservazioni ritardate.

Un modello ARMA Integrato di ordine d è un processo stocastico che diventa stazionario dopo essere stato differenziato d volte. Considerato che il processo stazionario può essere rappresentato con un modello ARMA(p,q) utilizzando i polinomi AR e MA nell'operatore di backward B, otteniamo un processo integrato che obbedisce ad equazioni del tipo:

$$\Phi(B)(I-B)^d Y_t = \Psi(B)\epsilon_t$$

Tutte le equazioni sono prese da (Fattore, 2020).
Il modello ARIMA è capace di adattare il modello se i dati sono stazionari, ovvero se la media e la deviazione standard sono costanti. Il parametro di differenziazione d è l'ordine di trasformazione per rendere il dataset stazionario.

C.2. PROPHET

Il modello FbProphet (Taylor and Letham, 2018) semplifica il lavoro dell'analista il quale, pur essendo esperto di dominio, non ha dimestichezza con l'analisi delle serie storiche o dei differenti metodi e modelli a disposizione.
Tale approccio ha il pregio di fornire parametri intuitivi che agevolano la comprensione dell'utente finale. Il modello ignora la dipendenza temporale dei dati e il training è usato solo come un esercizio per verificare che l'allineamento dei dati sia efficace. Il modello offre molti vantaggi come, ad esempio, accomodare stagionalità multiple e festività conosciute o personalizzate. Il modello mette a disposizione flessibilità offrendo due opzioni per quanto riguarda i trend: 1. Modello lineare; 2. Modello a crescita saturante (non lineare).

- Modello lineare: nel caso di problemi di previsione nei quali non si ha una saturazione nel parametro C, il modello usa un tasso costante di crescita nella funzione;

$$g(t) = (k + a(t)^t\delta)t + (m + a(t)^T\gamma)$$

- Non lineare: per problemi di previsione nei quali la capacità massima di crescita del parametro C non può essere definita, viene indicata come dipendente dal tempo t.

$$g(t) = \frac{C(t)}{1 + e^(- (k + a(t)^T\delta)(t - (m + a(t)^T\gamma)))}$$

L'incertezza del trend futuro viene modellata dando per scontato che il suo tasso rimanga costante. Assumiamo quindi che in futuro avremo la stessa frequenza e grandezza media nel tasso di cambiamento che abbiamo osservato nei dati del passato.
La componente stagionale viene specificata come una serie di Fourier, mentre le festività vengono aggiunte dall'utente tramite una lista di giorni di vacanza. Possiamo utilizzare una funzione indicatrice che rappresenta quando il tempo t cade in un giorno festivo i, al quale viene aggiunto un parametro k_i corrispondente al cambio nella previsione. Tale obiettivo viene raggiunto generando una matrice di regressori $Z(t)$, ottenendo:

$$h(t) = Z(t)k$$

La sua formulazione più generale è tale che:

$$y(t) = g(t) + s(t) + h(t) + e_t$$

dove:

- $g(t)$ è la funzione di trend (tasso di crescita) che modella i cambiamenti non periodici della serie temporale;

- $s(t)$ rappresenta i cambiamenti periodici (stagionalità giornaliere, settimanali, mensili e annuali);
- $h(t)$ rappresenta i periodi di vacanza;
- e_t rappresenta infine ogni cambiamento che il modello non è in grado di catturare, ovvero l' errore, che viene ipotizzato come normalmente distribuito.

4. I DATI

A. Data collection

La raccolta dei dati è stata effettuata dal sito GitHub (Protezione Civile, 2021) della Protezione Civile che è la principale fonte dei dati riguardanti la diffusione della pandemia da COVID-19 in Italia; i dati raccolti sono quelli relativi ai contagi e alla variazione nelle 24 ore delle province considerate. La protezione civile non rende disponibili i dati relativi alle singole città italiane, per ovviare a tale problema si è dovuto procedere con una scelta delle province, che ha seguito un criterio ben preciso: sono state in primo luogo selezionate le 5 province con popolazione più elevata di ogni regione italiana, e successivamente si è proceduto alla creazione di un database composto dai seguenti record:

1. Regione: regione di riferimento;
2. Provincia: province più popolose per ogni regione;
3. Abitanti provincia: numero di abitanti per l'intera provincia;
4. Abitanti città: numero di abitanti della sola città di riferimento;
5. Percentuale: $\frac{abitanti\_citta'}{abitanti\_provincia}100$;

Sono state considerate solo le province con percentuale superiore al 35% :

1. Bologna	9. Ferrara	16. Livorno
2. Milano	10. Parma	17. Terni
3. Rimini	11. Ravenna	18. Pescara
4. Roma	12. Reggio Emilia	19. Crotone
5. Torino	13. Trieste	20. Palermo
6. Asti	14. Firenze	21. Messina
7. Genova	15. Grosseto	22. Cagliari
8. La spezia		

In questo contesto utilizziamo il termine "provincia" anche come sinonimo del recentemente introdotto "città metropolitana", che si applica ad alcune città sopra riportate come Milano, Bologna e Roma.
E' stato adottato tale procedimento per evitare che le città facenti parte della provincia potessero compromettere le analisi successive. Dopo la selezione delle province si è proceduto alla raccolta dei dati del sito GitHub; ottenendo così due database separati: CONTAGI e VARIAZIONE 24H, i due database hanno la seguente struttura:

1. Colonna: numero identificativo della colonna, in formato numerico;
2. Data: data di riferimento in formato data;

Le colonne successive fanno riferimento alla lista di province sopra citate.

I dati presi in considerazione sono dal 24 febbraio 2020 al 23 dicembre 2021 per un totale di 669 giorni.
Nella parte iniziale di entrambi i database sono presenti dei valori mancanti dovuti al fatto che all'inizio della pandemia non in tutte le province erano presenti dei contagi.
I dati contenuti nei due data-set sono stati in primo luogo standardizzati dividendoli per la popolazione, onde evitare che la diversa numerosità della popolazione falsasse i nostri risultati; successivamente abbiamo proceduto anche a normalizzare i dati dividendoli per settimane e calcolando la media,visto che il numero dei tamponi effettuati nel week-end sono minori rispetto ai tamponi effettuati durante il resto della settimana. Di conseguenza durante il fine settimana figurano meno contagi quando semplicemente è stato effettuato un numero inferiore di tamponi.
Infine, i dati relativi all'indice Rt sono stati ricavati da (Guzzetta e Merler, 2020).

B. Data quality

Accuratezza: I dati della Protezione Civile si considerano attendibili; tuttavia tali dati non sono privi di errori, che possono essere attribuiti sia a errori di trascrizione sia al processo di raccolta (quindi di comunicazioni tardive o parziali da parte delle regioni).
Completezza: I dati si considerano completi poiché dopo un'attentata analisi non si è riscontrata la presenza di valori mancanti.
Consistenza:I dati della Protezione Civile garantiscono la consistenza in quanto sono significativi ed effettivamente utilizzabili.

C. City form

La forma urbana è principalmente osservata da come sono disposte le strade e le infrastrutture di trasporto; pertanto le città sono classificate in **lineari**, **a scacchiera** e **radiali**. Le città prese in considerazione si suddividono come segue:

Scacchiera	**Lineare**	**Radiale**
Torino	Asti	Parma
La Spezia	Rimini	Grosseto
Ferrara	Trieste	Milano
Piacenza	Palermo	Bologna
Ravenna	Cagliari	
Terni	Genova	
Roma	Livorno	
Pescara		
Crotone		
Messina		
Firenze		

5. ANALISI/PROCESSO DI TRATTAMENTO DEI DATI

A. Lavorazione dei dati

Come visto nel paragrafo precedente, la raccolta dei dati è stata compiuta tramite il sito GitHub della Protezione Civile. Si è proceduto a scaricare i dati relativi ai contagi e alla variazione nelle 24 ore per ogni singola città scelta per un periodo che va dal 24 febbraio 2020 al 23 dicembre 2021(669 giorni); procedendo successivamente ad integrare tutti i dati utilizzando la libreria Pandas, creando così due data-set separati, uno relativo ai contagi denominato CONTAGI e uno relativo alla variazione nelle 24 ore denominato VARIAZIONE 24H.
Inoltre, abbiamo proceduto a normalizzare i dati con la seguente formula:

$$\frac{Contagi}{Popolazione} * 1000$$

Abbiamo diviso per il numero della popolazione per evitare che la differenza della numerosità della popolazione fra due città ci portasse ad avere risultati non veritieri.

Successivamente è stata presa la decisione di considerare la media per sette giorni perché il numero dei tamponi è strettamente legato al numero dei contagi.
Questo si può osservare nei giorni lavorativi dove i contagiati tendono ad essere più elevati rispetto ai giorni feriali dove i tamponi eseguiti sono molti meno.
Come visto nel paragrafo precedente all'inizio dei due data-set ci sono dei valori mancanti che sono dovuti al fatto che all'inizio della pandemia non in tutte le province fossero presenti dei contagi.

I dati dell'indice Rt provenienti dal sito CovidStat INFN delle provincie interessate dallo studio presentavano alcuni valori mancanti che sono stati sostituiti attraverso l'utilizzo del filtro di Kalman, che permette di stimare i valori mancanti.

B. Clustering

Per effettuare le analisi di clustering sul data-set VARIAZIONE 24H abbiamo utilizzato R ed in particolare il pacchetto **TSclust**, le distanze che abbiamo impiegato sono :

- Distanza Euclidea: valuta la discrepanza da punto a punto enfatizzando le differenze di forme locali.
- CID: ha lo scopo di confrontare le complessità di forma.
- CDM: misura la dissimilarità in base alla dimensioni delle serie compresse.

Per ogni calcolo delle distanze sono stati utilizzati 3 diversi metodi di agglomerazione necessari per il calcolo della distanza dai cluster. I metodi utilizzati sono i seguenti:

- Complete: calcola tutte le dissomiglianze a coppie tra gli elementi nel cluster 1 e gli elementi nel cluster 2 e considera il valore più grande (cioè il valore massimo) di queste dissomiglianze come la distanza tra i due cluster. Tende a produrre grappoli più compatti.
- Mcquitty: Weighted Pair Group Method with Arithmetic Mean (WPGMA) metodo che costruisce un dendrogramma che riflette la struttura presente in una matrice di distanza a coppie (o matrice di similarità).
- Centroid: metodo che performa il baricentro sulla matrice delle distanze.

B.1. Valutazione dei cluster

Per le valutazioni delle distanze è stato sempre utilizzato il pacchetto **Tsclust** di R.
A seguito della lavorazione del data-set vista nel paragrafo 5.A abbiamo proceduto alla creazione di un cluster di confronto **true cluster**. Possiamo osservare i true cluster nel pagarafo 4.C ; essi sono stati individuati principalmente tramite un'attenta osservazione della forma urbana delle città nelle cartine.
I true cluster sono stati utilizzati per effettuare una valutazione sulla qualità dei cluster ottenuti con le varie metriche utilizzate.

C. Correlazioni e differenze tra le categorie di piante cittadine

Prima delle analisi è stato necessario procedere con il calcolo dei valori medi dell'*indice Rt* per ogni città e partendo dagli ultimi dati disponibili nel Dataset CONTAGI è stato estrapolato *l'infection rate cumulativo*. L'Rt indica l'indice di trasmissibilità, ovvero un parametro che misura la potenziale trasmissibilità di una malattia infettiva; questo valore indica quante persone possono essere contagiate da una sola persona in media e in un certo periodi di tempo. In questo calcolo sono inoltre presi in considerazione gli effetti delle misure di contenimento, in modo tale da permettere di valutare l'efficacia nel tempo delle misure adottate per limitare la diffusione del coronavirus(Guzzetta G., Merler S.,2020). Il motivo alla base della scelta di questo indice è che esso si basa su una semplice statistica riguardante i dati delle nuove infezioni giornaliere e non richiede particolari ipotesi sui meccanismi di trasmissione del virus. Per le stime vengono utilizzate le curve epidemiche dei casi sintomatici a partire dalla data di inizio dei sintomi. Per poter palare di indice di trasmissibilità si deve prima fare riferimento alla stima della trasmissibilità giornaliera da casi sintomatici $G(t)$, che è basata su un algoritmo Markov Chain Monte Carlo. Quest'ultimo permette di stimare la distribuzione a posteriori della trasmissibilità giornaliera $G(t)$ ad ogni tempo t.
Considerato che $G(t)$ può essere influenzata da eventi particolari, si considera più rappresentativo il valore medio dell'ultima settimana. Da questo viene ottenuto l'$R(t)$, sintetizzato della seguente formula, (Stewart, 2022):

$$R(t) = \frac{1}{7}\sum_{s=0}^{6} G(t-s)$$

I dati dell'$R(t)$ per provincia da utilizzare sono stati messi a disposizione dall'istituto Superiore di Sanità. L'acquisizione dei dati epidemiologici è affetta da una serie di ritardi, alcuni dei quali non trascurabili: in particolare, il tempo tra l'evento infettivo e lo sviluppo dei sintomi(tempo di incubazione), quello tra sintomi e l'esecuzione del tampone, quello tra l'esecuzione del tampone e la conferma di positività, e quello tra la conferma di positività e l'inserimento nel sistema di sorveglianza integrata dell'Istituto Superiore di Sanità. Prendendo in considerazione questi aspetti, abbiamo deciso di utilizzare l'infection rate cumulato che viene sintetizzato dalla seguente formula:

$$CumulativeInfectionRate = \frac{TotalCases}{PopulationSize\frac{1}{10^5}}$$

Questo tasso viene utilizzato per misurare la frequenza di insorgenza di nuovi casi di infezione all'interno di una popolazione durante un periodo di tempo specifico. Dopo aver definito il valore medio dell'indice Rt per ogni città presa in considerazione dallo studio si è passati alla valutazione della correlazione dell'indice con la forma della città, per cui attraverso la

libreria *Python scipy* e alla funzione *spearmanr* è stato possibile calcolare il valore della statistica F e il p-value abbinato. Successivamente è stata valutata anche per la variabile *infection rate cumulato* definito per ogni città, il valore della statistica F e il p-value associato. Quindi è stato possibile passare alla fase di valutazione mediante *test ANOVA* dell'ipotesi Nulla facente riferimento all'uguaglianza delle medie all'interno dei tre gruppi di riferimento. Quindi attraverso la libreria *Pandas* e utilizzando il metodo *corr* è stata valutata la correlazione attraverso il coefficiente di correlazione di *Pearson* tra le tre categorie presenti nello studio. Attraverso la libreria *scipy* e la funzione *chi2_contingency* è stato preso in considerazione l'ipotesi di indipendenza tra le coppie delle tabelle di contingenza , e utilizzando la funzione *shapiro* sempre proveniente dalla libreria Scipy è stato possibile testare la normalità dei dati all'interno dei gruppi ed infine è stato possibile utilizzare il test ANOVA. Poichè le assunzioni di *indipendenza, omoschedasticità e normalità* sono state rispettate. Per la variabile infection rate cumulato è stata seguita la stessa procedura e anche in questo caso le assunzioni sono state rispettate.

D. Forecasting

Per effettuare l'analisi predittiva riguardo al futuro andamento dei contagi abbiamo utilizzato i modelli ARIMA e FbProphet, per capire quale modello fosse il più adatto allo scopo. Raccogliendo i dati dall'inizio della pandemia (guardando agli aggiornamenti quotidiani e non a quelli cumulativi), abbiamo cercato delle differenze nelle curve dei contagi a seconda della city shape; per fare ciò, abbiamo preso in esame una città per ogni cluster, ovvero Bologna (scacchiera), Trieste (radiale), Rimini (lineare).
Prendiamo ora in esempio il modello ARIMA. Questo modello si presta ottimamente allo scopo di prevedere i dati futuri utilizzando le serie storiche. Come delineato da studi precedenti (Kumar e Susan, 2020) il modello ARIMA è il modello d'elezione per quanto riguarda l'analisi dei dati sulla pandemia tuttora in atto; questo poiché abbiamo numerosi e attendibili dati sulla stessa.
Abbiamo effettuato delle analisi preventive utilizzando le librerie Pandas, Numpy, Matplotlib e Statsmodels.
Innanzitutto, abbiamo decomposto le serie storiche al fine di individuare trend e stagionalità delle stesse.
Per adoperare il modello ARIMA, la serie storica deve essere stazionaria. Per assicurarci di ciò abbiamo impiegato il test di Dickey-Fuller; quando il p-value è minore di 0.05 possiamo considerare la serie storica come stazionaria. Abbiamo notato che il dataset da noi adoperato aveva un p-value più alto di 0.05 e abbiamo così proceduto ad integrare la serie storica.
Per individuare gli ordini p,i,q del modello abbiamo utilizzato la funzione auto_arima. Questa funzione è contenuta in Pmdarima, una libreria statistica designata per colmare il vuoto nella capacità d'analisi delle serie storiche di Python. Questa funzione è l'equivalente della funzione auto.arima di R, possiede numerose *utilties* come la differenziazione, una collezione di test statistici della stazionarietà e della stagionalità e *utilities* per la cross-validation.
La funziona auto_arima cerca e identifica i migliori parametri per il modello ARIMA, conducendo numerosi test per identificare il parametro di differenziazione i, e i parametri AR(p) e MA(q). In caso di stagionalità esso cerca di identificare anche i parametri P,Q e D.
Per trovare il miglior modello, la funzione ottimizza per una data informazione i valori di "aic", "bic","aicc", "oob", e "hqic", che corrispondono rispettivamente Akaike Information Criterion,Bayesian Information Criterion, Corrected Akaike Information Criterion, HannanQuinn Information Criterion.
Il migliore modello ARIMA è quello che minimizza uno di questi valori.
La metrica usata per il confronto è stata la "Mean Absolute Percentage Error" (MAPE), che misura l'accuratezza delle previsioni secondo la seguente formula:

$$MAPE = \frac{100}{n} \sum_{t=1}^{N} \left| \frac{A_t - F_t}{A_t} \right|$$

dove il parametro A_t rappresenta la serie dei valori reali mentre il parametro F_t rappresenta la serie fittizia delle previsioni.

Il secondo modello da noi utilizzato è FbProphet, da noi scelto per via della sua semplicità di utilizzo e implementazione.
Come per il modello precedente, abbiamo importato i dati utilizzando la libreria Pandas. Per adoperare questo modello bisogna inizialmente rinominare la colonna contenente la data delle osservazioni come "ds" e la colonna contenente i valori come "y".
Tramite FbProphet abbiamo implementato due modelli di forecasting.
Il primo è un modello semplice, che ha come unico parametro la stagionalità settimanale.
Nel secondo modello sono stati aggiunti in un primo momento i giorni festivi "rossi" italiani, per notare eventuali cambiamenti nel trend nei suddetti giorni; in un secondo momento abbiamo aggiunto anche alcuni periodi "notevoli" dovuti alle restrizioni governative, come il lockdown totale (2020/03/08 – 2020/05/18) ed alcuni periodi ricadenti sotto i decreti che hanno diviso l'Italia in zone di diversi colori con diverse restrizioni.
Abbiamo utilizzato la cross-validation per valutare entrambi i modelli. Questo è stato possibile perché la libreria Prophet include delle funzionalità per la cross-validation delle serie storiche per misurare l'errore nelle previsioni stando ai dati storici; ciò avviene selezionando punti di interruzione casuali nella serie storica, e adattando il modello per ciascuno fino al punto di interruzione.
Come esempio, abbiamo deciso di utilizzare la città di Bologna, scelta come "rappresentante" delle città a griglia.
Per avere un metro di giudizio sulle performance del modello abbiamo utilizzato MAPE, analogamente al caso del modello ARIMA, questa volta attraverso la cross-validation.
Tenendo conto sia dei valori di MAPE che della lettura dei grafici, conveniamo che il modello che ci permette di predire con maggiore efficacia i dati futuri è il secondo.

6. RISULTATI

A. Risultati del clustering

La valutazione delle metriche utilizzate ha restituito i seguenti risultati:

	Distanza euclidea	**CID**	**CDM**
Complete	0,473	0,425	0,466
Mcquitty	0,446	0,473	0,465
Centroid	0,471	0,473	0,465

I risultati derivanti dalla valutazione del processo di clustering sono compresi tra 0 e 1 e ammettono una semplice interpretazione: tanto più sono vicini a 1 tanto più le prestazioni delle metriche sono migliori.
Di seguito vengono riportati i cluster migliori per ogni distanza :

- Distanza euclidea: il metodo che si è rivelato essere il migliore è il *complete* e i cluster sono i seguenti:

```
X24H_Bologna    X24H_Milano    X24H_Rimini      X24H_Roma    X24H_Torino      X24H_Asti
           1              1              2              1              1              2
 X24H_Genova X24H_La.Spezia   X24H_Ferrara     X24H_Parma   X24H_Ravenna   X24H_Trieste
           1              2              2              1              2              3
X24H_Firenze  X24H_Grosseto   X24H_Livorno     X24H_Terni   X24H_Pescara   X24H_Crotone
           1              2              2              2              2              2
X24H_Palermo   X24H_Messina  X24H_Cagliari  X24H_Piacenza
           1              1              1              2
```

- CID : i metodi migliori sono risultati *mcquitty* e *centroid*, inoltre hanno custerizzato le città nello stesso modo in quanto hanno il medesimo risultato; i cluster sono i seguenti:

```
X24H_Bologna    X24H_Milano    X24H_Rimini      X24H_Roma    X24H_Torino      X24H_Asti
           1              1              2              1              1              2
 X24H_Genova X24H_La.Spezia   X24H_Ferrara     X24H_Parma   X24H_Ravenna   X24H_Trieste
           1              2              2              1              2              3
X24H_Firenze  X24H_Grosseto   X24H_Livorno     X24H_Terni   X24H_Pescara   X24H_Crotone
           1              2              2              2              2              2
X24H_Palermo   X24H_Messina  X24H_Cagliari  X24H_Piacenza
           1              1              1              2
```

- CDM : il metodo che si è rivelato essere il migliore è il *complete* e i cluster sono i seguenti:

```
X24H_Bologna    X24H_Milano    X24H_Rimini      X24H_Roma    X24H_Torino      X24H_Asti
           1              1              2              1              1              2
 X24H_Genova X24H_La.Spezia   X24H_Ferrara     X24H_Parma   X24H_Ravenna   X24H_Trieste
           1              2              2              1              2              3
X24H_Firenze  X24H_Grosseto   X24H_Livorno     X24H_Terni   X24H_Pescara   X24H_Crotone
           1              2              2              2              2              2
X24H_Palermo   X24H_Messina  X24H_Cagliari  X24H_Piacenza
           1              1              1              2
```

Nei cluster risultanti 1 indica ha una forma urbana a scacchiera, 2 indica che la città ha una forma urbana lineare e 3 indica che la città ha una forma urbana radiale.

B. Risultati correlazioni

Di seguito vengono mostrati i risultati del Coefficiente di correlazione per ranghi di Spearman e i relativi p-value:

- Correlazione - Coefficiente di correlazione per ranghi di Spearman - Indice Rt

	F value	**p-value**
Indice Rt	0.122	0.886

- Correlazione - Coefficiente di correlazione per ranghi di Spearman - Infection rate cumulato

	F value	**p-value**
Infection rate cumulato	0.114	0.611

I due test non sono statisticamente significativi, di conseguenza non possono essere rigettate l'ipotesi H0(presenza di correlazione tra le variabili e la pianta cittadina). Provando comunque a descrivere i risultati notiamo valori del coefficiente di Spearman vicini allo zero che indicano una assenza di correlazione. Da questi risultati non possono essere definite delle correlazioni che implichino che la forma delle città incida nel numero di contagi.

C. Risultati differenze tra le categorie di piante cittadine

Di seguito vengono mostrati i risultati delle assunzioni che permettono di applicare il test ANOVA:

- Correlazione - Coefficiente di correlazione di Pearson - Indice Rt

	Lineare	**Radiale**	**Scacchiera**
Lineare	1	0.775	-0.0508
Radiale	0.775	1	0.601
Scacchiera	-0.0508	0.601	1

- Correlazione - Coefficiente di correlazione di Pearson - Infection rate cumulato

	Lineare	**Radiale**	**Scacchiera**
Lineare	1	-0.398	-0.137
Radiale	-0.398	1	0.223
Scacchiera	-0.137	0.223	1

- Indipendenza - Test chi-quadro - Indice Rt

	chi2	**p-value**
Lineare e Radiali	0.050	0.213
Radiale e Scacchiera	0.050	0.213
Lineari e Scacchiera	0.050	0.227

- Indipendenza - Test chi-quadro - Infection rate cumulato

	chi2	**p-value**
Lineare e Radiali	0.050	0.213
Radiale e Scacchiera	0.050	0.213
Lineari e Scacchiera	0.050	0.227

- Normalità - Shapiro-Wilk Test - Indice Rt

	Shapiro-Wilk	**p-value**
Scacchiera	0.972	0.905
Radiale	0.981	0.906
Lineari	0.887	0.258

- Normalità - Shapiro-Wilk Test - Infection rate cumulato

	Shapiro-Wilk	**p-value**
Scacchiera	0.956	0.720
Radiale	0.850	0.124
Lineari	0.952	0.729

- Omoschedasticità - Levene test - Indice Rt

	Levene test value	**p-value**
Indice Rt	0.910	0.420

- Omoschedasticità - Levene test - Infection rate cumulato

	Levene test value	**p-value**
Infection rate cumulato	1.193	0.325

Le assunzioni per l'applicazione del test ANOVA sono state rispettate sia per l'indice Rt che per Infection rate cumulato. Di seguito possiamo osservare i risultati del test ANOVA:

- Differenza in Media - Test ANOVA - Indice Rt

	F value	**p-value**
Indice Rt	0.122	0.886

- Differenza in Media - Test ANOVA - Infection rate cumulato

	F value	**p-value**
Infection rate cumulato	0.695	0.512

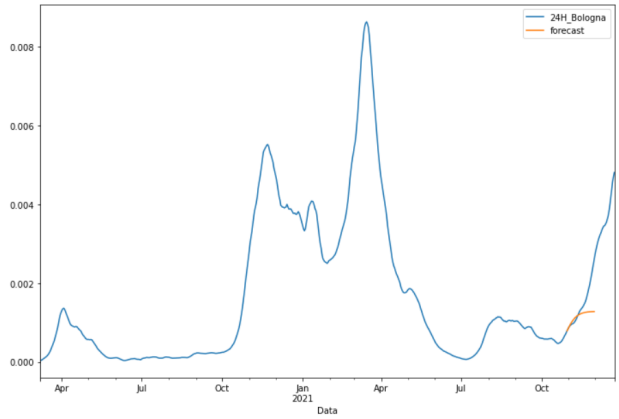
I risultati dei test ANOVA non sono significativi, poichè presentano un p-value eccessivamente grande, questo quindi non permette di rigettare l'ipotesi H0 (L'ipotesi nulla prevede che i dati di tutti i gruppi abbiano la stessa origine, ovvero la stessa distribuzione stocastica, e che le differenze osservate tra i gruppi siano dovute solo al caso).

D. Risultati Forecasting

Di seguito abbiamo riportato i risultati del forecasting.:

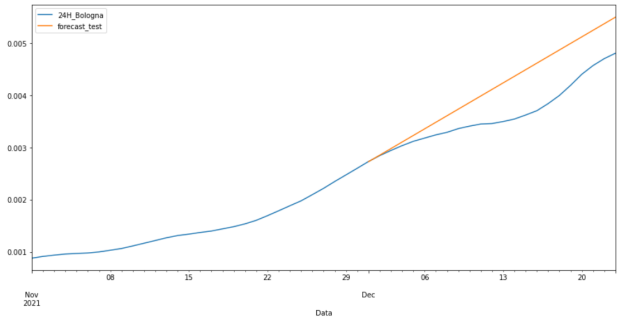
Come possiamo vedere nella **figura 1**, il modello ARIMA è stato utilizzato per "prevedere" l'andamento della pandemia tra la fine di ottobre e la fine di novembre 2021. Per questa predizione è stato utilizzato l'intero dataset a disposizione. Possiamo notare che il modello si adatta bene al mese di ottobre, mentre si distacca sensibilmente dai dati reali durante il mese di novembre.
Il valore del MAPE in questo caso è di circa 0.161822.
Un'ulteriore previsione è stata svolta tramite il modello ARIMA partendo da un dataset più ristretto, in modo tale da tentare di

Fig. 1. Modello ARIMA(5,1,0)



prevedere il picco dei contagi verificatosi nel periodo natalizio. In questo caso, il dataset su cui trainare il modello è stato ristretto ai soli mesi di ottobre, novembre e dicembre. Come possiamo notare dalla **figura 2**, dove la reale numerosità dei contagi viene sovrastimata, al contrario che nel precedente grafico.

Fig. 2. Modello ARIMA (2021/10/30 - 2021/12/30)

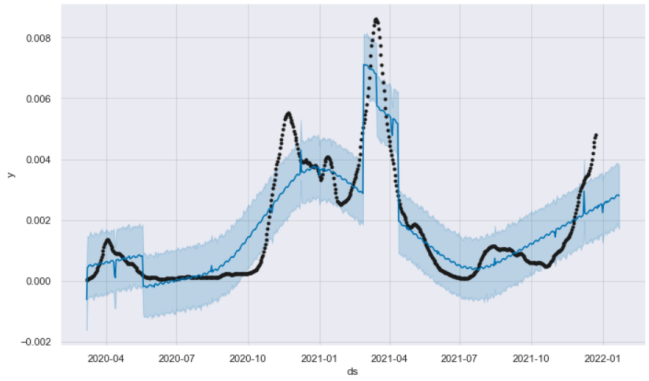


Abbiamo ricalcolato il valore del MAPE che è circa 0.133183.

Veniamo ora alla **figura 3**, in cui è riportato il modello generale prodotto da FbProphet. Sull'asse x sono riportate le date (chiamate appunto "ds" da FbProphet) mentre sull'asse y abbiamo i contagi nelle 24 h normalizzati. I punti neri nel grafico rappresentano i casi reali osservati, mentre la linea azzurra rappresenta i valori presentati dal modello Prophet.
Come possiamo vedere il modello si adatta abbastanza bene ai dati, tranne per alcuni picchi improvvisi come quello dell'autunno 2020. Possiamo inoltre constatare che, come per il modello precedente, anch'esso fatica a prevedere il reale andamento dei contagi futuri, sottostimandolo.
Abbiamo quindi ottenuto un valore di MAPE di circa 0.82603.

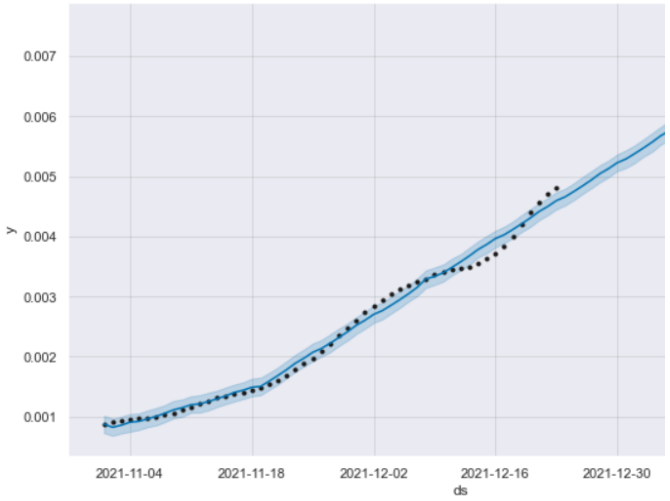
Anche in questo caso abbiamo provato a restringere il dataset, prendendo in considerazione i soli mesi di ottobre, novembre e dicembre.
Come per il modello ARIMA, anche FbProphet si adatta meglio ai dati prendendo in considerazione un intervallo di tempo ristretto. Questo può essere dovuto al fatto che, utilizzando dati che riguardano mesi di lento ma costante innalzamento dei casi di contagio, il modello non deve adattarsi o prevedere alcun

Fig. 3. Modello Prophet



picco improvviso o "anomalo". **Figura 4** Il valore MAPE qui corrisponde a 0,04879.

Fig. 4. Modello Prophet (2021/10/30 - 2021/12/30)



Nonostante il secondo modello di FbProphet sia capace di tenere in considerazione festività e eventi notevoli, reputiamo più funzionale il modello ARIMA per via degli incoraggianti valori numerici per la metrica MAPE.
Si può notare, tuttavia, che utilizzando un dataset ristretto il valore MAPE diventa inferiore per il modello Prophet, che quindi riteniamo più affidabile in questo caso.

	MAPE Trieste	**MAPE Rimini**
ARIMA	0.5054	0.3396
PROPHET	0.9686	0.7092

7. CONCLUSIONE E POSSIBILI SVILUPPI

Tirando le somme del lavoro da noi svolto, possiamo innanzitutto affermare che lo studio dello sviluppo e l'espansione di una malattia epidemica rappresenta un'ardua sfida per chiunque si cimenti in questo compito. Abbiamo potuto notare, attraverso i nostri modelli di previsione, che gli sviluppi futuri della curva dei contagi può contenere inaspettate quanto dolorose sorprese; mentre la curva prevista dal nostro modello era tutto sommato modesta, nella realtà abbiamo avuto un aumento dei casi di COVID-19 ben al di sopra delle aspettative. Il comportamento umano e le nuove mutazioni dei vari ceppi del virus Sars-Cov-2 possono andare oltre ciò che si può calcolare utilizzando una mole nient'affatto ridotta di dati.
Procedendo con il calcolo della correlazione tra l'indice Rt e la forma delle città prese in esame, e poi tra quest'ultima e l'indice infection rate, abbiamo potuto notare che una correlazione tra i due fattori risulta non esistere. Questo potrebbe anche accadere poiché, non avendo usato un disegno multifattoriale, non abbiamo avuto modo di analizzare tutte le variabili in gioco in un caso così delicato.
Utilizzando un maggior numero di variabili, si potrebbe probabilmente avere una conoscenza più accurata della reale correlazione tra le variabili da noi studiate. Si potrebbero inoltre esaminare molte altre caratteristiche dei centri urbani che causano, più o meno direttamente, variazioni nella curva dei contagi: peculiarità della popolazione, come l'età media o la percentuale di anziani; indici che misurano il grado di affollamento di una città come la densità abitativa e perfino la qualità e l'offerta del trasporto pubblico possono determinare una diversa diffusione del virus in questione.
Non vogliamo comunque smentire del tutto la veridicità della nostra ipotesi, avanzata peraltro da numerosi studi, e ci piace pensare che uno sviluppo urbano più razionale e sostenibile potrebbe aumentare la qualità della vita di ognuno di noi e aiutarci a scongiurare il pericolo di una futura diffusione capillare di un'altra pandemia.

8. RIFERIMENTI BIBLIOGRAFICI

AbouKorin S.A., Haoying H., Mahran M.(2020). *Role of urban planning characteristics in forming pandemic resilient cities – Case study of Covid-19 impacts on European cities within England, Germany and Italy*. Disponibile su reasearchgate.net.

Fattore M., 2020. *Fundamentals of time series analysis, for the working data scientist*. Resa disponibile per il corso di laurea in Data Science.

Guzzetta G., Merler S. (2020) *Stime della trasmissibilità di SarS-CoV-2 in Italia*. Disponibile nella sezione "Open Data" di www.epicentro.iss.it.

Kumar N., Susan S.. (2020). *COVID-19 Pandemic Prediction using Time Series Forecasting Models*, 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020.

Presidenza del Consiglio dei Ministri - Dipartimento della Protezione Civile ,(2021). *dati-province Covid*. Disponibile su https://github.com/pcm-dpc/COVID-19/tree/master/dati-province

Montero P., Vilar J. (2014). *TSclust: An R Package for Time Series Clustering*, Journal of Statistical Software, Vol.62, Issue 1.

Véron J. (2008). *L'urbanizzazione del mondo*, Il Mulino.

Stewart C., (2022). *COVID-19 infection rate in Italy as of January 2022, by region*. Disponibile su www.statista.com.

Taylor SJ, Letham B. (2017). *Forecasting at scale*. Disponibile su peerj.com.