

Analisi di serie storiche per ristoranti

Giorgio Carbone (matr. 811974)¹, Gianluca Cavallaro (matr. 826049)², Remo Marconzini (matr. 883256)³, Gianluca Scuri (matr. 886725)⁴

Sommario

Il progetto proposto ha come obiettivo quello di analizzare l'andamento delle vendite di 6 ristoranti collocati in Lombardia ed Emilia-Romagna, in un periodo temporale che va dal gennaio 2018 all'aprile 2022. Attraverso i dati forniti, opportunamente integrati con ulteriori informazioni utili, si vuole cercare di rispondere a tre domande: presenza di pattern significativi nelle serie storiche, fornire una stima delle perdite accusate durante il periodo di chiusura dovuto alla pandemia da COVID-19 e fornire una stima dell'andamento futuro dei ristoranti. Viene dapprima condotta una analisi esplorativa delle serie storiche dei diversi ristoranti, con l'obiettivo di individuare e spiegare alcuni comportamenti, quali ad esempio stagionalità e trend. Successivamente, vengono implementati diversi modelli per l'analisi e la previsione di serie storiche (ARIMA & SARIMA, UCM e Random Forest). I più robusti ed accurati per questa applicazione si sono rivelati essere i modelli SARIMA e Random Forest. Questi hanno permesso di stimare delle perdite comprese tra il 16% e il 21% rispetto al fatturato annuo del 2019 a causa del lockdown nella primavera del 2020, ed hanno permesso di stimare l'andamento futuro della serie storica.

Keywords

Data Science – Time Series Analysis – ARIMA – UCM – Random Forest

^{1,2,3,4}Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli studi di Milano-Bicocca, Milano, Italia

Indice

1	Introduzione	1
2	Domande di ricerca	2
3	Dati	2
3.1	Data Integration	2
4	Analisi esplorativa	3
4.1	Ristorante 1	3
	Missing values • Vendite • Scontrini • Correlazione con altri attributi	
4.2	Altri ristoranti	6
5	Analisi predittiva	7
5.1	Analisi della serie storica pre-Covid ai fini di stimare le possibili perdite subite durante il primo lockdown	7
	Analisi esplorativa serie storica pre-Covid • Modelli ARIMA • Modelli UCM • Modelli di Machine Learning • Time Series Cross-Validation: valutazione accuratezza puntuale di previsione e confronto tra modelli • Previsioni e considerazioni	
5.2	Analisi della serie storica post primo lockdown ai fini di prevedere il livello di fatturato futuro	17
	Analisi esplorativa della serie storica post 1° lockdown • Modelli ARIMA • Modelli UCM • Modelli di Machine Learning • Time Series Cross-Validation: valutazione accuratezza puntuale di previsione e confronto tra modelli • Previsioni e considerazioni	
6	Conclusioni	24

1. Introduzione

Con la diffusione del COVID-19 all'inizio del 2020, molte attività commerciali sono state costrette ad affrontare un periodo economicamente molto complicato a causa delle misure restrittive implementate per contenere la pandemia. Nonostante il primo caso accertato in Cina risalga al 6 gennaio 2020, in Italia gli effetti della pandemia si sono manifestati dalla seconda metà

di febbraio 2020, con le prime chiusure di scuole e università, preludio del lockdown completo iniziato il 9 marzo. I due anni successivi sono stati caratterizzati da continue chiusure e riaperture, nonché modifiche delle misure restrittive. Fra tutti i settori colpiti duramente dagli effetti della pandemia, quello della ristorazione rappresenta un caso emblematico. Dapprima le chiusure, dopodiché una lenta ripartenza, spesso con posti limitati, oltre che evidenti cambiamenti nelle abitudini dei consumatori, hanno causato ingenti perdite e, nei casi più gravi, chiusure definitive. Per questo motivo, l'obiettivo di questo progetto è prendere in esame i 6 ristoranti a disposizione per quantificare l'impatto del COVID e per prevedere l'andamento futuro delle vendite, nel periodo di ritorno alla normalità.

Per condurre le analisi desiderate è stata dapprima eseguita una fase di pre-processing dei dati a disposizione, con l'obiettivo principale di integrare il dataset iniziale con informazioni potenzialmente utili alla caratterizzazione della serie storica.

Successivamente, la fase di data exploration ha permesso di valutare l'andamento delle vendite dei singoli ristoranti, confrontando diversi periodi temporali ed analizzandone la correlazione con le informazioni inserite.

Infine, verranno implementati diversi modelli per l'analisi delle serie storiche, con l'obiettivo di prevedere l'andamento della serie nel periodo pandemico e realizzare previsioni future.

2. Domande di ricerca

Le domande di ricerca proposte sono le seguenti:

1. Quali sono le principali caratteristiche dei diversi ristoranti? Tramite un'analisi esplorativa, è possibile individuare dei pattern significativi nelle serie storiche a disposizione?
2. Quale sarebbe stato il livello atteso di vendite per i vari ristoranti senza la pandemia da COVID-19? È possibile stimare, implementando diversi modelli per l'analisi di serie storiche, le perdite accusate dai ristoranti durante il periodo pandemico?
3. Quale sarà il livello delle vendite dei ristoranti nei mesi successivi ad Aprile 2022? Qual è il modello più affidabile nella stima del livello futuro?

3. Dati

Le serie storiche riguardano 6 ristoranti situati nel nord Italia, fra Lombardia ed Emilia-Romagna, in particolare a Piacenza, Montebello della Battaglia, Voghera e Stradella. Il dataset iniziale contiene 4 attributi:

- **data**: attributo contenente la data a cui fanno riferimento le informazioni, espressa in formato *yyyy-mm-dd*;
- **scontrini**: attributo quantitativo discreto che rappresenta il numero di scontrini emessi nel dato giorno;
- **lordototale**: attributo contenente l'incasso lordo giornaliero;
- **id_ristorante**: attributo qualitativo contenente il codice di riferimento per ciascun ristorante.

3.1 Data Integration

Al fine di eseguire un'analisi completa delle serie storiche a disposizione, sono stati individuati fattori potenzialmente correlati all'andamento degli incassi. L'integrazione del dataset iniziale ha visto l'aggiunta dei seguenti attributi:

- **Location, Regione e Provincia**: attributi generati a partire dall'attributo **id_ristorante** e dal match con le informazioni fornite circa la posizione del ristorante;
- **Giorno, Mese e Anno**: attributi qualitativi generati a partire da **data**;
- **Weekend**: attributo booleano che assume valore **True** se il giorno in questione è un sabato o una domenica, **False** altrimenti;
- **Festivo**: attributo booleano che assume valore **True** se il giorno in questione è un festivo, **False** altrimenti. Fra i festivi sono state considerate le domeniche, i giorni riconosciuti come feste a livello nazionale e le feste dei patroni locali;
- **Colore**: attributo qualitativo che, in base alla regione di appartenenza del ristorante, indica il colore attribuito alla regione come misura di contenimento del COVID-19, a seconda del giorno considerato;

- **Pioggia:** attributo booleano che assume valore **True** se nel giorno in questione ci sono state precipitazioni, **False** altrimenti. I dati meteo sono stati presi da ARPA Lombardia [inserire riferimento] e ARPA Emilia-Romagna [inserire riferimento], considerano le stazioni di rilevazione più vicine alla posizione dei ristoranti;

Tutti gli attributi aggiunti sono stati definiti sia per il periodo coperto dai dati forniti sia per il periodo successivo, consentendo di effettuare le opportune previsioni.

4. Analisi esplorativa

Come prima cosa, è stata condotta un'analisi esplorativa approfondita dei dati dei singoli ristoranti. Di seguito viene riportata l'analisi effettuata sul primo ristorante. In sezione 4.2 vengono invece riportate ulteriori considerazioni e differenze degli altri ristoranti analizzati.

4.1 Ristorante 1

4.1.1 Missing values

Inizialmente, è stata indagata la presenza di valori mancanti nel dataset. In particolare, sono stati individuati 300 valori mancanti per l'attributo **scontrini**. I primi 235 dipendono dal fatto che i dati del lordo totale per i primi 8 mesi sono aggregati mensilmente e, quindi, presentano un solo valore al mese. I restanti valori mancanti sono da ricondurre a chiusure dovute al COVID-19 e ad alcune festività.

Per quanto appena detto, per tutte le analisi esplorative successive sono stati scartati i dati dei primi 8 mesi.

4.1.2 Vendite

La prima analisi è stata condotta sull'attributo **lordototale**. Viene mostrata sia la serie storica con dati giornalieri che quelle con dati medi settimanali e mensili.

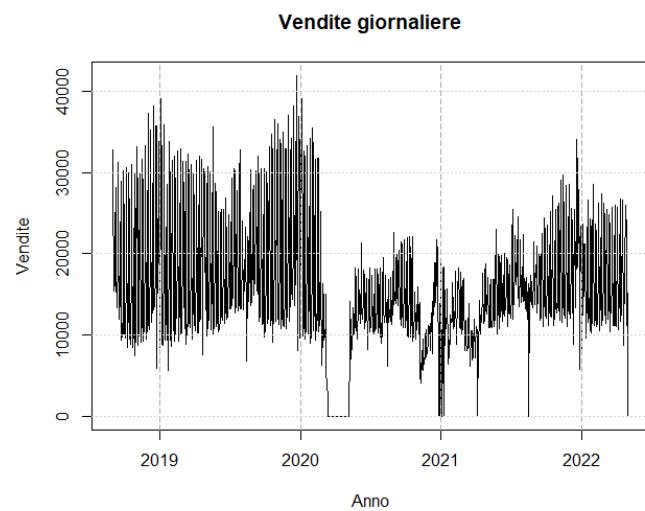


Figura 1. Vendite giornaliere

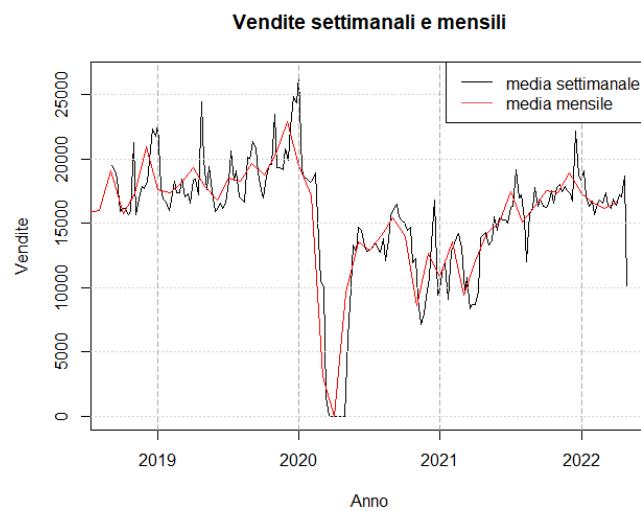


Figura 2. Vendite medie mensili e settimanali

Il periodo più interessante è sicuramente quello dei primi mesi del 2020. A causa dello scoppio della pandemia, si nota dapprima una diminuzione delle vendite e poi un periodo di chiusura completa, corrispondente al primo lockdown. Le vendite ripresero nella seconda metà del maggio 2020, riprendono a rilento, a causa della situazione generale di incertezza che ha caratterizzato la fine del 2020 e tutto il 2021: i valori di vendita, infatti, risultano decisamente inferiori al periodo precedente alla pandemia. Negli ultimi mesi del 2020 si registra un ulteriore abbassamento delle vendite, in corrispondenza della seconda ondata pandemica.

Lo stesso andamento si nota anche osservando le medie settimanali e mensili, che mostrano una lenta risalita solo nella seconda parte del 2021.

Successivamente sono stati analizzati gli andamenti dei singoli anni. Dal 2018 al 2022 sono state generate delle serie storiche per le vendite medie settimanali e mensili. In entrambi i grafici si può notare una certa stagionalità. In figura 1 si possono notare dei picchi intorno alla 17esima, alla 29esima, alla 42esima e alla 51esima settimana dell'anno. Tali picchi si mostrano

sia per gli anni precedenti alla pandemia che per quelli colpiti dal COVID, ad eccezione di quelli appartenenti alle ondate più significative. Si può evidenziare come il 2021, specialmente nelle ultime settimane dell'anno, abbia segnato un ritorno a livelli simili a quelli precedenti alla pandemia, con una tendenza proseguita poi nei primi mesi del 2022.

Anche considerando la media mensile delle vendite si può notare un andamento coerente con quello appena descritto. I picchi principali, registrati nei mesi di gennaio, aprile, luglio, settembre e dicembre, sono presenti sia prima che durante la pandemia, nonostante il livello raggiunto nel 2020, 2021 e 2022 sia inferiore rispetto agli anni precedenti. Di nuovo, sono assenti i picchi di aprile per il 2020 e 2021, essendo due dei periodi più critici della pandemia.

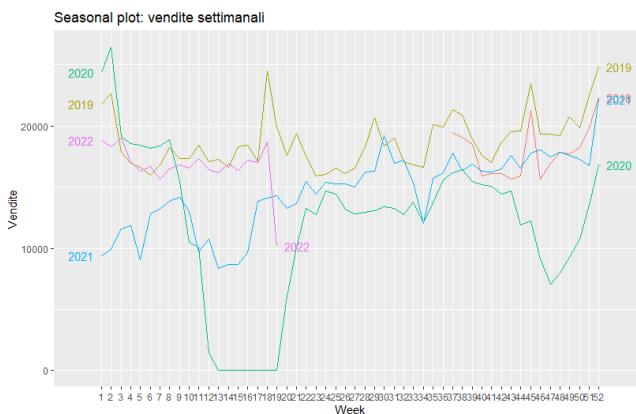


Figura 3. Stagionalità settimanale

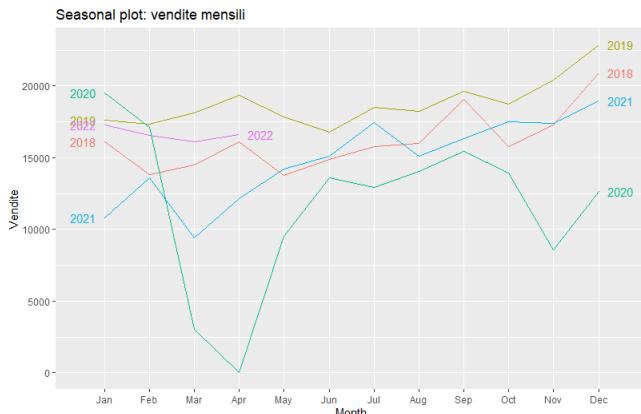


Figura 4. Stagionalità mensile

A conferma del comportamento appena descritto, sono stati realizzati anche dei *seasonal subseries plot*, un grafico che permette di enfatizzare la stagionalità all'interno dei dati. Vengono riportati sia il grafico riferito all'intero periodo dei dati che quello riferito solo al periodo precedente alla pandemia:

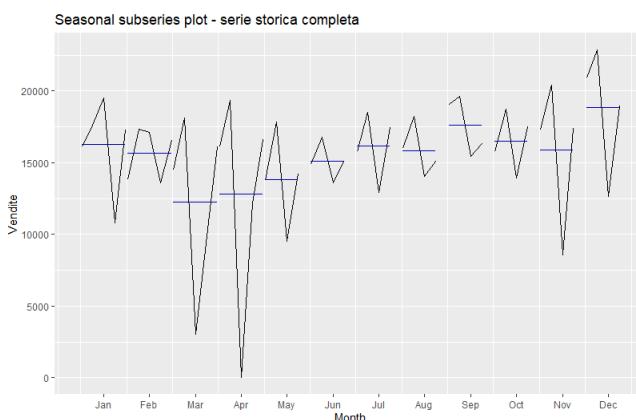


Figura 5. Seasonal sub-plot completo, periodo 2018-2022

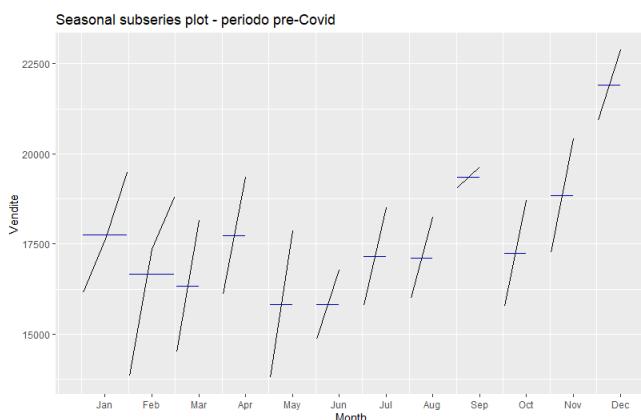


Figura 6. Seasonal sub-plot pre-Covid, anni 2018 e 2019

Dalla figura 5 possiamo notare come i mesi caratterizzati da valori medi di vendita più elevati siano proprio gennaio, luglio, settembre e dicembre. Dalla figura 6, oltre a quelli appena citati, emerge anche il mese di aprile, assente in figura 5 per via dell'influenza dei mesi di aprile 2020 e 2021.

4.1.3 Scontrini

Una ulteriore analisi condotta è quella legata al prezzo medio per scontrino. Viste le pesanti perdite accusate durante il lockdown, è lecito aspettarsi che i ristoranti siano corsi ai ripari cercando di risanare il bilancio attraverso un aumento dei prezzi dei propri prodotti. Il grafico 9 mostra l'andamento del prezzo medio per scontrino, separando il periodo precedente e quello successivo all'inizio della pandemia e senza considerare il periodo di chiusura.

Come atteso, la discrepanza è netta: c'è una differenza di 7,62 € fra il prezzo medio per scontrino fra il post pandemia e il pre pandemia. Per altro, dopo la riapertura dei ristoranti si nota come l'andamento del prezzo medio sia decisamente più variabile, con i valori più elevati registrati nei periodi immediatamente successivi alle riaperture, a maggio e novembre 2020.

Una possibile spiegazione di questo aumento può essere legata all'introduzione di un servizio di consegna a domicilio che comporta l'introduzione di un minimo d'ordine e le eventuali spese di trasporto.

È stato effettuato un confronto fra l'andamento delle vendite e quello degli scontrini. L'andamento dei due attributi risulta essere analogo, come confermato dalla figura 8 dove possiamo notare una correlazione molto elevata, dal momento che le osservazioni si dispongono a formare una retta. Tuttavia, si può notare come ci sia però una evoluzione fra il periodo precedente e quello successivo allo scoppio della pandemia: in particolare, come già sottolineato, l'aumento del prezzo medio per scontrino fa sì che durante la pandemia, a parità di numero di scontrini, il totale dell'incasso sia maggiore.

Per quanto appena detto, le analisi esplorative condotte sulla variabile **scontrini** non vengono riportate, essendo analoghe a quelle mostrate per la variabile **lordototale**.

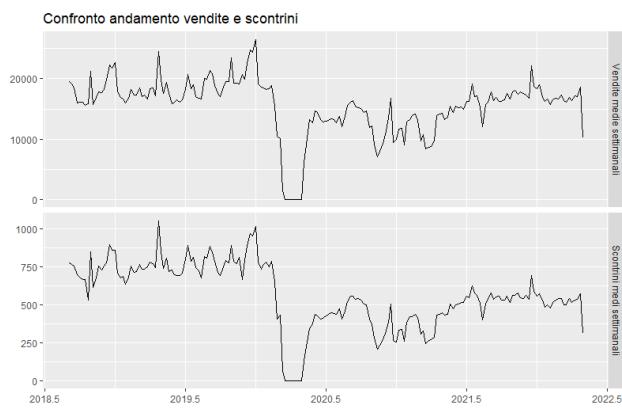


Figura 7. Andamento medio settimanale per vendite (sopra) e scontrini (sotto)

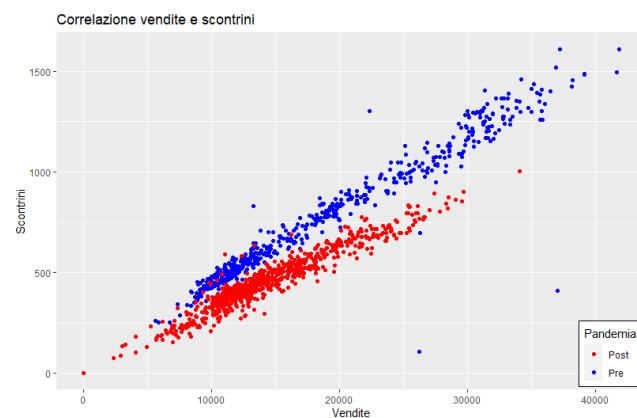


Figura 8. Correlazione fra vendite e scontrini

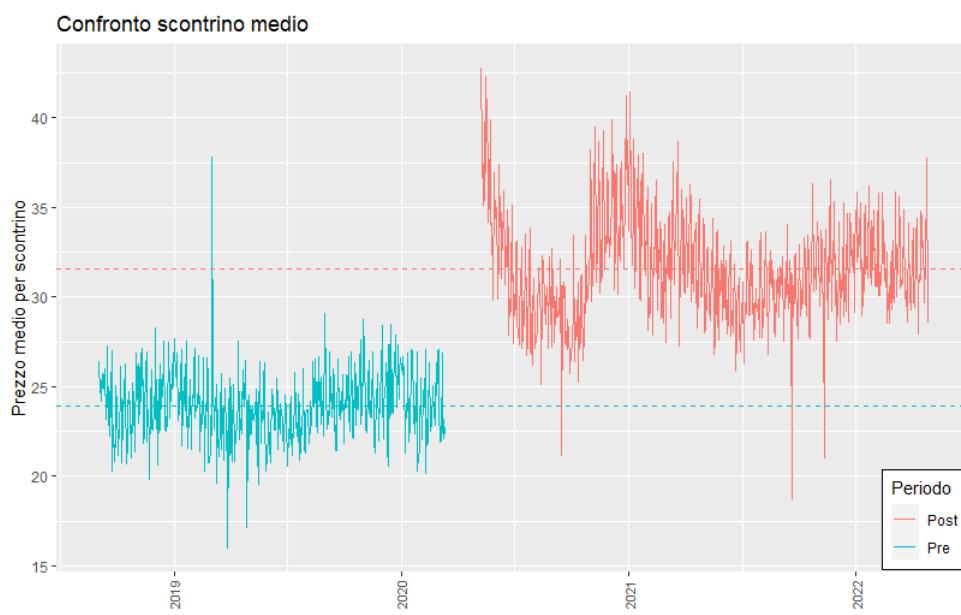
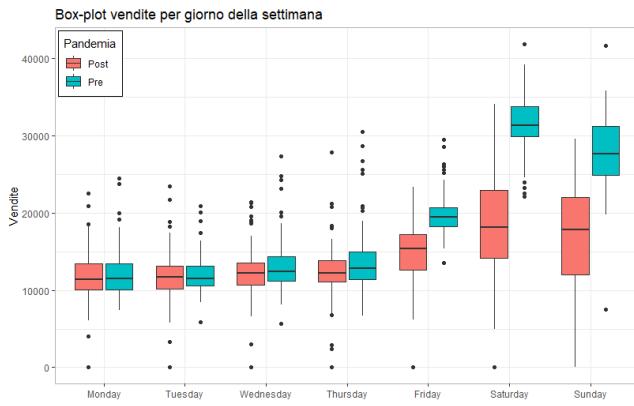
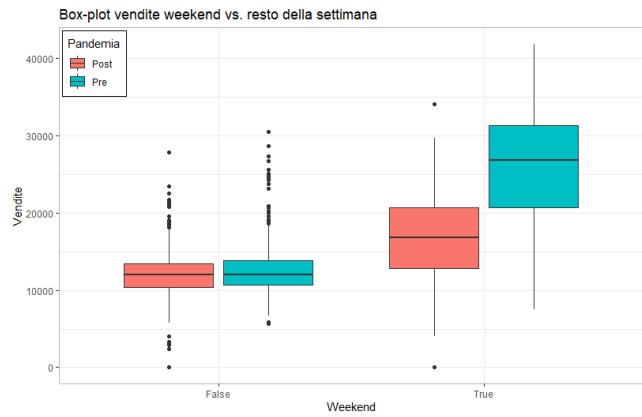


Figura 9. Prezzo medio per scontrino pre/post pandemia

4.1.4 Correlazione con altri attributi

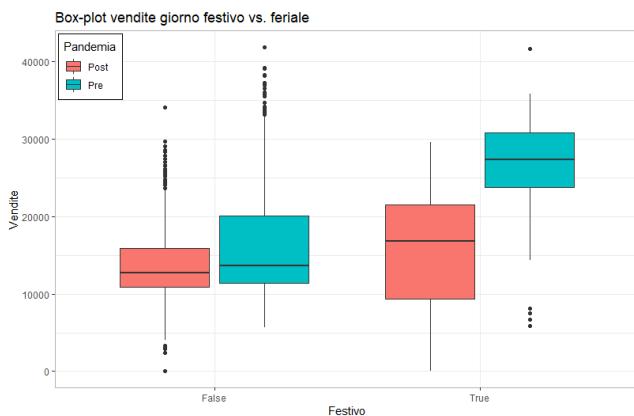
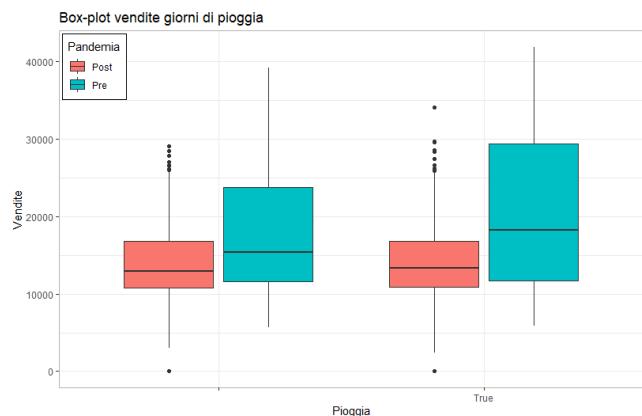
Nell'ultima parte dell'analisi esplorativa è stata indagata la dipendenza delle vendite da alcune degli attributi aggiunti nella fase di data integration. Sono stati realizzati dei box-plot condizionati da alcune delle variabili categoriali inserite, suddividendo i dati fra pre e post inizio della pandemia, così da evidenziare eventuali comportamenti discordanti. Di seguito vengono riportate alcune conclusioni interessanti.

**Figura 10.** Box-plot vendite per giorno della settimana**Figura 11.** Box-plot vendite weekend

Da figura 10 possiamo notare come i giorni che, in media, portano un incasso maggiore sono venerdì, sabato e domenica. Questa tendenza è valida sia per il periodo precedente che per quello successivo all'inizio della pandemia, seppur dopo di esso il valore medio sia significativamente inferiore. Questo andamento porta a incassi medi decisamente maggiori nel weekend rispetto al resto della settimana.

In figura 12, invece, si nota una differenza fra il valore medio di incasso fra giorni festivi e feriali. Nei giorni festivi, quali le domeniche e le varie festività, la gente è più portata a recarsi al ristorante. Tale comportamento è molto più netto nel periodo precedente alla pandemia, mentre nel periodo successivo la differenza in media, seppur significativa, è di molto inferiore.

L'ultimo box-plot, riportato in figura 13, distingue fra giorni con e senza precipitazioni. Tale attributo è spesso preso in considerazione per l'analisi delle vendite di un ristorante, dal momento che le persone sono meno inclini a uscire di casa in giorni di brutto tempo. In questo caso, tuttavia, le vendite medie registrate nei giorni con e senza presenza di precipitazioni non presentano differenze significative, tanto nel pre quanto nel post pandemia.

**Figura 12.** Box-plot vendite giorni festivi e feriali**Figura 13.** Box-plot giorni di pioggia

4.2 Altri ristoranti

Le analisi condotte sul ristorante 1 sono state replicate anche per i restanti 5 ristoranti. Le conclusioni tratte per il primo ristorante restano generalmente valide anche per gli altri, anche se con alcune differenze.

In figura 14 vengono mostrate le serie storiche delle vendite settimanali medie complete per i singoli ristoranti, mentre in figura 15 si mostra solo l'andamento del periodo successivo al primo lockdown, a partire dal 6 maggio 2020.

Per il ristorante 2 mancano i dati dal 27 settembre al 25 ottobre 2021, probabilmente a causa di un periodo di chiusura, mentre quelli del ristorante 3 partono dall'8 novembre 2019, cosa che suggerisce come tale ristorante sia stato aperto proprio in quel periodo. Tra gli altri, il ristorante 4 presenta molti più giorni di chiusura completa intorno alle festività principali, come Natale o Pasqua.



Figura 14. Vendite medie settimanali per ristorante



Figura 15. Vendite medie settimanali per ristorante, post-lockdown

L'andamento delle serie storiche dei singoli ristoranti risulta simile per tutti i ristoranti, con un periodo di chiusura completa nella primavera del 2020 a causa della pandemia e una successiva ripresa delle attività, prima lentamente e poi tornando sui livelli precedenti allo scoppio della pandemia (grafici 14 e 15).

In relazione a ciò, interessante è il comportamento dei ristoranti 4, 5 e 6. In particolare, analizzando i Seasonal Plot si può notare come tali ristoranti si siano ripresi molto meglio dalla pandemia rispetto al ristorante 1 e 2, facendo registrare valori medi di vendita pari o superiori a quelli degli anni precedenti alla pandemia. Di seguito si riporta il caso del ristorante 5 e del ristorante 6:

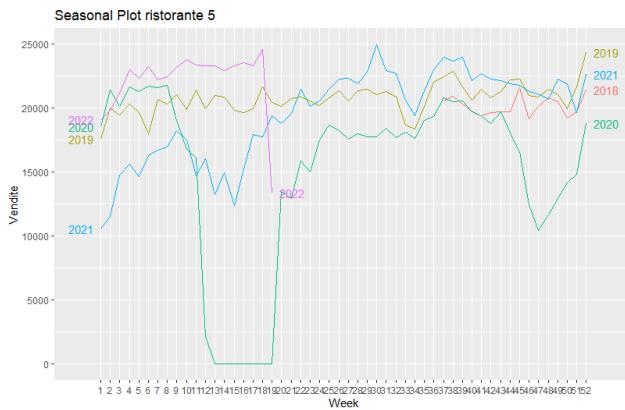


Figura 16. Vendite settimanali per anno, ristorante 5

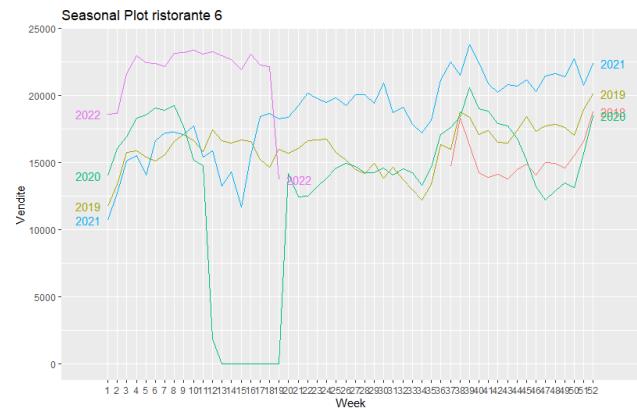


Figura 17. Vendite settimanali per anno, ristorante 6

Va sottolineato che quanto appena detto vale anche per il terzo ristorante. Tuttavia, avendo esso aperto solo alla fine del 2019, non è possibile affermare con certezza che il ristorante 3 abbia sofferto meno della pandemia. Per i ristoranti 4, 5 e 6, invece, il comportamento appena menzionato mostra come essi abbiano resistito meglio alle difficoltà portate dalla pandemia, andando addirittura a migliorare le proprie vendite. Nonostante la mancanza di informazioni circa la tipologia di ristorante, è possibile ipotizzare che si tratti di ristoranti che si sono meglio adattati al cambiamento delle abitudini dei consumatori, per esempio rendendo disponibile un servizio di *delivery*.

5. Analisi predittiva

In questa sezione viene riportata l'analisi effettuata per tentare di rispondere ai quesiti di ricerca riportati nella sezione 2. Tutte le analisi mostrate di seguito fanno quindi riferimento a previsioni della variabile *lordototale* e si riferiscono solo al *Ristorante 1*, i risultati possono poi essere intuitivamente estesi agli altri ristoranti.

5.1 Analisi della serie storica pre-Covid ai fini di stimare le possibili perdite subite durante il primo lockdown
La seconda domanda di ricerca vuole stimare gli incassi attesi per il *Ristorante 1* nella primavera del 2020, ai fini di determinare le perdite subite nel periodo del primo lockdown. Dapprima, viene condotta un'analisi esplorativa della serie storica pre-Covid

che permette di comprendere le sue componenti, quali trend e stagionalità.

Dopodiché, sono state utilizzate diverse tecniche di previsione, appartenenti alle famiglie dei modelli lineari (ARIMA, UCM, sezione 5.1.2 e 5.1.3) e dei modelli di Machine Learning (Random Forest, sezione 5.1.4). L'orizzonte predittivo considerato si estende dal 24 Febbraio 2020, giornata in cui sono state sancite le prime misure di contenimento della pandemia, al 06 Maggio 2020, primo giorno di riapertura, per un totale di 73 step di previsione.

Al fine di identificare il miglior modello tra quelli stimati sono state confrontate le performance di accuratezza di previsione mediante due processi di valutazione:

1. Procedura di *hold-out*, con suddivisione dei dati disponibili in *training* (466 osservazioni) e *test set* (73 osservazioni, in modo coerente con l'orizzonte di previsione), e determinazione di diverse metriche di accuratezza, per ogni modello considerato, quali RMSE, MAPE e MAE.
2. Procedura di *time series cross-validation* con *rolling forecasting origin*: un particolare tipo di cross-validation per previsioni di serie storiche, che tiene per conto della dipendenza temporale tra le osservazioni e che preserva tale dipendenza durante la fase di test. I modelli sono addestrati e valutati più volte, rispetto a quanto accade con la procedura di hold-out, ottenendo delle stime delle metriche di accuratezza di previsione puntuale più consistenti.

5.1.1 Analisi esplorativa serie storica pre-Covid

I dati a disposizione per il periodo precedente alla pandemia ricoprono il periodo Settembre 2018 - Febbraio 2020.

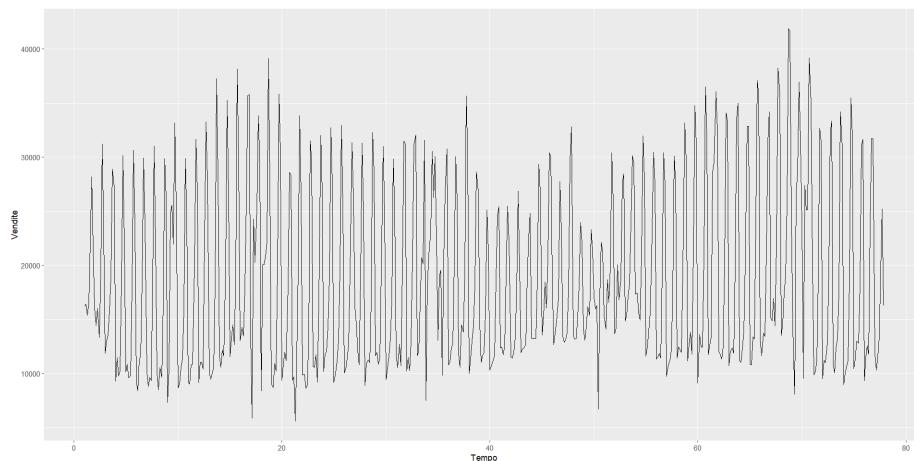


Figura 18. Incassi giornalieri pre-Covid per il primo ristorante

Come possiamo notare, l'andamento della serie storica risulta essere molto irregolare. Ciò è dovuto principalmente alla presenza di stagionalità nei dati.

I dati considerati sono giornalieri e, quindi, presentano stagionalità multiple. Nello specifico: una stagionalità annuale, che riflette l'andamento tipico delle attività ristorative che di norma incassano maggiormente in alcuni giorni dell'anno (i.e. Pasqua, Natale) e meno in altri periodi (i.e. Ferragosto), e una stagionalità settimanale, già analizzata nella sezione 4.1.2.

Passiamo ora ad analizzare le varie componenti della serie storica, ovvero trend, stagionalità e parte residua, tramite l'algoritmo *mstl*[1] (*multiple seasonal decomposition*, figura 19). Notiamo come la stagionalità a 365 giorni (annuale, ndr) non viene stimata per insufficienza di dati, dal momento che servono almeno due periodi completi, ovvero almeno due anni. Di conseguenza, la stima del trend risulta essere irregolare, poiché include la parte di stagionalità che l'algoritmo non è stato in grado di stimare. Infine, possiamo notare un trend in leggera crescita nella seconda parte della serie storica e un trend in netto calo nell'ultima parte, sintomo di come i dati siano influenzati dalla pandemia anche prima dell'imposizione delle prime restrizioni. Per quanto riguarda la stagionalità, si nota come nella parte centrale i picchi settimanali siano meno marcati rispetto al resto della serie storica.

5.1.2 Modelli ARIMA

I primi modelli usati per la previsione del fatturato sono i modelli ARIMA (e la loro estensione, SARIMA). I modelli ARIMA(p, d, q) fanno parte della famiglia dei *processi stocastici lineari non stazionari* e risultano essere una estensione dei modelli ARMA(p, q), a cui vengono applicate d differenziazioni per rendere stazionario il processo.

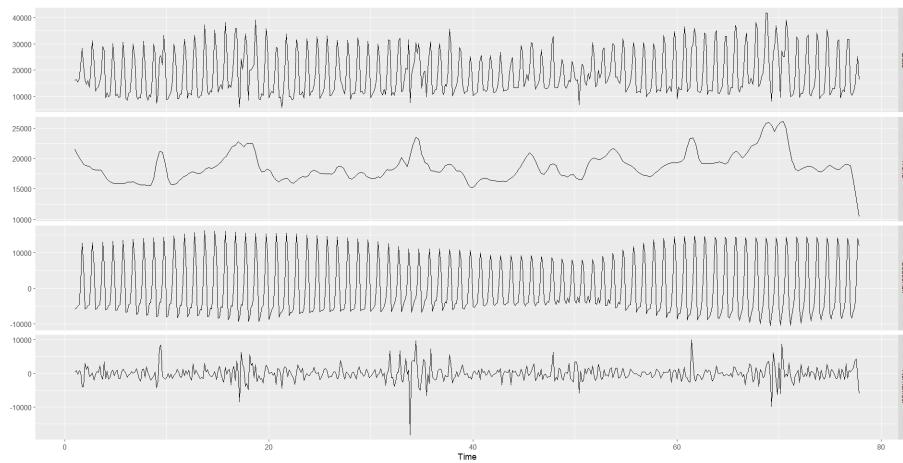


Figura 19. Decomposizione della serie storica pre-Covid

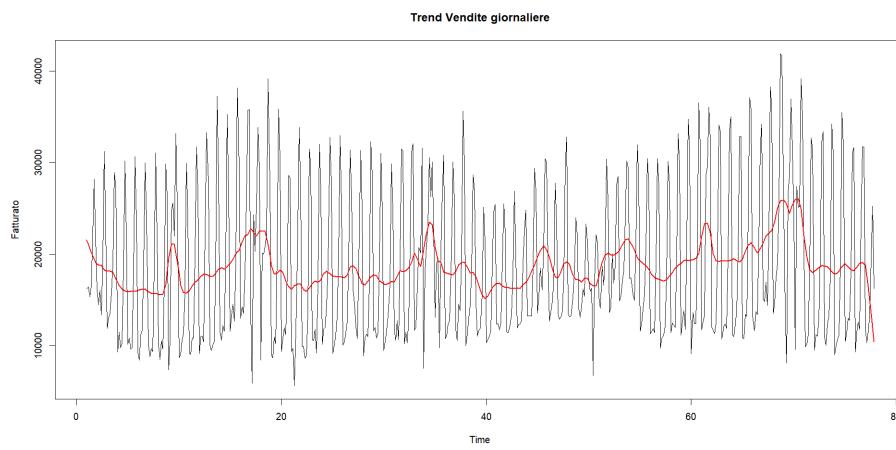


Figura 20. Stima trend della serie storica pre-Covid

Aspetti metodologici

La principale assunzione dei modelli ARMA è che il processo generatore dei dati sia un processo stazionario. Le condizioni di stazionarietà (debole) riguardano i primi due momenti del processo, ovvero funzione media e funzione varianza, le quali non devono dipendere dal tempo; inoltre, viene richiesto che la funzione di autocovarianza dipenda solo dal lag temporale tra le due osservazioni, e non dall'istante temporale specifico.

Ai fini di rendere stazionaria la serie storica in esame, sarà necessaria l'applicazione di alcune trasformazioni, quali differenziazioni (stagionali o non stagionali) per soddisfare il requisito di stazionarietà in media o trasformazioni di Box-Cox[2], come la trasformazione logaritmica, per soddisfare il requisito di stazionarietà in varianza.

In merito all'identificazione del corretto modello ARIMA (o SARIMA, ndr) a partire dall'osservazione dei dati, viene applicata la procedura di *Box-Jenkins*[3]. Tale procedura assume, in primo luogo, che la serie storica sia già stata resa stazionaria attraverso una serie di, eventuali, trasformazioni preliminari dei dati. I successivi passi della metodologia Box-Jenkins sono:

- Identificazione dei parametri $(p,d,q)(P,D,Q)_s$
- Stima dei parametri del modello
- Controllo diagnostico

Se il controllo diagnostico non è soddisfacente il modello viene specificato nuovamente in maniera iterativa fino al raggiungimento di risultati accettabili.

Per procedere con l'*identificazione dei parametri*, è necessario valutare se la serie storica necessita di trasformazioni, attraverso l'utilizzo di vari test, ai fini di renderla stazionaria. Il valore del parametro d o del parametro λ (trasformazione di Box-Cox)

derivano quindi da tali analisi.

Successivamente, attraverso l'analisi dell'ACF (*auto-correlation function*) e PACF (*partial auto-correlation function*) vengono scelti i parametri $(p, q)(P, Q)_s$ che si riferiscono rispettivamente alla parte auto-regressiva (riflette la dipendenza tra un'osservazione e p osservazioni ritardate) e alla parte a media mobile (che riflette la dipendenza tra un'osservazione e q residui ritardati), sia stagionali che non stagionali.

Successivamente si passa alla stima dei parametri scelti nella fase precedente. Fra i diversi metodi possibili, viene usato

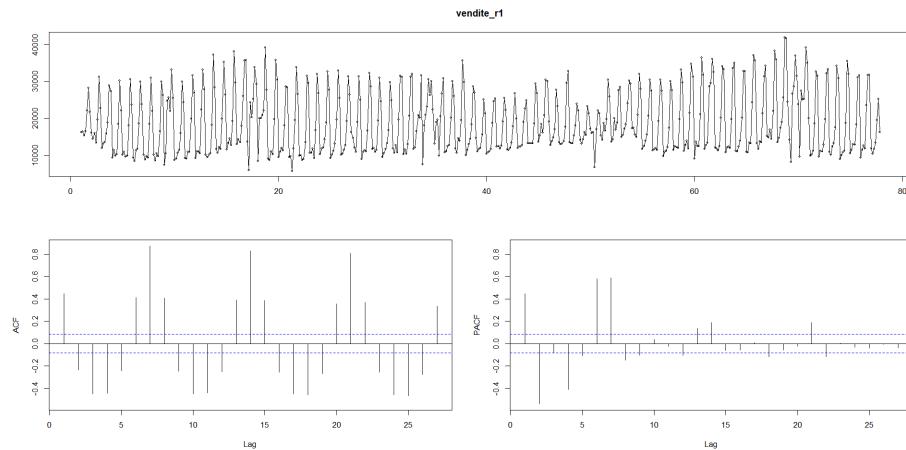


Figura 21. ACF e PACF della serie storica pre-Covid

il metodo dei *minimi quadrati condizionati* per trovare il punto di partenza, e successivamente il metodo della *massima verosimiglianza*.

Infine, una volta stimati i parametri, se ne verifica l'adeguatezza del modello in termini di:

- Significatività dei parametri
- Verifica delle condizioni di stazionarietà e invertibilità del modello
- Parsimonia del modello (criterio AIC, ndr)
- Analisi dei residui, in quanto si assume che siano *white noise*

Una ulteriore estensione dei modelli ARIMA sono i modelli ARIMAX, che introducono al loro interno dei regressori atti a modellare comportamenti che i parametri p, q, P, Q non riescono a catturare. L'introduzione di regressori porta un aumento della complessità del modello, comportando un aumento del parametro AIC, ma può migliorare le capacità predittive del modello. Nel caso in cui si decida di introdurli nel modello è quindi necessario valutarne la significatività statistica.

Stazionarietà della serie storica

Come si evince dai precedenti grafici, la serie storica è non stazionaria, per i seguenti motivi:

- Presenza di stagionalità (annuale e settimanale);
- Presenza di un trend (osservazioni sistematicamente sopra e sotto la media);
- Possibile non stazionarietà in varianza.

Attraverso i test *ndiffs* e *nsdiffs*, che valutano la necessità di una differenziazione (stagionale o non stagionale), e il test *BoxCox.lambda*, che stima il parametro λ ottimale, si evince che sono necessarie una differenza stagionale e l'applicazione di una trasformazione di Box-Cox con $\lambda = 0.59$.

Dalla figura 22 notiamo come l'ACF non decade più a zero lentamente, e attraverso alcuni test che valutano la stazionarietà della serie storica (i.e. *Ljung-Box test*, *Kwiatkowski-Phillips-Schmidt-Shin test*) possiamo affermare che la serie storica è ora stazionaria.

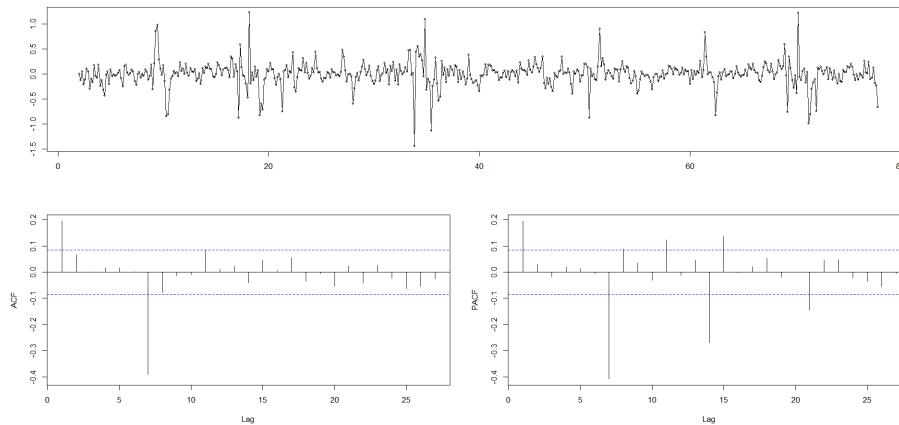


Figura 22. ACF e PACF della serie storica trasformata pre-Covid, differenziata stagionalmente

Stima e valutazione dei modelli

Per tale tipologia, sono stati considerati i seguenti modelli, a cui è stata applicata dapprima la procedura Box-Jenkins per l'individuazione del modello corretto e successivamente è stata usata la funzione `auto.arima()`, confrontandone i risultati.

1. **SARIMA(1,0,0)(0,1,2)**, $\lambda = 0.5969$. La modellazione di tutte le componenti viene lasciata ai soli parametri autoregressivi e a media mobile.
2. **ARIMAX(2,1,3)**, $\lambda = 0.5969$, regressori = dummy giornaliere. In questo modello si cerca di modellare la stagionalità settimanale attraverso l'introduzione di variabili dummy indicanti i giorni della settimana.
3. **SARIMAX(3,1,2)(1,0,1)**, $\lambda = 0.5969$, regressori = termini di Fourier. In questo modello si cerca di modellare sia la stagionalità settimanale che quella annuale attraverso l'introduzione dei termini di Fourier (3 per la stagionalità settimanale, 15 per quella annuale).
4. **SARIMAX(3,1,3)(1,0,0)**, $\lambda = 0.5969$, regressori = dummy festività annuali, settimanali + altri regressori. In questo modello la stagionalità annuale e settimanale sono gestite attraverso variabili dummy (per quella annuale sono state inserite dummy per ogni festività, aggiungendo anche una variabile che tenga conto di eventuali ponti); inoltre, sono inseriti altri regressori, come una dummy che identifica i giorni di pioggia, per cercare di modellare i comportamenti di natura non stagionale, cercando di aumentare l'accuratezza delle previsioni.

Di seguito vengono riportate le varie misure di performance su dati di *training* e *test* per ogni modello (tabella 1 e 2):

Modello	RMSE	MAE	MAPE	AIC
SARIMA(1,0,0)(0,1,2)	3003.55	1979.22	12.34	4096.91
ARIMAX(2,1,3)	2977.31	2035.70	12.54	4131.82
SARIMAX(3,1,2)(1,0,1)	2646.37	1776.44	10.97	4077.01
SARIMAX(3,1,3)(1,0,0)	2139.70	1603.14	9.53	3846.98

Tabella 1. Metriche di performance dei modelli sui dati di **training**

Modello	RMSE	MAE	MAPE
SARIMA(1,0,0)(0,1,2)	4989.80	3602.51	19.10
ARIMAX(2,1,3)	5004.92	3602.51	24.51
SARIMAX(3,1,2)(1,0,1)	3986.36	3032.06	17.86
SARIMAX(3,1,3)(1,0,0)	4869.57	3765.96	19.94

Tabella 2. Metriche di performance dei modelli sui dati di **test**

Seguendo la procedura di *hold-out*, il modello che sui dati di *training* risulta avere le performance migliori è il modello **SARIMAX(3,1,3)(1,0,0)**, con MAPE pari a 9.53 % e AIC pari a 3846.98. Tali performance non vengono però confermate sui

dati di test, sintomo che il modello soffre di *overfitting*. Il modello che generalizza meglio sui dati di test è quello che sfrutta i termini di Fourier come regressori, con MAPE pari a 17.86 % e RMSE pari a 3986.36.

5.1.3 Modelli UCM

Una seconda serie di modelli utilizzati per la decomposizione e previsione delle serie storiche è quella dei modelli a componenti non osservabili. A differenza dei modelli ARIMA i modelli UCM non necessitano dell'assunzione di stazionarietà della serie storica e permettono di decomporre la serie storica in trend, stagionalità e cicli. Per applicare questi modelli è necessario innanzitutto assegnare il valore NA ai dati che si vogliono predire per permettere al filtro di Kalman di predire i valori sulla base delle componenti non osservate. In seguito è necessario valutare i risultati dell'analisi esplorativa della serie storica per valutare le componenti da inserire, in particolare da questa è emerso:

- Trend variabile
- Forte stagionalità settimanale
- Possibile stagionalità annuale

Sono quindi stati realizzati tre modelli:

1. **UCM con ritardi** è costruito sfruttando unicamente le informazioni contenute nella serie storica degli incassi

$$Y_t = LLT + SEAS_{7gg} + SEAS_{365gg}$$

LLT corrisponde al Local Linear Trend ed assume che sia la media che la pendenza del trend siano influenzate da rumore, la stagionalità a 7 giorni è descritta con dummy stocastiche mentre la stagionalità a 365 giorni è descritta con funzioni trigonometriche. Per descrivere la stagionalità annuale sono state effettuate diverse prove considerando diversi numeri di armoniche fondamentali, i risultati migliori si ottengono considerandone 10 di queste.

2. **UCM con ritardi e regressori** sviluppato aggiungendo il regressore indicante le festività, un attributo binario che distingue i giorni festivi e feriali. Questa aggiunta è motivata dal fatto che il training set presenta poco più di un anno di osservazioni e la componente della stagionalità a 365 giorni fatica a distinguere gli eventi ricorrenti dal rumore. Il modello risulta quindi composto in questo modo:

$$Y_t = LLT + SEAS_{7gg} + Fest$$

In questo caso la stagionalità a 7 giorni è stata ottenuta con le componenti trigonometriche poiché permette una miglior interpretazione dell'andamento settimanale.

3. **UCM con ritardi e dummy** realizzato sull'idea che le varie festività possano incidere in modo diverso sugli incassi, vengono dunque passate al modello come regressori individuali. Questo comporta un aumento dei parametri ma un teorico miglioramento delle prestazioni nonostante per diverse festività ci sia solo un'occorrenza nel train set.

Di seguito, vengono riportati i valori di performance dei differenti modelli:

Modello	RMSE	MAE	MAPE
UCM ritardi	4175.09	3060.12	17.58
UCM ritardi + regressori	5072.25	4065.36	24.10
UCM ritardi + dummy	4593.67	3527.28	18.81

Tabella 3. Metriche di performance dei modelli UCM

5.1.4 Modelli di Machine Learning

Come terza e ultima classe di modelli sono stati considerati i modelli di Machine Learning. Nel caso di questi modelli, l'idea è quella di riuscire ad apprendere l'andamento della serie storica direttamente dai dati, senza doverli necessariamente manipolare o definire componenti specifiche.

Nel caso specifico, è stato scelto il modello Random Forest come rappresentante di questa classe di algoritmi. Si tratta di uno dei modelli più semplici ma in grado di raggiungere una buona accuratezza in molti casi reali.

L'informazione temporale contenuta nei dati viene implementata nell'algoritmo passando come regressori un certo numero di ritardi. Complessivamente, vengono realizzati 3 modelli:

1. **Random Forest con ritardi**, che considera come regressori del modello i soli ritardi della serie storica;
2. **Random Forest ritardi + regressori**, dove vengono aggiunti alcuni regressori utili: un regressore a 7 valori che indica il giorno della settimana, un regressore a 366 valori che indica il giorno dell'anno, un regressore binario che distingue fra giorni feriali e festivi e un regressore binario che identifica i giorni in cui si sono verificate precipitazioni;
3. **Random Forest ritardi + dummy**, che, sulla base dell'idea seguita in sezione 5.1.3, viene sviluppato identificando ogni singola festività con un regressore indipendente.

In Tabella 4, vengono riportati i valori di performance dei differenti modelli: le metriche ottenute individuano nel secondo modello quello più performante.

Modello	RMSE	MAE	MAPE
RF ritardi	4910.38	3311.81	20.69
RF ritardi + regressori	4379.09	2910.40	14.59
RF ritardi + dummy	4392.08	2936.24	14.68

Tabella 4. Metriche di performance dei modelli Random Forest

5.1.5 Time Series Cross-Validation: valutazione accuratezza puntuale di previsione e confronto tra modelli

Time Series Cross-Validation: definizione della procedura di validazione [4][5][6][7][8]

Le performance di previsione puntuale possono essere determinate solo considerando come i diversi modelli si comportino su nuovi dati che non siano stati utilizzati in fase di training, in modo da ottenere delle previsioni confrontabili con i dati reali e quindi poter quantificare l'errore di previsione commesso.

L'approccio più semplice al processo di valutazione dei modelli, eseguito nelle sezioni precedenti, è quello di *hold-out*, il quale prevede la separazione dei dati disponibili in *training data*, utilizzati per la stima dei parametri di un modello di previsione, e in *test data* (temporalmente consequenti a quelli di training e, in numero, coerenti con l'orizzonte di previsione scelto), utilizzati per la valutazione dell'accuratezza di previsione.

Se tale procedura viene eseguita una sola volta, si parla di valutazione con *fixed origin*. Per rendere più consistente la valutazione dei modelli, si è invece scelto di adottare un approccio alla validazione alternativo, impostando una procedura di *cross validation* con *rolling forecasting origin*, detta anche *time series cross-validation*.

Con time series cross-validation si intende un particolare tipo di cross-validation per previsioni di serie storiche, che tiene però conto della dipendenza temporale tra le osservazioni e che preserva tale dipendenza durante la fase di test. Questa tipo di validazione ci ha permesso di confrontare i diversi modelli in termini di performance di previsione, potendo ottenere delle misure di accuratezza, con il vantaggio poi di poter allenare il modello scelto sull'intero dataset.

Specificatamente si è costruita una procedura di time series cross-validation con *rolling forecasting origin, increasing in-sample e non-constant holdout*, adattata per previsioni *multi-step*.

Data la serie storica $y = \{y_1, \dots, y_T\}$, la procedura prevede l'iniziale definizione di due parametri: l'orizzonte di previsione H , ovvero il numero di *step* di previsione in avanti che si vogliono generare (e quindi testare) con i modelli e la finestra minima di training m , la quale dimensione deve essere maggiore o uguale al numero minimo di osservazioni necessarie per stimare il modello e che coincide con la dimensione iniziale del *training set*. Durante la prima iterazione del processo il modello viene stimato sui dati y_1, \dots, y_t (e eventuali regressori), dove $t=m$, e vengono predette le osservazioni successive $\hat{y}_{t+h|t}$ per $h = 1, \dots, H$. Date le osservazioni del test set y_{t+1}, \dots, y_{t+h} , per $h = 1, \dots, H$, queste sono confrontate con le h -esime previsioni ottenendo gli errori di previsione assoluti $e_{t+h} = y_{t+h} - \hat{y}_{t+h|t}$ e percentuali $e\%_{t+h} = 100 * e_{t+h} / y_{t+h}$, per $h = 1, \dots, H$. Tale processo viene ripetuto per $t = m, \dots, T-1$, l'origine del training set si sposta di una osservazione ogni step, ingrandendo iterativamente la dimensione del training set (*increasing in-sample*). Il modello viene stimato a ogni step e testato sul nuovo test set, il quale iterativamente perde l'osservazione iniziale e acquisisce un'osservazione alla fine della serie. Si ottengono due matrici degli errori (assoluti e percentuali), con un numero di righe pari al numero $T-m-1$ di iterazioni del processo e H colonne, dove la h -esima colonna contiene gli errori per l'orizzonte di previsione h . Inoltre, nel caso specifico si è deciso di adottare una strategia del tipo *non-constant holdout sample size*, utile nel caso di serie storiche di piccole dimensioni, la quale prevede che la procedura di rolling origin continui anche quando il test set abbia un numero di osservazioni minore dell'orizzonte di previsione H , riducendo progressivamente l'orizzonte di previsione quando l'origine del training set si avvicini alla fine della serie storica (come mostrato in Figura 23).

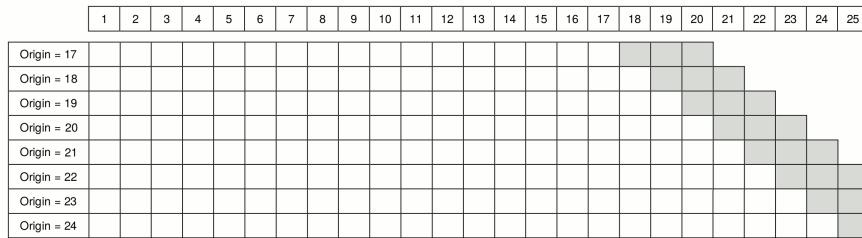


Figura 23. Esempio di procedura di *Time Series Cross-Validation* su una serie storica di 25 osservazioni, con *rolling forecasting origin* ($m = 17$), *increasing in-sample, non-constant holdout* e adattata per previsioni *multi-step* ($H = 3$). Da [9].

Ottenute le matrici degli errori di previsione, gli errori e_{m+1}, \dots, e_T , per i $h = 1, \dots, H$ orizzonti temporali sono riassunti attraverso diverse metriche per la valutazione dell'accuratezza di previsione puntuale. Dalla matrice degli errori assoluti (errori *scale dependent*) si ottiene: *Root Mean Squared Error RMSE* (1), misura minimizzata dalla media degli errori, e *Mean Absolute Error MAE* (2), metrica invece minimizzata dalla mediana degli errori, quindi più robusta agli *outlier*. Mentre dagli errori percentuali si computa il *Mean Absolute Percentage Error MAPE* (3), indipendente dalla scala dei dati e di facile interpretazione.

$$(1) \text{RMSE}_h = \sqrt{\frac{1}{T-m-1} \sum_{t=m}^{T-1} (y_{t+h} - \hat{y}_{t+h|t})^2} \quad (2) \text{MAE}_h = \frac{1}{T-m-1} \sum_{t=m}^{T-1} |y_{t+h} - \hat{y}_{t+h|t}| \quad (3) \text{MAPE}_h = \frac{1}{T-m-1} \sum_{t=m}^{T-1} \left| \frac{y_{t+h} - \hat{y}_{t+h|t}}{y_{t+h}} \right|$$

La procedura di time series cross-validation è stata implementata in *R*, utilizzando, per la famiglia di modelli ARIMA, una versione leggermente modificata della funzione `tscv()` [10] del pacchetto `forecast` e definendo invece delle funzioni inedite per le famiglie di modelli UCM e Random Forest.

Time Series Cross-Validation: definizione dei parametri

In Tabella 5 sono riportati i valori dei parametri di time series cross validation, H (orizzonte o numero di step di previsione) e m (dimensione della finestra iniziale di training), selezionati per i diversi modelli stimati.

Modello	Orizzonte/step di previsione (H)	Osservazioni nella finestra iniziale di training (m)
SARIMA(1,0,0)(0,1,2)	73	126
ARIMAX(2,1,3)	73	126
SARIMAX(3,1,2)(1,0,1)	73	365
SARIMAX(3,1,3)(1,0,0)	73	365
RF ritardi	73	120
RF ritardi + regressori	73	120
RF ritardi + dummy	73	365
UCM ritardi	73	365
UCM ritardi + regressori	73	120
UCM ritardi + dummy	73	120

Tabella 5. Parametri di time series cross validation per la validazione dei modelli stimati sulla serie storica pre-Covid.

H è uguale a 73 per tutti i modelli, valore coerente con il numero di step di previsione da effettuare a partire dalla fine della serie storica pre-Covid e che identifica il periodo di previsione che inizia il giorno 24 Febbraio 2020 e termina il 6 Maggio 2020. Il parametro m , il numero di osservazioni nella finestra iniziale di training, ovvero la porzione di serie storica su cui non si esegue la cross-validation, ha invece subito una fase di ottimizzazione che ha portato all'ottenimento di valori differenti per i diversi modelli. Si è definito inizialmente il limite inferiore del parametro, coincidente con il valore dell'orizzonte di previsione H . Inoltre, i diversi modelli richiedono un numero minimo di osservazioni affinché possano essere stimati, valore che incrementa con complessità degli stessi, con la presenza di regressori che si realizzano annualmente (come le variabili dummy relative alle festività) e con il fatto che sia modellata una stagionalità annuale. Per buona parte dei modelli più complessi si è quindi deciso di impostare un valore minimo di m uguale a 365. Definito il valore minimo per il parametro si sono poi testati iterativamente diversi valori dello stesso: parte dei modelli ottiene performance di accuratezza migliori con una finestra iniziale di previsione ampia, mentre altri modelli raggiungono accuratezze migliori con valori di m minori, probabilmente a causa della presenza di *outlier* negli errori, il quale impatto sulle metriche di accuratezza viene mitigato effettuando un maggior numero di iterazioni di cross-validation, quindi ottenendo un numero di errori, per orizzonte di previsione, maggiore.

Time Series Cross-Validation: risultati e selezione del modello più performante

Tabella 6. Metriche per la valutazione dell’accuratezza di previsione puntuale, dei modelli stimati sulla serie storica pre-Covid, in funzione dell’orizzonte di previsione h (campionati ogni 7-step) e medi ($h_i - h_f \text{ avg}$): RMSE (a), MAE (b) e MAPE (c). I valori relativi alle performance migliori sono evidenziati.

(a) Root Mean Squared Error (RMSE) in funzione dell’orizzonte di previsione h .

h	SARIMA (1,0,0)(0,1,2)	ARIMAX (2,1,3)	SARIMAX (3,1,2)(1,0,1)	SARIMAX (3,1,3)(1,0,0)	RFr	RFr+r	RFr+d	UCMr	UCMr+r	UCMr+d
1	3313	3398	3676	2745	3330	3140	3218	4017	3848	3662
7	3636	3643	3348	3328	3683	3423	3477	3907	3821	3776
14	3877	4046	3531	4196	3946	3575	3598	4455	4507	4521
21	3973	4301	3640	4771	4277	3784	3798	4524	4749	4977
28	4043	4450	3766	5135	4494	3950	3961	4560	4978	6422
35	4119	4567	3891	5481	4557	4018	4031	4896	5204	6690
42	4195	4507	3978	5658	4617	3997	4045	5090	5144	7779
49	4281	4490	3997	5993	4628	4028	4036	5350	5310	10000
63	4451	4637	3927	5141	4821	4320	4352	4258	5550	16184
73	4572	5040	4202	4859	4897	4487	4513	5988	6366	14097
1-7 avg	3582	3628	3600	3158	3602	3340	3399	4393	4464	4666
7-30 avg	3977	4285	3639	4609	4256	3762	3774	5062	5115	5698
30-73 avg	4338	4671	4042	5495	4699	4159	4168	5641	5938	10615
1-73 avg	4152	4449	3873	4992	4454	3955	3970	5339	5537	8495

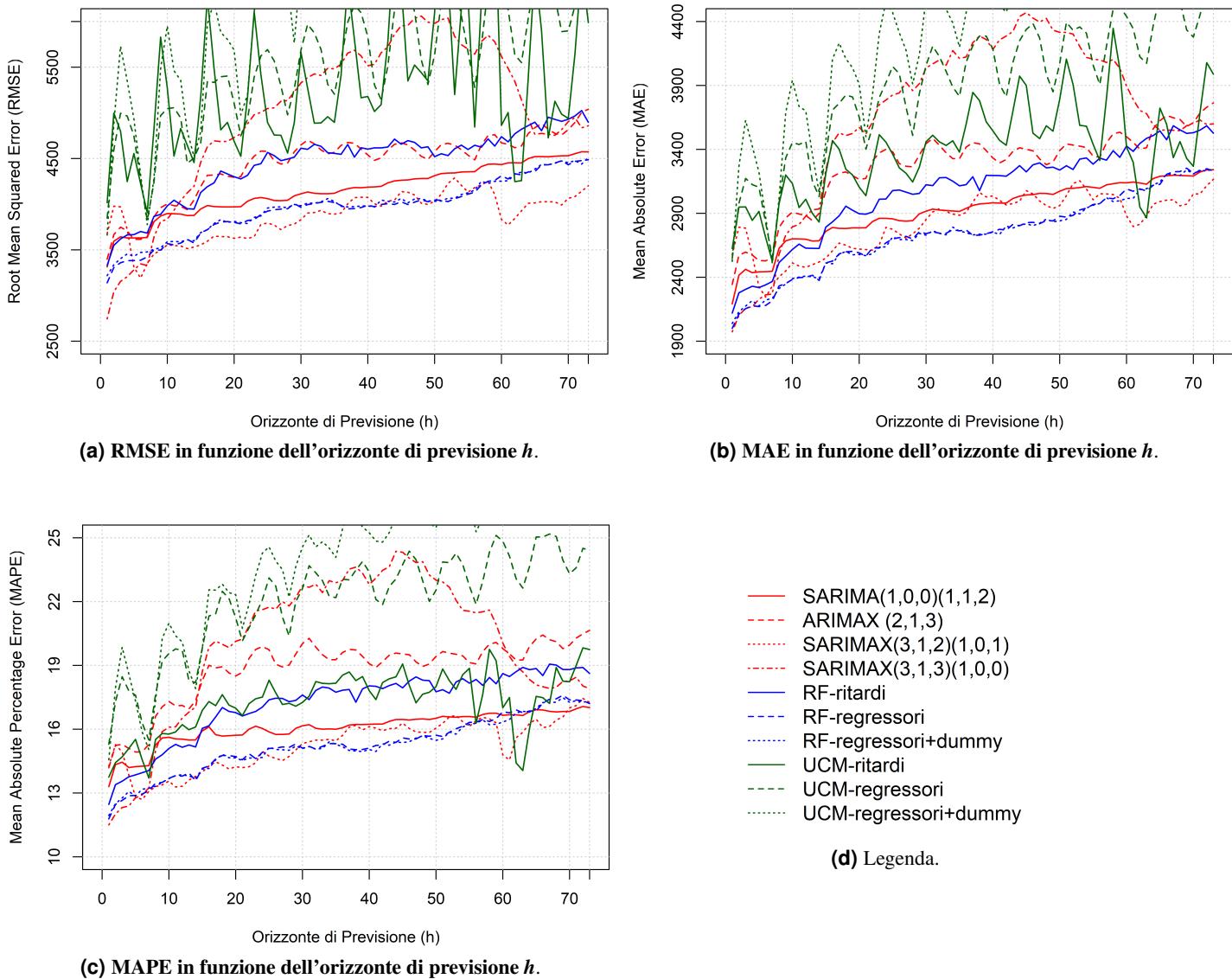
(b) Mean Absolute Error (MAE) in funzione dell’orizzonte di previsione h .

h	SARIMA (1,0,0)(0,1,2)	ARIMAX (2,1,3)	SARIMAX (3,1,2)(1,0,1)	SARIMAX (3,1,3)(1,0,0)	RFr	RFr+r	RFr+d	UCMr	UCMr+r	UCMr+d
1	2191	2344	2577	1973	2120	2000	2039	2625	2547	2525
7	2447	2554	2332	2269	2367	2216	2235	2524	2552	2507
14	2685	2899	2519	2934	2625	2375	2388	2832	3031	3063
21	2787	3173	2617	3520	2893	2565	2569	3038	3402	3500
28	2842	3264	2739	3860	3049	2699	2687	3145	3510	3912
35	2915	3327	2885	4111	3115	2763	2761	3379	3814	4200
42	2980	3321	2965	4287	3190	2769	2800	3432	3812	4474
49	3048	3302	2937	4332	3261	2840	2834	3542	3850	4689
63	3128	3413	2880	3679	3486	3075	3092	2864	4073	5173
73	3241	3763	3166	3599	3527	3231	3251	3988	4527	5706
1-7 avg	2405	2528	2518	2166	2294	2144	2170	2788	2901	3063
7-30 avg	2782	3148	2623	3415	2861	2560	2552	3200	3602	3919
30-73 avg	3078	3457	2994	4035	3324	2948	2940	3592	4243	5013
1-73 avg	2920	3271	2831	3660	3079	2749	2744	3391	3912	4481

(c) Mean Absolute Percentage Error (MAPE) in funzione dell’orizzonte di previsione h .

h	SARIMA (1,0,0)(0,1,2)	ARIMAX (2,1,3)	SARIMAX (3,1,2)(1,0,1)	SARIMAX (3,1,3)(1,0,0)	RFr	RFr+r	RFr+d	UCMr	UCMr+r	UCMr+d
1	13.3	14.2	14.3	11.5	12.5	11.8	11.9	13.7	15.3	14.6
7	14.3	15.3	13.1	13.2	14.1	13.1	13.2	13.7	15.5	15.0
14	15.5	17.3	13.8	17.0	15.2	13.6	13.7	16.1	18.2	17.9
21	15.7	18.7	14.2	20.2	16.6	14.6	14.5	16.6	20.1	20.7
28	15.8	19.0	14.7	21.8	17.3	15.1	15.0	17.2	20.4	22.2
35	16.0	19.3	15.5	22.9	17.7	15.3	15.3	18.2	22.2	24.1
42	16.2	19.2	15.9	23.5	17.9	15.3	15.3	18.2	21.9	24.8
49	16.5	19.1	15.8	23.3	18.3	15.7	15.6	17.7	22.0	25.7
63	16.6	19.3	15.9	18.7	18.9	16.8	16.8	14.1	22.6	27.1
73	17.0	20.6	17.3	17.9	18.6	17.2	17.3	19.7	24.4	31.4
1-7 avg	14.2	15.0	13.8	12.5	13.6	12.7	12.8	14.5	16.9	17.0
7-30 avg	15.8	18.5	14.2	19.5	16.4	14.5	14.4	16.9	20.8	21.8
30-73 avg	16.5	19.6	16.1	21.4	18.3	16.1	16.0	18.0	23.5	27.3
1-73 avg	16.1	18.8	15.3	20.0	17.2	15.3	15.2	17.3	22.0	24.6

Figura 24. Metriche per la valutazione dell’accuratezza di previsione puntuale, RMSE (a), MAE (b) e MAPE (c), dei modelli stimati sulla serie storica pre-Covid, in funzione dell’orizzonte di previsione h e legenda (d). L’asse y è troncato, al fine di mostrare meglio i modelli più performanti.



I risultati del processo di time series cross validation sui modelli stimati sulla serie storica pre-Covid sono riportati in Figura 24 e in Tabella 6. In particolare, in Tabella 6 si mostrano: le diverse metriche per la valutazione dell’accuratezza di previsione puntuale in funzione dell’orizzonte di previsione h (non sono presenti le accuratezze per tutti i 73 gli step di previsione, ma sono stati campionati i valori ogni 7 step); i valori delle accuratezze medie $h_i - h_f \text{ avg}$ (per gli intervalli di step 1-7, 7-30 e 30-73), utili a verificare le performance di previsione a breve, medio e lungo termine e i valori di accuratezza media lungo tutto l’orizzonte di previsione ($h_1 - h_{73} \text{ avg}$). Analogamente in Figura 24 sono rappresentate le diverse metriche per la valutazione dell’accuratezza di previsione puntuale in funzione dell’orizzonte di previsione h (per tutti i 73 step).

Dai risultati si evince:

- Come atteso le performance di accuratezza di previsione puntuale tendono a peggiorare all’aumentare dell’orizzonte di previsione h .
- Le metriche di performance medie (1-73 avg), ottenute mediante time series cross validation, sono paragonabili o migliori rispetto a quelle ottenute mediante la procedura di hold-out per i modelli delle famiglie ARIMA e Random Forest. I modelli UCM invece esibiscono dei valori, soprattutto per quanto riguarda la metrica RMSE, sensibilmente peggiori. Da

una analisi degli errori è stato possibile evidenziare la presenza di outliers, negli stessi, in determinate iterazioni del processo, legati a forti sovrastime del valore del *lordo totale*, i quali portano a un peggioramento del valore di RMSE, particolarmente sensibile ai valori estremi. Tali outlier potrebbero essere imputabili alla sensibilità di questo tipo di modelli alla posizione dell'ultima osservazione del training set.

- L'utilizzo delle variabili *dummy* per indicare le singole festività, rispetto a un'unica variabile indicante se il giorno è o meno festivo, in tutti i modelli e per tutte le metriche, non migliora le performance di previsione dei modelli e per buona parte di essi si ottiene forte peggioramento delle stesse. La matrice dei regressori dummy è sparsa (le festività si realizzano annualmente), è quindi probabile che la quantità di dati di training disponibili non sia sufficiente per permettere ai modelli di apprendere adeguatamente i dati.
- I modelli Random Forest (ritardi + regressori), Random Forest (ritardi + dummy) e SARIMAX(3,1,2)(1,0,1) sovraprofornano gli altri modelli in termini di metriche medie (1-73 avg) e lungo i diversi orizzonti di previsione. Il modello SARIMAX(3,1,2)(1,0,1), rispetto ai due modelli Random Forest, performa peggio per previsioni a breve termine, mentre presenta accuratezze migliori per previsioni a lungo termine. Si è scelto quindi di utilizzare per la fase di previsione i modelli SARIMAX(3,1,2)(1,0,1) e Random Forest (ritardi + regressori). Infatti, le performance dei due modelli Random Forest sono molto simili, si è quindi optato per il modello più parsimonioso.

5.1.6 Previsioni e considerazioni

I modelli selezionati vengono utilizzati per prevedere gli incassi del ristorante nel periodo fra il 24 febbraio 2020 e il 6 maggio 2020. Di seguito i risultati:

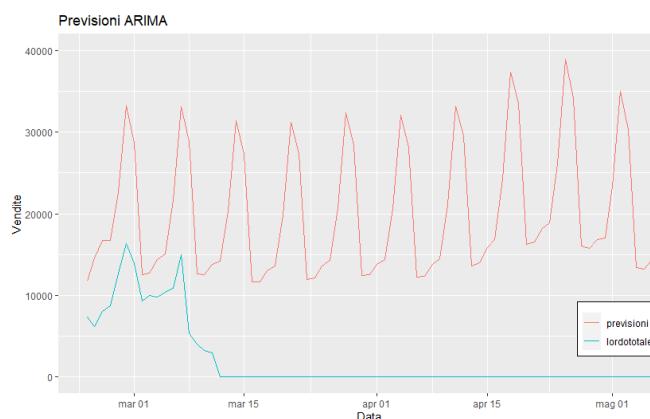


Figura 25. Previsioni modello ARIMA

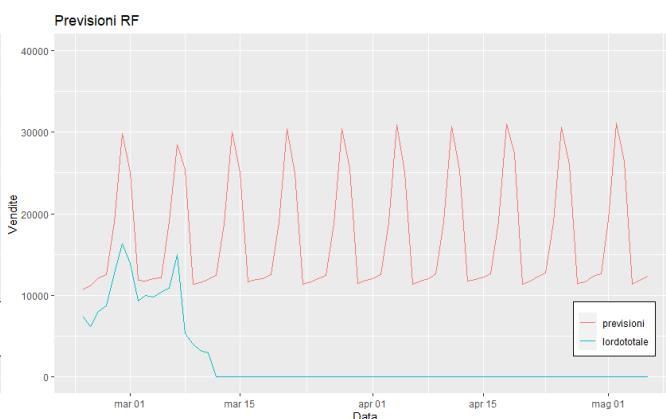


Figura 26. Previsioni modello Random Forest

Come si evince dai grafici, anche nel periodo fra la fine di febbraio e l'inizio di marzo, antecedente al lockdown completo, il ristorante accusa un calo nelle vendite.

La perdita totale stimata dal modello Random Forest ammonta a 1.108.234€, pari al 16% del fatturato totale dell'anno precedente. Nel caso del modello ARIMA, invece, la perdita stimata ammonta a 1.441.490€, pari al 21% del fatturato totale del 2019. La discrepanza fra i due modelli sta nel fatto che, come si nota dalle figure 25 e 26, le previsioni del modello ARIMA presentano una crescita nelle ultime 3 settimane considerate, con picchi stimati che sfiorano incassi di 40.000€.

5.2 Analisi della serie storica post primo lockdown ai fini di prevedere il livello di fatturato futuro

La terza domanda di ricerca vuole stimare gli incassi futuri per il Ristorante 1. Dapprima, viene condotta un'analisi esplorativa della serie storica post 1° lockdown che permette di comprendere le sue componenti, quali trend e stagionalità. Dopodiché, sono state utilizzate diverse tecniche di previsione, appartenenti alle famiglie dei modelli lineari (ARIMA, UCM, sezione 5.2.2 e 5.2.3) e dei modelli di Machine Learning (Random Forest, sezione 5.2.4). L'orizzonte predittivo considerato si estende dal 29 Aprile 2022 al 27 Giugno 2022, per un totale di 60 step di previsione.

Come per la precedente analisi predittiva (sezione 5.1), anche in questo caso prima si procede ad identificare un insieme di possibili modelli tramite procedura *hold out*, per poi scegliere il modello migliore tramite procedura di *time series cross validation* con *rolling forecasting origin*.

5.2.1 Analisi esplorativa della serie storica post 1° lockdown

I dati a disposizione per il periodo successivo al primo lockdown ricoprono il periodo Maggio 2020 - Aprile 2022.

Come possiamo notare dalla decomposizione della serie storica (figura 27) l'andamento risulta essere molto irregolare, a causa soprattutto:

- Presenza di stagionalità, annuale e settimanale
- Influenza delle restrizioni per il contenimento dei contagi, che entravano in vigore in base al colore della regione in un determinato momento

L'influenza delle restrizioni si nota maggiormente andando a visualizzare il trend della serie storica indicando i vari cambi di colore (figura 28). Notiamo come il trend cambi di direzione, con diverse pendenze, ogni qualvolta la regione cambia di colore; inoltre per i periodi di maggiori contagi (e maggiori restrizioni) la stagionalità settimanale risulta essere molto meno marcata (figura 27). Si noti come l'Irregolarità del trend è dovuta anche alla presenza di stagionalità annuale al suo interno, non stimabile dall'algoritmo *mtsl* per insufficienza di dati. Infine, si evidenzia come l'allentamento quasi totale delle restrizioni (periodo dove la regione non ha mai cambiato di colore) abbia portato ad un aumento generale del trend delle vendite, con conseguente aumento dell'intensità della stagionalità settimanale.

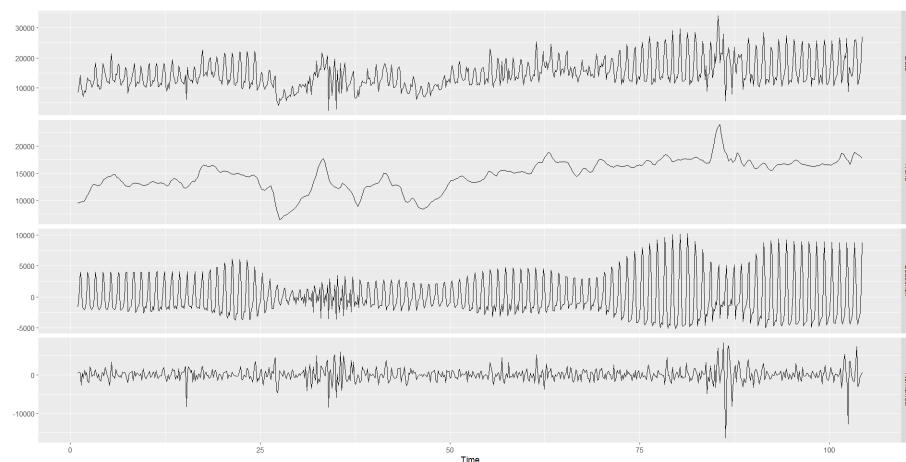


Figura 27. Decomposizione della serie storica post 1° lockdown

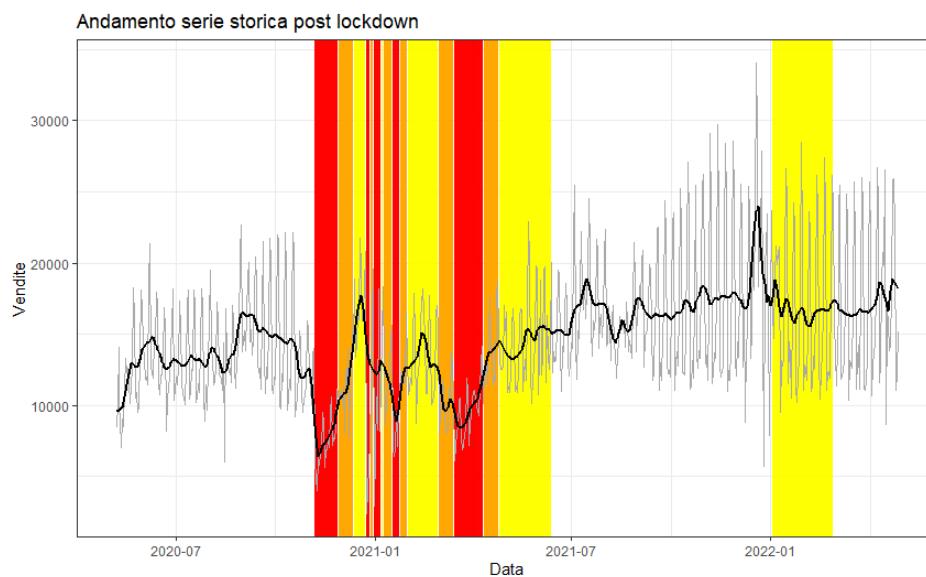


Figura 28. Trend serie storica al cambiamento del colore della regione

Un problema da tenere in considerazione in tale serie storica è la presenza di valori mancanti. Nello specifico, il numero

di *missing values* è pari a 6 (figura 29), e corrispondono principalmente a giorni festivi. Essendo che nella serie storica precedentemente analizzata il ristorante è risultato aperto in tutte le festività, e quindi non conoscendo la reale causa di tali valori mancanti, si decide di imputarli.

L'imputazione viene effettuata tramite la libreria *imputeTS*[11] utilizzando l'algoritmo *Kalman Smoothing*.

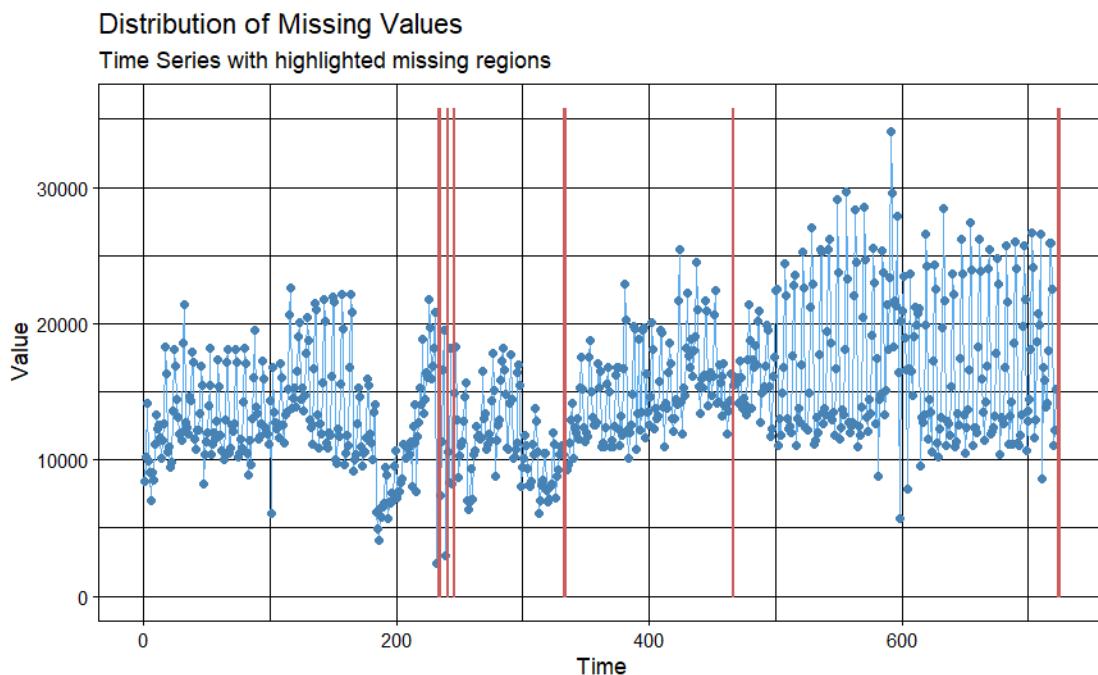


Figura 29. Distribuzione *missing values* serie storica post 1° lockdown

5.2.2 Modelli ARIMA

La prima classe di modelli utilizzata per prevedere il fatturato futuro è la classe dei modelli Arima , presentati nella sezione 5.1.2. Come per la precedente analisi predittiva, anche in questo caso è stata applicata la procedura di *Box-Jenkins*[3], assieme all'analisi di stazionarietà della serie storica.

I modelli considerati sono i seguenti:

1. *SARIMA*(1,0,2)(0,1,2), $\lambda = -0.10$. La modellazione di tutte le componenti viene lasciata ai soli parametri autoregres-sivi e a media mobile.
2. *SARIMAX*(1,1,3)(1,0,0), $\lambda = -0.10$, regressori = dummy giornaliere. In questo modello si cerca di modellare la stagionalità settimanale attraverso l'introduzione di variabili dummy indicanti i giorni della settimana.
3. *SARIMAX*(1,1,2)(1,0,1), $\lambda = -0.10$, regressori = termini di Fourier. In questo modello si cerca di modellare sia la stagionalità settimanale che quella annuale attraverso l'introduzione dei termini di Fourier (3 per la stagionalità settimanale, 15 per quella annuale).
4. *SARIMAX*(1,0,2)(1,0,2), $\lambda = -0.10$, regressori = termini di Fourier, dummy relative al colore della regione + altri regressori. In questo modello la stagionalità annuale e settimanale sono gestite attraverso i termini di Fourier, viene inserita una variabile che indica il colore della regione e infine vengono inseriti altri regressori, come una dummy che identifica i giorni di pioggia, per cercare di modellare i comportamenti di natura non stagionale, cercando di aumentare l'accuratezza delle previsioni.

Di seguito vengono riportate le varie misure di performance su dati di *training* e *test* per ogni modello (tabella 7 e 8):

Modello	RMSE	MAE	MAPE	AIC
SARIMA(1,0,2)(0,1,2)	2382.50	1694.17	13.86	-2160.11
SARIMAX(1,1,3)(1,0,0)	2343.58	1657.06	13.54	-2199.87
SARIMAX(1,1,2)(1,0,1)	2295.63	1602.45	12.84	-2190.60
SARIMAX(1,0,2)(1,0,2)	2003.77	1467.00	11.26	-2405.11

Tabella 7. Metriche di performance dei modelli sui dati di **training**

Modello	RMSE	MAE	MAPE
SARIMA(1,0,2)(0,1,2)	4008.10	2720.22	17.11
SARIMAX(1,1,3)(1,0,0)	4275.59	3277.05	20.31
SARIMAX(1,1,2)(1,0,1)	5025.21	3903.59	27.34
SARIMAX(1,0,2)(1,0,2)	4763.11	3567.35	24.80

Tabella 8. Metriche di performance dei modelli sui dati di **test**

Seguendo la procedura di *hold-out*, il modello che sui dati di training risulta avere le performance migliori è il modello **SARIMAX(1,0,2)(1,0,2)**, con MAPE pari a 11.26 % e AIC pari a -2405.11. Tali performance non vengono però confermate sui dati di test, sintomo che il modello soffre di *overfitting*. Il modello che generalizza meglio sui dati di test è **SARIMA(1,0,2)(0,1,2)**, con MAPE pari a 17.11 % e RMSE pari a 4008.10.

5.2.3 Modelli UCM

Per modellare l'andamento post primo lockdown sono quindi stati realizzati quattro modelli:

1. **UCM con ritardi** è costruito sfruttando unicamente le informazioni contenute nella serie storica degli incassi

$$Y_t = LLT + SEAS_{7gg}$$

La stagionalità a 7 giorni è descritta con funzioni trigonometriche.

2. **UCM con ritardi e regressori** sviluppato aggiungendo il regressore indicante le festività, un attributo binario che distingue i giorni festivi e feriali.

$$Y_t = LLT + SEAS_{7gg} + Fest$$

3. **UCM con ritardi e regressori e colori regioni** uguale al modello precedente con l'aggiunta di un attributo per modellare i diversi colori delle regioni.

4. **UCM con ritardi e dummy** realizzato modellando individualmente le festività.

Di seguito, vengono riportati i valori di performance dei differenti modelli:

Modello	RMSE	MAE	MAPE
UCM ritardi	3084.06	1731.38	11.95
UCM ritardi + regressori	3218.39	1767.95	12.38
UCM ritardi + regressori + colori reg.	3622.55	2462.94	15.61
UCM ritardi + dummy	3114.03	1746.41	11.85

Tabella 9. Metriche di performance dei modelli UCM

5.2.4 Modelli di Machine Learning

Per quanto riguarda le previsioni future con il modello Random Forest, l'approccio è stato analogo a quello seguito per le previsioni del primo lockdown, sviluppando 3 modelli identici a quelli descritti in sezione 5.1.4. La sola differenza sta nell'aggiunta di regressori che modellano il colore della regione per l'emergenza Covid.

I risultati ottenuti sono i seguenti:

Modello	RMSE	MAE	MAPE
RF ritardi	3805.39	1706.86	11.81
RF ritardi + regressori	3059.49	1694.95	11.39
RF ritardi + dummy	3125.89	1699.85	11.51

Tabella 10. Metriche di performance dei modelli Random Forest

5.2.5 Time Series Cross-Validation: valutazione accuratezza puntuale di previsione e confronto tra modelli

Time Series Cross-Validation: definizione della procedura di validazione [4][5][6][7][8]

In fase di validazione è stata costruita una procedura di time series cross-validation con *rolling forecasting origin, increasing in-sample e non-constant holdout*, adattata per previsioni *multi-step*, con caratteristiche del tutto analoge a quelle descritte nella sezione 5.1.5.

Anche in questo caso la procedura di time series cross-validation è stata implementata in *R*, utilizzando, per la famiglia di modelli ARIMA, una versione leggermente modificata della funzione `tsCV()` [10] del pacchetto `forecast` e definendo invece delle funzioni inedite per le famiglie di modelli UCM e Random Forest.

Time Series Cross-Validation: definizione dei parametri

In Tabella 11 sono riportati i valori dei parametri di time series cross validation, H (orizzonte o numero di step di previsione) e m (dimensione della finestra iniziale di training), selezionati per i diversi modelli stimati.

Modello	Orizzonte/step di previsione (H)	Osservazioni nella finestra iniziale di training (m)
SARIMA(1,0,2)(0,1,2)	60	180
SARIMAX(1,1,3)(1,0,0)	60	180
SARIMAX(1,1,2)(1,0,1)	60	540
SARIMAX(1,0,2)(1,0,2)	60	540
RF ritardi	60	365
RF ritardi + regressori	60	365
RF ritardi + dummy	60	365
UCM ritardi	60	365
UCM ritardi + regressori	60	365
UCM ritardi + regressori + colori reg.	60	540
UCM ritardi + dummy	60	365

Tabella 11. Parametri di time series cross validation per la validazione dei modelli stimati sulla serie storica post 1° lockdown.

H è uguale a 60 per tutti i modelli, valore coerente con il numero di step di previsione da effettuare a partire dalla fine della serie storica post 1° lockdown e che identifica il periodo di previsione che inizia il giorno 29 aprile 2022 e termina il 27 giugno 2022. Il parametro m , il numero di osservazioni nella finestra iniziale di training, ovvero la porzione di serie storica su cui non si esegue la cross-validation, ha invece subito una fase di ottimizzazione che ha portato all'ottenimento di valori differenti per i diversi modelli. Le considerazioni fatte in fase di ottimizzazione sono analoghe a quelle espresse in Sezione 5.1.5.

Time Series Cross-Validation: risultati e selezione del modello più performante

I risultati del processo di time series cross validation sui modelli stimati sulla serie storica post 1° lockdown sono riportati in Figura 30 e in Tabella 12. In particolare, in Tabella 12 si mostrano: le diverse metriche per la valutazione dell'accuratezza di previsione puntuale in funzione dell'orizzonte di previsione h (non sono presenti le accuratezze per tutti e 60 gli step di previsione, ma sono stati campionati i valori ogni 7 step); i valori delle accuratezze medie $h_i - h_f \text{ avg}$ (per gli intervalli di step 1-7, 7-30 e 30-60), utili a verificare le performance di previsione a breve, medio e lungo termine e i valori di accuratezza media lungo tutto l'orizzonte di previsione ($h_1 - h_{60} \text{ avg}$). Analogamente in Figura 30 sono rappresentate le diverse metriche per la valutazione dell'accuratezza di previsione puntuale in funzione dell'orizzonte di previsione h (per tutti i 60 step).

Tabella 12. Metriche per la valutazione dell'accuratezza di previsione puntuale, dei modelli stimati sulla serie storica post 1° lockdown, in funzione dell'orizzonte di previsione h (campionati ogni 7-step) e medi ($h_i - h_f \text{ avg}$): RMSE (a), MAE (b) e MAPE (c). I valori relativi alle performance migliori sono evidenziati.

(a) Root Mean Squared Error (RMSE) in funzione dell'orizzonte di previsione h .

h	SARIMA (1,0,2)(0,1,2)	SARIMAX (1,1,3)(1,0,0)	SARIMAX (1,1,2)(1,0,1)	SARIMAX (1,0,2)(1,0,2)	RFr	RFr+r	RFr+d	UCMr	UCMr+r	UCM r+r+c	UCMr+d
1	3190	3127	4368	4018	2919	2802	2805	3013	3272	4169	2652
7	3326	3292	5159	4498	3091	2970	2958	3352	3392	4607	2969
14	3752	3708	6006	5001	3387	3179	3192	3496	3437	4661	3333
21	4070	4003	6067	5239	3691	3427	3410	3822	3794	5205	3721
28	4301	4231	5636	5139	3922	3577	3542	4013	3903	5623	3863
35	4414	4345	5404	5065	3948	3686	3650	4143	4022	6003	4023
42	4391	4297	4924	5099	4075	3686	3689	4209	4067	6088	4154
49	4363	4253	4866	5354	4003	3734	3726	4279	4129	6261	4278
60	4105	4143	4479	4630	4162	3871	3900	4455	4331	5879	4453
1-7 avg	3312	3273	4827	4295	3019	2915	2915	3235	3419	4526	2824
7-30 avg	3984	3961	5795	5111	3641	3371	3359	3821	3829	5282	3624
30-60 avg	4292	4276	4973	5096	4060	3728	3730	4267	4137	6074	4234
1-60 avg	4059	4038	5271	5008	3778	3496	3493	3976	3935	5590	3836

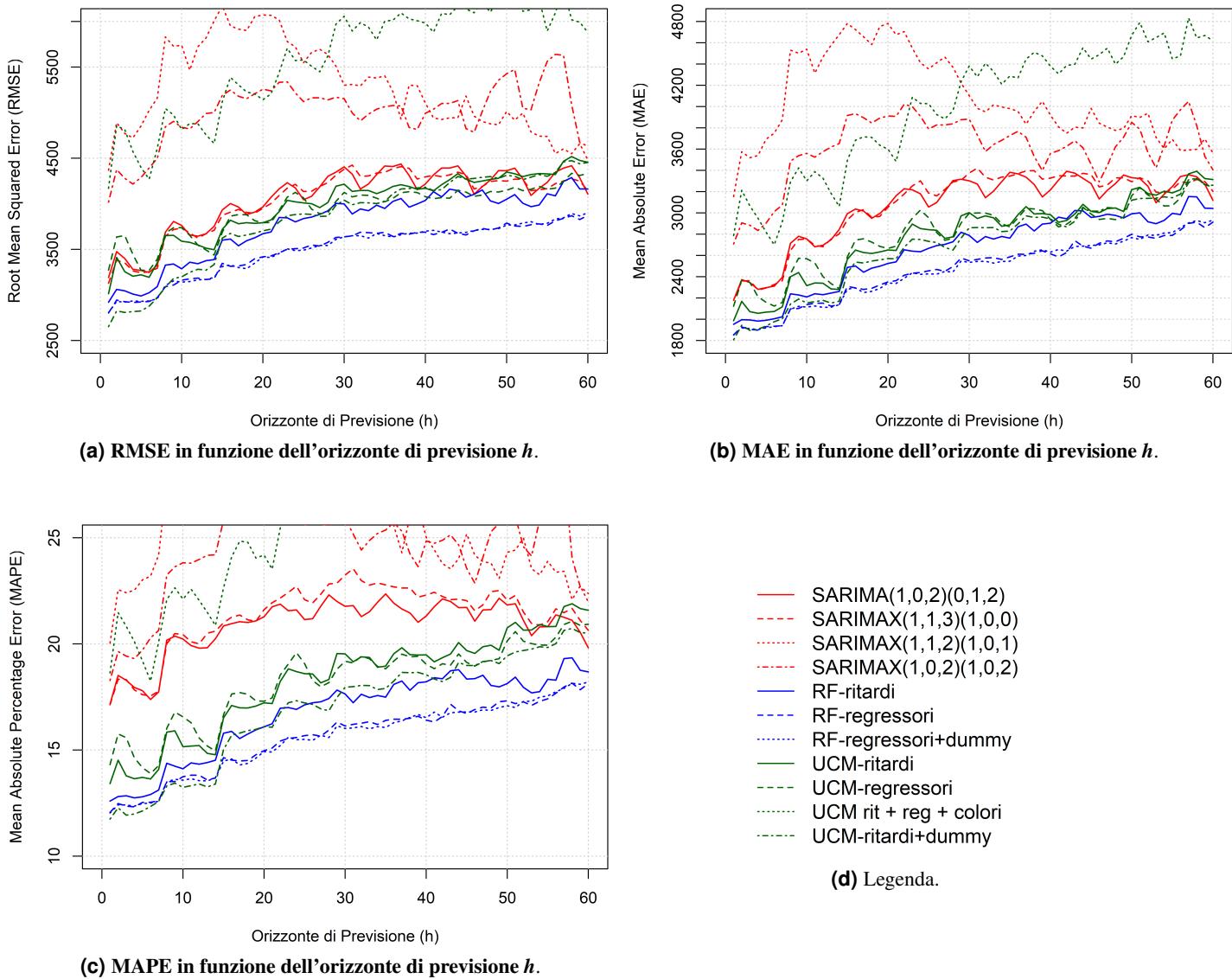
(b) Mean Absolute Error (MAE) in funzione dell'orizzonte di previsione h .

h	SARIMA (1,0,2)(0,1,2)	SARIMAX (1,1,3)(1,0,0)	SARIMAX (1,1,2)(1,0,1)	SARIMAX (1,0,2)(1,0,2)	RFr	RFr+r	RFr+d	UCMr	UCMr+r	UCM r+r+c	UCMr+d
1	2178	2178	3153	2707	1952	1852	1851	1985	2122	2787	1806
7	2394	2362	3875	3074	2022	1938	1939	2116	2157	2916	2003
14	2844	2802	4676	3652	2261	2135	2136	2279	2287	3060	2169
21	3161	3109	4674	3907	2539	2362	2336	2646	2687	3483	2567
28	3303	3291	4368	3837	2728	2495	2444	2774	2759	3960	2649
35	3381	3345	3980	3727	2762	2586	2545	2886	2868	4286	2829
42	3391	3346	3748	3626	2944	2627	2640	2928	2893	4386	2867
49	3355	3328	3792	3711	2951	2760	2734	3012	2958	4468	3013
60	3118	3203	3562	3405	3042	2912	2931	3313	3243	4619	3259
1-7 avg	2311	2308	3586	2903	1991	1912	1910	2076	2212	2921	1923
7-30 avg	3014	3017	4527	3792	2497	2321	2301	2631	2694	3678	2499
30-60 avg	3270	3313	3826	3642	2935	2710	2709	3082	3045	4489	3032
1-60 avg	3060	3082	4067	3613	2657	2468	2460	2792	2813	3995	2698

(c) Mean Absolute Percentage Error (MAPE) in funzione dell'orizzonte di previsione h .

h	SARIMA (1,0,2)(0,1,2)	SARIMAX (1,1,3)(1,0,0)	SARIMAX (1,1,2)(1,0,1)	SARIMAX (1,0,2)(1,0,2)	RFr	RFr+r	RFr+d	UCMr	UCMr+r	UCM r+r+c	UCMr+d
1	17.2	17.1	20.0	18.3	12.6	12.1	12.0	13.4	14.3	18.6	11.7
7	17.7	17.8	24.2	20.6	13.1	12.6	12.6	14.1	14.3	20.1	12.6
14	20.2	20.6	28.3	24.2	14.5	13.7	13.7	14.8	15.0	20.9	13.4
21	21.7	21.9	28.9	25.8	16.2	15.1	14.9	17.2	17.5	23.5	16.1
28	22.3	22.8	27.7	25.6	17.4	15.9	15.6	18.3	18.2	27.2	16.8
35	22.4	22.8	25.9	25.4	17.5	16.4	16.1	19.0	18.8	29.6	17.9
42	22.1	22.2	23.8	24.9	18.4	16.6	16.5	19.3	19.1	30.4	18.2
49	22.1	22.5	23.1	25.7	18.0	17.3	17.0	19.7	19.3	30.9	19.1
60	19.8	20.6	22.0	22.4	18.7	18.1	18.2	21.6	20.9	30.0	20.5
1-7 avg	17.8	17.8	22.6	19.6	12.8	12.4	12.4	13.8	14.7	19.8	12.1
7-30 avg	21.0	21.5	28.2	25.3	16.0	14.8	14.7	17.1	17.5	24.7	15.6
30-60 avg	21.4	22.1	24.4	25.0	18.2	17.0	16.9	20.1	19.6	30.3	19.1
1-60 avg	20.9	21.3	25.7	24.5	16.7	15.6	15.5	18.2	18.2	26.9	16.9

Figura 30. Metriche per la valutazione dell’accuratezza di previsione puntuale, RMSE (a), MAE (b) e MAPE (c), dei modelli stimati sulla serie storica post 1° lockdown, in funzione dell’orizzonte di previsione h e legenda (d). L’asse y è troncato, al fine di mostrare meglio i modelli più performanti.



Dai risultati si evince:

- Come atteso le performance di accuratezza di previsione puntuale tendono a peggiorare all’aumentare dell’orizzonte di previsione h .
- Le metriche di performance medie (1-60 avg), ottenute mediante time series cross validation, sono paragonabili o, nella maggior parte dei casi, peggiori rispetto a quelle ottenute mediante la procedura di hold-out per quasi tutte le famiglie di modelli.
- Anche in questo caso, l’utilizzo delle variabili *dummy* per indicare le singole festività, rispetto a un’unica variabile indicante se il giorno è o meno festivo, in tutti i modelli e per tutte le metriche, non sembra migliorare sensibilmente le performance di previsione dei modelli e per buona parte di essi si ottiene un peggioramento delle stesse.
- I modelli Random Forest (ritardi + regressori), Random Forest (ritardi + dummy) sovraprofornano gli altri modelli in termini di metriche medie (1-60 avg) e lungo i diversi orizzonti di previsione, anche se il modello UCM ritardi + dummy presenta performance leggermente migliori per previsioni a breve termine. Si è scelto quindi di utilizzare per la fase di

previsione il modello Random Forest (ritardi + regressori). Infatti, le performance dei due modelli Random Forest sono molto simili, si è quindi optato per il modello più parsimonioso.

5.2.6 Previsioni e considerazioni

Fra i modelli proposti, le metriche calcolate hanno portato alla selezione del modello Random Forest ritardi + regressori. Con questo modello sono state eseguite le previsioni per il fatturato del ristorante 1 per un orizzonte temporale di 60 giorni a partire dall'ultimo dato a disposizione. Di seguito i risultati:

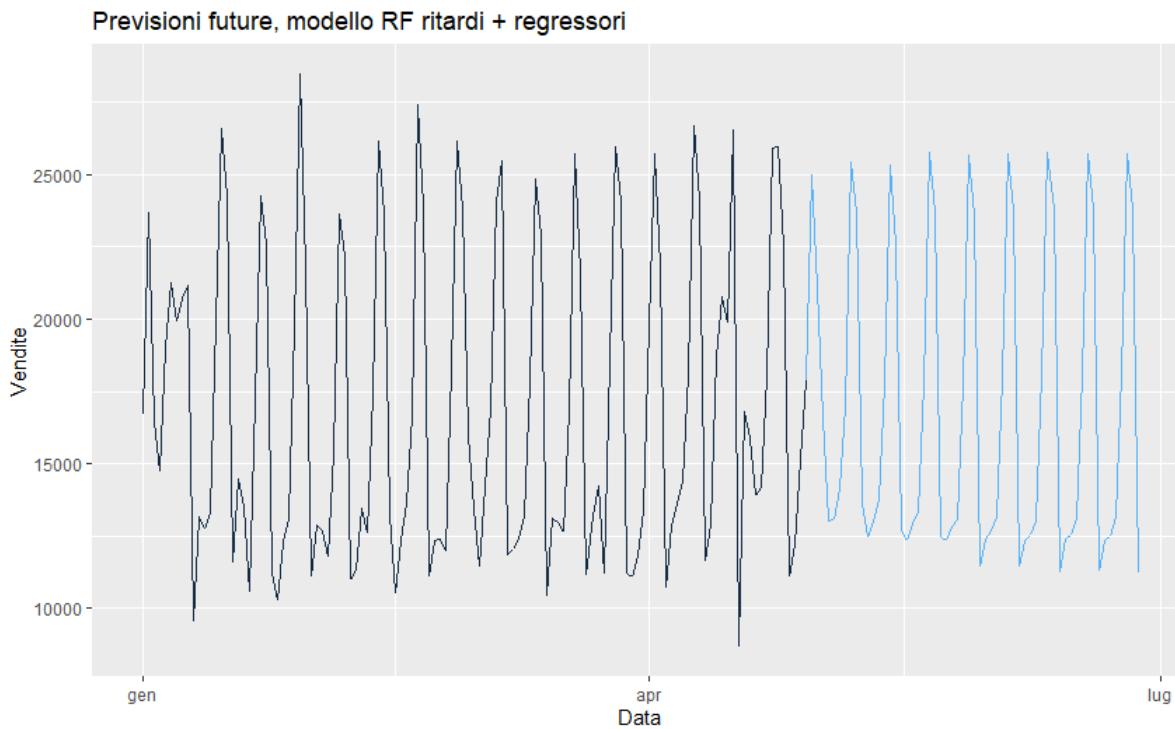


Figura 31. Previsioni fatturato per i futuri 60 giorni

Le previsioni continuano a mostrare l'andamento periodico settimanale caratteristico della serie storica. Il modello utilizzato, probabilmente per una quantità insufficiente di dati, non riesce comunque a prevedere andamenti particolari come quelli che si possono notare all'inizio di gennaio o nel mese di aprile per i dati reali.

Inoltre, confrontando queste stime con quelle della sezione 5.1.6, possiamo notare come il trend della serie storica si sia notevolmente abbassato. Nonostante il ristorante sia tornato ad un andamento regolare, infatti, è possibile apprezzare come non sia ancora ritornato ai livelli precedenti alla pandemia.

6. Conclusioni

A seguito delle analisi condotte, è possibile trarre alcune conclusioni rispetto alle domande di ricerca proposte.

L'analisi esplorativa iniziale ha permesso di identificare alcune caratteristiche peculiari delle serie storiche a disposizione, utili per la successiva realizzazione dei modelli predittivi. Ad esempio, è stato possibile apprezzare il forte impatto avuto dalla pandemia da Covid-19 sugli incassi, alla quale i ristoranti hanno risposto con un aumento dei prezzi del ristorante, questo lo si è notato da un innalzamento del prezzo medio per scontrino.

Il secondo obiettivo del presente lavoro era quello di stimare le perdite accusate nel periodo iniziale della pandemia. Innanzitutto sono stati valutati differenti modelli e sono stati selezionati i più performanti ed affidabili in base ai dati a disposizione. Le previsioni, ottenute con i migliori modelli, hanno quindi confermato il forte danno causato dalle chiusure forzate di inizio 2020. I due modelli considerati hanno previsto un mancato incasso nel range tra 21% e 16% del fatturato totale dell'anno precedente per il solo periodo del lockdown.

Infine, è stata realizzata una stima delle vendite nei 60 giorni successivi ai dati a disposizione. Individuare un modello in grado di prevedere in modo efficace l'andamento degli incassi rappresenta uno strumento utile per l'analisi di mercato, che consente, per esempio, una miglior gestione delle risorse.

Il lavoro proposto può comunque essere ulteriormente esteso analizzando le serie storiche degli altri ristoranti e integrando ulteriori dati. Le previsioni proposte potrebbero essere ulteriormente migliorate e rese consistenti avendo a disposizione una quantità maggiore di dati della serie storica. Inoltre, avendo a disposizione informazioni riguardanti, per esempio, il listino prezzi del locale, sarebbe possibile effettuare un'analisi più approfondita dell'evoluzione delle spese attraverso il periodo pandemico.

Inoltre, si potrebbe pensare di esplorare più a fondo la classe dei modelli di Machine Learning alla ricerca di algoritmi più efficaci del Random Forest proposto.

Bibliografia e sitografia

- [1] *Multiple seasonal decomposition*. URL: <https://search.r-project.org/CRAN/refmans/forecast/html/mstl.html>.
- [2] G. E. P. Box e D. R. Cox. “An Analysis of Transformations”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 26.2 (1964), pp. 211–243.
- [3] Stanley Jere e Edwin Moyo. “Modelling Epidemiological Data Using Box-Jenkins Procedure”. In: *Open Journal of Statistics* 06 (gen. 2016), pp. 295–302.
- [4] R.J. Hyndman e G. Athanasopoulos. *FORECASTING: PRINCIPLES AND PRACTICE (2nd Ed)*. OTexts: Melbourne, Australia., 2018. URL: <https://otexts.com/fpp2/>.
- [5] R.J. Hyndman e G. Athanasopoulos. *FORECASTING: PRINCIPLES AND PRACTICE (3rd Ed)*. OTexts: Melbourne, Australia., 2021. URL: <https://otexts.com/fpp3/>.
- [6] Elizabeth Holmes. *Fisheries Catch Forecasting*. bookdown, 2020. URL: <https://fish-forecast.github.io/Fish-Forecast-Bookdown/>.
- [7] Ivan Svetunkov. *Forecasting and Analytics with ADAM*. bookdown, 2022. URL: www.openforecast.org/adam.
- [8] Leonard J. Tashman. “Out-of-sample tests of forecasting accuracy: An analysis and review.” In: *International Journal of Forecasting* 16 (2000), pp. 437–450. DOI: [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0).
- [9] Ivan Svetunkov e Fotios Petropoulos. “Old dog, new tricks: a modelling view of simple moving averages.” In: *International Journal of Production Research* 56 (18) (2018), pp. 6034–6047. DOI: <https://doi.org/10.1080/00207543.2017.1380326>.
- [10] Rob J. Hyndman. *tsCV: Time series cross-validation*. URL: <https://www.rdocumentation.org/packages/forecast VERSIONS/8.9/topics/tsCV>. (accessed: January 4, 2023).
- [11] *imputeTS: Time Series Missing Value Imputation*. URL: <https://steffenmoritz.github.io/imputeTS/>.