Ristorazione, Covid-19 e previsioni future

Lorgna Lorenzo[1,2], Marzorati Stefano[1,3], Palomba Eleonora[1,4]

Sommario

Il seguente lavoro è stato realizzato con l'obiettivo di analizzare l'impatto che il diffondersi dell'epidemia da Covid-19 ha avuto sul settore della ristorazione. Tale risultato è stato ottenuto a partire dallo studio dell'andamento dei dati relativi a vendite e scontrini giornalieri di un campione costituito da sei ristoranti, collocati in un territorio compreso tra le regioni Lombardia ed Emilia-Romagna. Nello specifico, i dati utilizzati sono riconducibili alle vendite effettuate da gennaio 2017 ad aprile 2021, ragion per cui è stato possibile effettuare una comparazione tra l'andamento delle vendite registrate prima che le misure governative nazionali predisponessero il blocco completo delle attività, e quelle registrate in seguito alle riaperture. Inoltre, si è cercato di individuare le variabili più significative che permettessero di implementare alcuni modelli predittivi (*SARIMA & SARIMAX*, *Random Forest*, *Prophet*, *HoltWinters*, *TBATS*), utili per determinare i valori delle vendite che si sarebbero potuti registrare durante il periodo pandemico o in alternativa valori futuri che potrebbero assumere le vendite nei mesi successivi ad aprile 2021. Le analisi condotte hanno mostrato che la pandemia ha avuto delle ripercussioni economiche in negativo sulle vendite e sugli scontrini per tutti gli esercizi di ristorazione in questione. Considerando, invece, i modelli previsionali, le cui migliori performance sui dati settimanali sono state registrate dalla famiglia dei modelli *ARIMA*, è emersa la presenza di una ripresa, seppur contenuta, dei ristoranti per il periodo successivo ai mesi di lockdown.

Keywords

R — timeseries — Restaurants' sales — Covid — forecast — SARIMA — SARIMAX — Random Forest — Prophet — HoltWinters — TBATS

[1] *Corso di Laurea Magistrale in Data Science, Università degli Studi di Milano-Bicocca*
[2] *l.lorgna@campus.unimib.it - 829776*
[3] *s.marzorati11@campus.unimib.it - 830272*
[4] *e.palomba4@campus.unimib.it - 876479*

Indice

Introduzione

La quasi totalità dei settori produttivi è stata impattata duramente in seguito al diffondersi dell'epidemia da Covid-19. In particolare, la predisposizione di alcune drastiche, sebbene necessarie, misure di sicurezza come il blocco totale delle attività commerciali subentrato in tutta Italia dal 9 marzo 2020, ha provocato ingenti cali di fatturato, registrati omogeneamente in tutto il territorio nazionale. Alcune tipologie di settori, tuttavia, hanno risentito degli effetti di tali provvedimenti in maniera decisamente più significativa di altri, fra cui il settore della ristorazione: basti pensare che nel 2020 sono state registrate perdite di fatturato pari a circa il 46% rispetto al 2019. Essendo dunque risultato uno tra i più colpiti, il settore della ristorazione costituisce un caso particolar-

mente rappresentativo di come la pandemia abbia modificato le vendite registrate nel corso del tempo, dato che a sua volta lascia desumere altrettante interessanti informazioni circa i cambiamenti delle abitudini dei consumatori. A tal proposito, tra gli obiettivi primari dell'analisi vi è stato quello di delineare in modo analitico le ripercussioni economiche per i sei ristoranti presi in esame. A supporto di ciò si è voluto analizzare come sarebbero andate le vendite durante il periodo della pandemia nel quale i ristoranti sono stati costretti alla chiusura. In aggiunta alle analisi svolte sono state effettuate anche delle previsioni per date per cui non si hanno ancora a disposizione le vendite effettive. Il periodo in questione è da considerarsi dalla seconda metà di aprile 2021. L'obiettivo di tale previsione è stato quello di comprendere meglio la portata della ripresa economica dei sei esercizi.

Per condurre le analisi desiderate, sono stati effettuati alcuni passaggi conseguenziali: dapprima è stato eseguito un pre-processing dei dati allo scopo di renderli più fruibili, in seguito sono state effettuate diverse integrazioni del dataset originario con una serie di dataset esterni allo scopo di aggiungere informazioni relative a dati meteorologici, eventi sportivi, nonché al giorno della settimana, mese, stagione e se il periodo considerato potesse essere considerato festivo o meno; inoltre, allo scopo di inserire all'interno delle analisi una variabile che potesse essere considerata come riassuntiva della maggior o minore presenza di contagi (comprendendo anche le conseguenti misure contenitive), sono stati considerati i colori delle zone predisposti come da decreto nei diversi periodi.
Successivamente è stata condotta una fase di esplorazione dei dati in cui, grazie all'ausilio di alcune visualizzazioni, è stato possibile iniziare a visionare l'andamento delle vendite nel corso del tempo per ognuno dei sei ristoranti. In aggiunta è stato realizzato anche un confronto per ogni singolo ristorante prendendo in esame l'estate 2019 e l'estate 2020, cercando di comprendere l'influenza della pandemia sulle vendite in periodo estivo. In conclusione a questa fase di esplorazione è stato analizzato per i sei ristoranti anche il trend degli scontrini, con l'intento di capire se a seguito delle misure restrittive ci fosse stata una diminuzione o meno del numero di scontrini e del loro valore medio.
La parte centrale dello studio ha riguardato invece l'implementazione di diversi modelli che consentissero un'analisi accurata delle serie storiche dei sei esercizi di ristorazione. In particolare sono stati utilizzati i modelli *SARIMA*, *SARIMAX*, *Random Forest*, *Prophet*, *HoltWinters* e *TBATS*. I seguenti modelli sono stati utilizzati per realizzare previsioni durante il periodo Covid e per il periodo successivo ad aprile 2021, rispettivamente per immaginare come sarebbero andate le vendite se non ci fosse stata la pandemia e per avere una previsione della ripresa economica nella seconda metà del 2021.

1. Obiettivi

Le domande di ricerca che sono state formulate sono le seguenti:

- *Considerando l'intero periodo di attività dei sei ristoranti emergono similarità e/o differenze nell'andamento delle vendite? In particolare, quale effetto ha avuto la pandemia sull'attività economica dei sei ristoranti presi in esame? Ci sono ristoranti che hanno avuto una reazione migliore in seguito al periodo di chiusure?*
- *Ci sono state sostanziali differenze tra l'estate 2019 e quella del 2020, la prima influenzata dalla pandemia? Nonostante il Covid-19, con i primi allentamenti delle misure restrittive, i ristoranti in estate sono tornati operativi, tuttavia la situazione di incertezza ha giocato un ruolo rilevante nell'andamento delle vendite.*
- *Quali insight si possono ricavare analizzando il numero di scontrini effettuati e il valore delle vendite? Quali strategie sono state implementate dai gestori delle attività per far fronte alle perdite?*
- *Come sarebbero andate le vendite senza l'effetto del Covid? Riuscire a quantificare le perdite dovute al periodo di lockdown, per i gestori di un'attività, può essere utile per avere una visione più chiara del bilancio di attività. Come sarà l'andamento delle vendite nei mesi successivi ad aprile 2021? Comprendere con anticipo i trend futuri permetterebbe ai ristoratori di pianificare al meglio i primi mesi di ripartenza, gestendo in maniera più oculata gli approvvigionamenti e il personale.*

2. Aspetti metodologici

Allo scopo di ottenere una migliore comprensione dell'andamento delle vendite nel corso del tempo ed elaborare previsioni, è stato scelto di utilizzare diversi modelli, quali:

- *SARIMA e SARIMAX*

- *Random Forest*
- *Prophet*
- *HoltWinters*
- *TBATS*

SARIMA e SARIMAX Nello specifico, allo scopo di ottenere una stima del fatturato relativo al periodo Covid, è stato inizialmente effettuato un tentativo utilizzando un modello del tipo *SARIMA* (seasonal autoregressive integrated moving avarage) [4, 5]. Un modello *SARIMA (p,d,q) (P,D,Q)* è un processo lineare non stazionario impiegato nello studio di serie storiche che presentano determinate caratteristiche. Esso risulta essere un'estensione del modello *ARIMA* (autoregressive integrated moving average) dal momento in cui viene modellata anche la componente stagionale. Il modello *ARIMA* viene derivato, a sua volta, dai modelli *ARMA* (autoregressive moving avarage) a cui vengono applicate le differenze di ordine d, allo scopo di ottenere la condizione di stazionarietà (media, varianza e autocorrelazione costanti). I parametri p e q indicano, rispettivamente, la misura con cui il modello si basa su valori passati e il valore ritardato degli errori di previsione nel modello. Più nel dettaglio, tali parametri risultano dalle componenti del modello:

- *AR (Auto Regressive)*: riflette la dipendenza tra un'osservazione e n osservazioni ritardate
- *I (Integraged)*: per esplicitare che la serie è resa stazionaria tramite differenziazione
- *MA (Moving Average)*: riflette la dipendenza tra un'osservazione ed un errore residuo da un modello a media mobile applicato alle osservazioni ritardate

Nel caso dei modelli *SARIMA* sono da considerare anche i parametri stagionali: P, D, Q e m.
Nello specifico, considerando la famiglia dei modelli *ARIMA*, la componente *Autoregressive* risulta essere:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \varepsilon_1 \quad (1)$$

mentre la componente *Moving Average*:

$$Y_t = \alpha + \varepsilon + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + ... + \phi_q \varepsilon_{t-q} \quad (2)$$

Nella fase di implementazione del modello si cerca di ricavare i valori dei parametri del modello manualmente, attraverso l'ispezione dei dati a disposizione. In particolare:

- L'ordine del parametro p viene ricavato osservando il grafico *PACF* (Partial Autocorrelation Function): qualsiasi autocorrelazione presente in una serie stazionaria può essere corretta aggiungendo termini *AR* sufficienti. Inizialmente, dunque, viene posto l'ordine del termine *AR* uguale al numero di ritardi che attraversano il limite di significatività nel grafico *PACF* .
- L'ordine d viene determinato in modo che il grafico *ACF* raggiunga lo zero abbastanza rapidamente: se le autocorrelazioni sono positive per molti lag (10 o più), allora la serie necessiterà di ulteriori differenziazioni.
- L'ordine del parametro q viene viene determinato visionando il grafico *ACF* (Autocorrelation Function): un termine *MA* può essere definito come l'errore della previsione ritardata e il grafico *ACF* indica quanti termini *MA* sono necessari per rimuovere qualsiasi autocorrelazione dalla serie stazionaria. Ad esempio, se dal grafico emergesse che un paio di ritardi sono al di sopra della linea di significatività, q verrebbe provvisoriamente fissato uguale a 2.

Procedendo in modo analogo ma prendendo in considerazione la componente stagionale del modello *SARIMA* sono stati individuati anche i valori dei parametri P, D, Q e m.
Effettuando diversi tentativi è possibile variare i parametri alla ricerca del modello migliore: generalmente viene preferito il modello con valore di *AIC* (Akaike Information Criteria) minore. Si tratta di una statistica ampiamente utilizzata, in quanto in grado di quantificare sia la bontà del fit che la parsimonia del modello.

$$\mathrm{AIC} = 2k - 2\ln\widehat{L} \quad (3)$$

$$\mathrm{AICc} = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2} \quad (4)$$

Successivamente, si è deciso di continuare le analisi utilizzando la funzione *auto.arima* del pacchetto *forecast*. In tal modo è possibile automatizzare il processo di selezione delle coordinate dei parametri presentati precedentemente, al fine di ottenere i valori ottimali per lo specifico set di dati a disposizione ed ottenere previsioni migliori.
Per la misurazione della qualità del modello ottenuto

con *auto.arima*, oltre alla valutazione dell'indice *AIC* citato in precedenza, si è provveduto a plottare i grafici *ACF* e *PACF* dei residui e a calcolare il *RMSE* (Root Mean Square Deviation) e il *MAPE* (Mean Absolute Percentage Error).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{5}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i} \tag{6}$$

Per cercare di modellare con più precisione i dati a disposizione sono stati considerati anche i modelli *SARIMAX* (seasonal autoregressive integrated moving avarage exogenous), ovvero dei modelli *SARIMA* che considerano dei regressori esterni.

Random Forest Il *Random Forest* [6] è un algoritmo di apprendimento supervisionato. Si tratta dunque di una metodologia versatile di machine learning utilizzata solitamente per affrontare problemi di classificazione e regressione. Per poter applicare tale algoritmo nel contesto delle serie storiche è bene prima modellare i dati in modo tale da avere un problema di apprendimento supervisionato. Nella pratica il *Random Forest* combina molti alberi decisionali in unico modello e il risultato finale risulta essere la media del risultato numerico restituito dai diversi alberi. La strategia che consiste nell'utilizzare un insieme di modelli che portano ad avere un unico risultato migliore è nota come apprendimento ensemble, di cui il *Random Forest* rappresenta uno dei casi più famosi. Nella costruzione del modello è possibile considerare le sole variabili ritenute importanti, tramite l'utilizzo di un comando specifico in R. Per misurare le performance dell'algoritmo vengono utilizzati *RMSE* e *MAPE*, metriche presentate precedentemente.

Prophet *Prophet* [7] è una libreria open-source progettata da Facebook per fare previsioni per set di dati in serie temporali univariate. Sebbene di recente introduzione, viene spesso utilizzato perché semplice da implementare e perché consente di effettuare previsioni soddisfacenti anche per dati con trend e struttura stagionale. È inoltre molto resistente alla presenza di missing values nel set di dati, gestendo bene anche la presenza di outliers. Si basa sull'utilizzo di un modello di serie temporale composto da tre elementi: trend, stagionalità e festività. L'equazione generale del modello è, pertanto, la seguente:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{7}$$

in cui $g(t)$ è la funzione di tendenza che modella i cambiamenti non periodici nel valore della serie storica, $s(t)$ rappresenta i cambiamenti periodici (ad esempio, stagionalità settimanale e annuale) e $h(t)$ rappresenta gli effetti delle vacanze che si verificano con tempistiche potenzialmente irregolari nell'arco di uno o più giorni. Il termine di errore ε_t rappresenta eventuali cambiamenti idiosincratici che non sono carpiti dal modello; inoltre, il modello si basa sull'assunzione parametrica che ε_t sia distribuito normalmente. I vantaggi nell'utilizzo di tale modello sono molteplici:

- A differenza dei modelli *SARIMA*, le misurazioni non necessitano di essere "spaziate" regolarmente e non è necessario interpolare i valori mancati, ad esempio rimuovendo eventuali outliers; inoltre, permette di considerare componenti di stagionalità multiple insite nelle serie temporali.
- Adattamento e flessibilità elevati.
- Presenza di parametri facilmente interpretabili, che possono agilmente essere modificati per fare ipotesi sulla previsione oppure per includere nuove componenti all'interno del modello.

HoltWinters *HoltWinters* [8] è una delle tecniche di previsione più ampiamente utilizzate per la previsione di serie temporali. Nonostante non sia di recente introduzione, attualmente trova ancora applicazione nell'ambito del monitoraggio, dove viene utilizzato per scopi come il rilevamento di anomalie o pianificazione della capacità. Si basa sull'utilizzo di un lisciamento esponenziale per codificare i dati passati e utilizzarli per prevedere valori "tipici" per presente e futuro. Si tratta dunque di una versione decisamente più sofisticata della decomposizione tramite media mobile. In questo caso, i tre elementi essenziali che definiscono il comportamento delle serie temporali (media, tendenza e stagionalità) sono espressi come tre tipi di lisciamento esponenziale:

- Lisciamento esponenziale singolo: adatto per prevedere dati privi di trend o struttura stagionale, in cui però il livello dei dati potrebbe cambiare nel corso del tempo (lt)
- Lisciamento esponenziale doppio: per prevedere dati in cui esiste la componente di trend (bt)

- Lisciamento esponenziale triplo: per prevede dati in cui è presente trend e/o stagionalità (*st*)

Ne consegue che il modello *HoltWinters* è appropriato per serie storiche di tipo non-stazionarie; infatti, le previsioni sono ricavate calcolando gli effetti combinati delle tre componenti descritte. Si riportano di seguito le equazioni che compongono il modello nella sua forma additiva:

$$\begin{aligned} \hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t+h-m(k+1)} \\ \ell_t &= \alpha\,(y_t - s_{t-m}) + (1-\alpha)\,(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^*\,(\ell_t - \ell_{t-1}) + (1-\beta^*)\,b_{t-1} \\ s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \end{aligned} \tag{8}$$

Per quanto riguarda, invece, la sua forma moltiplicativa:

$$\begin{aligned} \hat{y}_{t+h|t} &= (\ell_t + hb_t)\,s_{t+h-m(k+1)} \\ \ell_t &= \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)\,(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^*\,(\ell_t - \ell_{t-1}) + (1-\beta^*)\,b_{t-1} \\ s_t &= \gamma\frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m} \end{aligned} \tag{9}$$

Il metodo additivo è preferito quando le variazioni stagionali sono pressoché costanti lungo la serie, mentre il metodo moltiplicativo è preferito quando le variazioni stagionali cambiano proporzionalmente al livello della serie. Con il metodo additivo la componente stagionale è espressa in termini assoluti nella scala delle serie osservate, e nell'equazione di livello la serie è destagionalizzata sottraendo la componente stagionale. All'interno di ogni anno, la componente stagionale raggiungerà approssimativamente lo zero. Con il metodo moltiplicativo la componente stagionale è espressa in termini relativi (percentuali) e la serie è destagionalizzata dividendo per la componente stagionale.

TBATS Il modello *TBATS* [9] è in grado di gestire stagionalità complesse quali, ad esempio, stagionalità non lineari, multiple, "non-nidificate", su grandi periodi e, in generale, senza particolari vincoli. Il nome di questo modello deriva dalle sue componenti principali:

- *Trigonometric seasonality*
- *Box-Cox transformation*
- *ARMA errors*
- *Trend*
- *Seasonal components*

Il motivo per cui viene proposta una formulazione trigonometrica è da attribuirsi al fatto che questa scomposizione porta all'identificazione e all'estrazione di componenti stagionali complesse che altrimenti non sarebbero evidenti
Per la scelta del modello migliore, basata anch'essa tramite *AIC*, *TBATS* considererà diverse alternative e proverà a fittare diversi modelli, quali:

- con la trasformazione Box-Cox e senza
- con e senza trend
- con e senza trend damping
- con e senza procedimenti *ARMA(p,q)* per modellare i residui
- modelli non stagionali
- diversi tipi di analisi armoniche per modellare gli effetti stagionali

Anche questo modello, come quello precedente, trae origine dai metodi basati sul lisciamento esponenziale e può essere descritto dalle seguenti equazioni:

$$\begin{aligned} y_t^{(\omega)} &= \begin{cases} \frac{y_t^{\omega}-1}{\omega}; & \omega \neq 0 \\ \log y_t & \omega = 0 \end{cases} \\ y_t^{(\omega)} &= \ell_{t-1} + \phi b_{t-1} + \sum_{i=1}^{M} s_{t-m_i}^{(i)} + d_t \\ \ell_t &= \ell_{t-1} + \phi b_{t-1} + \alpha d_t \\ b_t &= \phi b_{t-1} + \beta d_t \\ s_t^{(i)} &= s_{t-m_i}^{(i)} + \gamma_i d_t \\ d_t &= \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \end{aligned} \tag{10}$$

3. I dati

I dati a disposizione, forniti dall'Università degli Studi di Milano Bicocca, sono relativi alle vendite registrate da sei ristoranti localizzati nel nord Italia, tra Lombardia ed Emilia-Romagna. I dati in questione si riferiscono ad un arco temporale che va dal 1 gennaio 2017 al 12 aprile 2021. Il dataset è costituito da 1.563 osservazioni e da 13 features. Quest'ultime nello specifico sono:

- *data*: data a livello giornaliero
- *vendite 1*: vendite registrate per il primo ristorante
- *scontrini 1*: numero di scontrini registrati per il primo ristorante
- *vendite 2*: vendite registrate per il secondo ristorante

- *scontrini 2*: numero di scontrini registrati per il secondo ristorante
- *vendite 3*: vendite registrate per il terzo ristorante
- *scontrini 3*: numero di scontrini registrati per il terzo ristorante
- *vendite 4*: vendite registrate per il quarto ristorante
- *scontrini 4*: numero di scontrini registrati per il quarto ristorante
- *vendite 5*: vendite registrate per il quinto ristorante
- *scontrini 5*: numero di scontrini registrati per il quinto ristorante
- *vendite 6*: vendite registrate per il sesto ristorante
- *scontrini 6*: numero di scontrini registrati per il sesto ristorante

Eccetto la prima variabile che fa riferimento alla data di rilevazione, le restanti variabili sono di tipo quantitativo. Sono presenti numerosi missing values attribuili in parte a periodi di inattività dei ristoranti, come ad esempio le settimane di lockdown durante marzo 2020 e i giorni corrispondenti alle principali festività (Pasqua, Ferragosto, Natale), e in parte ad errori o mancate registrazioni dei valori di vendite.

3.1 Data Integration

Per poter essere nelle condizioni di individuare delle risposte alle domande di ricerca precedentemente presentate si è provveduto a procedere con una fase di integrazione dati. In particolar modo sono state ricavate le seguenti nuove variabili, decidendo di prendere in considerazione la regione Emilia-Romagna:

- *rapporto vendite scontrini*: rapporto tra vendite e scontrini giornalieri
- *stagione*: stagione in cui viene registrata la vendita
- *mese*: numero del mese in cui viene registrata la vendita
- *weekday*: numero del giorno di settimana in cui viene registrata la vendita
- *is weekend*: se il giorno di settimana in cui viene registrata la vendita è weekend o meno (si tratta di una variabile binaria che assume rispettivamente valore 1 o 0)
- *is holiday*: se il giorno in cui viene registrata la vendita corrisponde con una festività o meno (si tratta di una variabile binaria che assume rispettivamente valore 1 o 0)
- *rossa Emilia Romagna*: se il colore della regione è rosso o meno (si tratta di una variabile binaria che assume rispettivamente valore 1 o 0). I dati sono stati ricavati da: `https://github.com/imcatta/restrizioni_regionali_Covid`
- *solo asporto Emilia Romagna*: se l'unica attività possibile per il ristorante è l'asporto (si tratta di una variabile binaria che assume rispettivamente valore 1 o 0)
- *pioggia*: se il giorno in cui viene registrata la vendita corrisponde con un giorno di pioggia o meno (si tratta di una variabile binaria che assume rispettivamente valore 1 o 0). I dati sono stati ricavati da: `https://www.ilmeteo.it/portale/archivio-meteo/emilia+romagna`
- *tot vaccini Emilia Romagna*: numero totale di vaccini eseguiti nella regione a livello giornaliero. I dati sono stati ricavati da: `https://github.com/italia/Covid19-opendata-vaccini`
- *Covid*: se la data è successiva al 9 marzo 2020, giorno in cui è stato decretato il lockdown a livello nazionale e che ha sancito, in maniera simbolica, l'inizio delle ripercussioni legate al diffondersi della pandemia (si tratta di una variabile binaria che assume rispettivamente valore 1 o 0)

Durante la fase di data integration sono state considerate anche altre variabili rappresentanti ad esempio le vacanze scolastiche, i giorni dei saldi e gli eventi sportivi. Tuttavia, in seguito ad alcune analisi, tali variabili non sono state considerate utili ai fini dell'implementazione dei modelli.

4. Analisi e processo di trattamento dei dati

Nella fase di trattamento dei dati sopra presentati si è proceduto inizialmente con un'esplorazione approfondita e una decomposizione delle vendite relative al primo ristorante. A seguire sono state fatte delle osservazioni per i restanti ristoranti, cercando di mettere in evidenza le principali caratteristiche delle serie storiche a disposizione. Successivamente, sempre considerando il primo ristorante, si è posto il focus sulle differenze tra l'estate 2019 e l'estate 2020. In aggiunta è stato analizzato anche il trend degli scontrini. Per concludere sono state effettuate delle previsioni considerando le vendi-

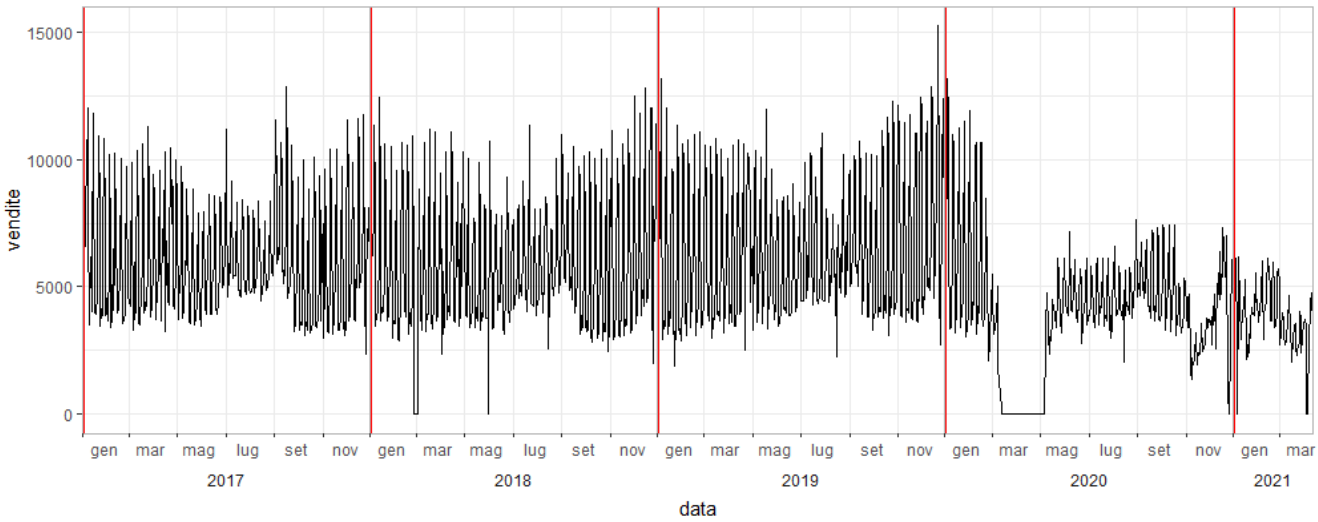

Figura 1. Vendite giornaliere ristorante 1

te del primo ristorante, sperimentando i diversi modelli precedentemente presentanti.

4.1 Ristorante 1

4.1.1 Esplorazione

La prima serie storica analizzata è stata quella riguardante le vendite giornaliere del primo ristorante. L'evoluzione delle vendite è visibile nella Figura 1.
I primi mesi dell'anno 2020 sono quelli che maggiormente attirano l'attenzione, in quanto rappresentano i dati riguardanti il periodo del Covid. In particolare già da gennaio si nota un leggero calo nelle vendite che, tuttavia, potrebbe essere assimilabile al periodo successivo alle feste natalizie. La situazione inizia però a precipitare alla fine di febbraio, quando iniziano ad essere istituite alcune zone rosse. La discesa continua fino al 9 marzo quando i ristoranti vengono chiusi e tutta Italia entra in lockdown. Si ha una ripresa nelle vendite a partire dal 18 maggio, in corrispondenza della riapertura di bar e ristoranti. Dopo diversi mesi in cui vi è stato un allentamento delle restrizioni, le vendite diminuiscono in maniera netta ad ottobre, periodo in cui l'Italia si appresta ad entrare nella seconda ondata della pandemia. Il 19 ottobre si impongono, perciò, misure restrittive per bar e ristoranti. Successivamente, le vendite registrano una ripresa, seppur la situazione di incertezza generale tende a proseguire.
Oltre al particolare periodo relativo al lockdown, il ristorante è stato chiuso nel periodo tra il 26 febbraio e il 2 marzo 2018. A seguito di una ricerca, tali chiusure si possono attribuire ad un'ondata di gelo e neve che ha colpito il nord Italia in quei giorni.

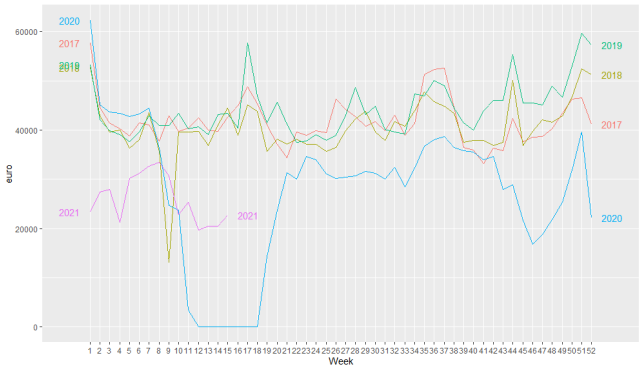

Figura 2. Stagionalità settimanale vendite ristorante 1

Nella Figura 2 sono messe a confronto le serie storiche riguardanti i diversi anni di vendite del primo ristorante. In questo caso la granularità dei dati è settimanale. Nel grafico si può notare una certa stagionalità nei dati: in particolare, sono presenti dei picchi nelle vendite nelle settimane vicino alla 15esima, vicino alla 37esima e infine nelle ultime settimane dell'anno.
Con granularità differente è possibile osservare lo stesso comportamento anche nella Figura 3 dove per gli anni precedenti al Covid-19 si hanno picchi di vendite per i mesi di aprile, luglio, settembre e dicembre. Analizzando le serie storiche per gli anni 2020 e 2021, dunque considerando l'effetto Covid, è interessante osservare che seppur su un livello inferiore di vendite i picchi si sono verificati negli stessi mesi degli anni precedenti, eccetto il mese di aprile essendo quello più critico.

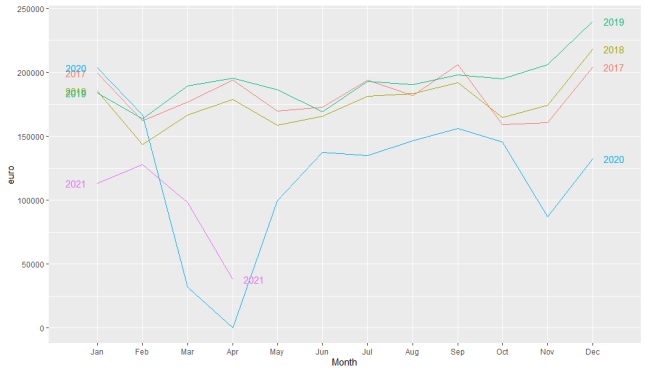

Figura 3. Stagionalità mensile vendite ristorante 1

In figura 4 viene mostrato un *seasonal subseries plot*, grafico che permette di enfatizzare la stagionalità dei dati e di vederne la sua variazione nel tempo. In questo caso il periodo di tempo scelto è quello mensile ed è possibile confermare ciò che è stato precedentemente osservato nei grafici riguardanti la stagionalità. Considerando le medie delle vendite di ciascun mese le più alte risultano essere quelle di gennaio, settembre e dicembre. Nonostante nel grafico precedente anche aprile sembrava essere uno dei mesi migliori in termini di vendite, osservando questo grafico è possibile notare che la media non è tra le migliori; questo perché risulta influenzata dagli ultimi due anni, dove aprile è stato al centro delle chiusure per il Covid-19. Non considerando la media ma osservando i singoli valori per il mese di aprile si può notare che esso negli anni 2017, 2018 e 2019 risulta essere uno dei mesi con valori di vendite maggiori.

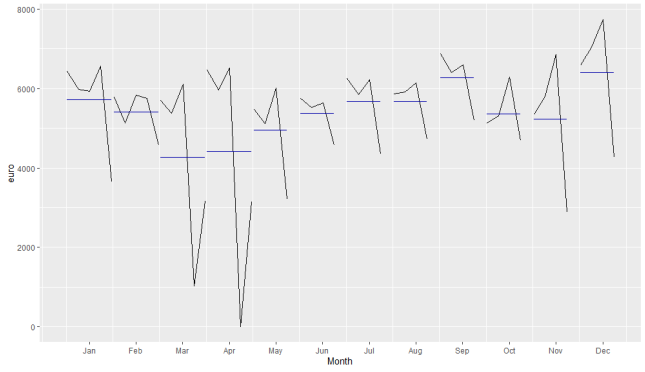

Figura 4. Seasonal subseries plot mensile vendite ristorante 1

Le osservazioni fatte per il mese di aprile sono riscontrabili anche in miglior modo nella Figura 5 che mostra la stessa tipologia di grafico della figura precedente ma con i dati che interessano solo il periodo non influenzato da Covid-19, fino al 31 dicembre del 2019. Si può dunque notare che ad aprile la media delle vendite subisce un picco.

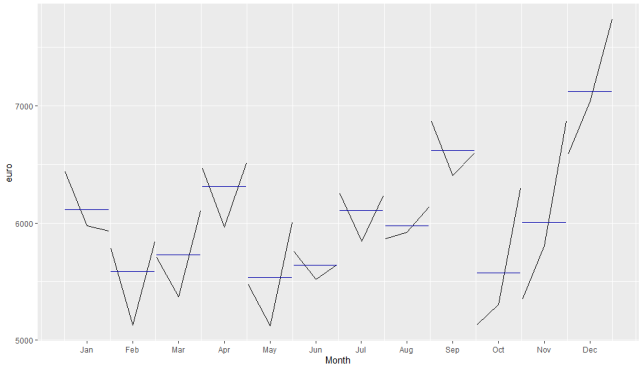

Figura 5. Media mensile vendite pre Covid-19 ristorante 1

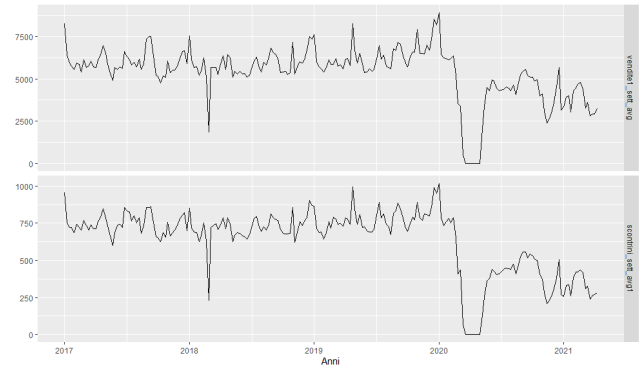

Figura 6. Confronto serie vendite e scontrini ristorante 1

Come descritto nella sezione riguardante i dati, nel dataset a disposizione sono presenti anche i dati riguardanti gli scontrini giornalieri effettuati dal ristorante. Seppur non verranno utilizzati nella sezione riguardante i modelli previsionali, è stata analizzata la correlazione che è presente tra le vendite e gli scontrini effettuati.
In un primo momento è stato creato un plot che mette a confronto la serie storica delle vendite (parte superiore Figura 6) e degli scontrini effettuati (parte inferiore Figura 6), notando che le serie storiche mostrano un comportamento praticamente analogo. L'esplorazione è proseguita con la creazione di un *qq-plot* (in Figura 7) che mostra sull'asse delle ascisse il valore delle vendite e sulle ordinate quello degli scontrini. Il risultato ottenuto porta alla conclusione che vi è un'alta correlazione tra le due serie storiche, in quanto le osservazioni si dispongono quasi a formare una retta.

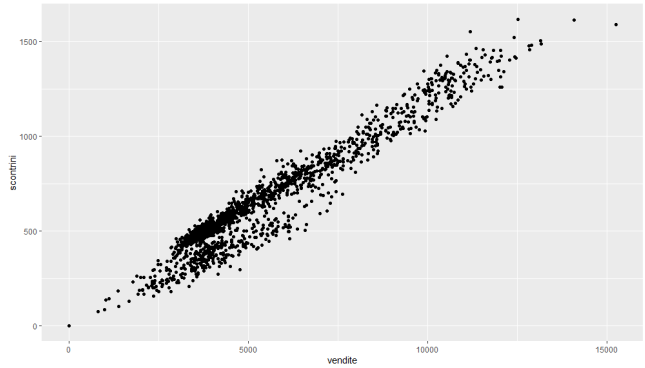

Figura 7. Correlazione tra vendite e scontrini ristorante 1

4.1.2 Decomposizione

Successivamente all'esplorazione si è proceduto con la decomposizione delle serie storiche, considerando i dati relativi alle vendite con granularità giornaliera. Questa operazione è stata necessaria in quanto le serie storiche hanno diversi pattern al loro interno e grazie all'utilizzo di particolari modelli è possibile dividere la serie storica nelle diverse componenti che la costituiscono. Nel caso preso in esame le componenti risultano essere trend, stagionalità e residui. All'interno di questo progetto la decomposizione è stata effettuata per migliorare la comprensione della serie storica esplorata andando a studiare i diversi pattern che la compongono.

All'interno dei dati a granularità giornaliera sono presenti due stagionalità a periodo differente: una prima a livello settimanale, quindi con periodo 7 giorni, e una seconda con periodo 365 giorni, quindi annuale.

Per effettuare la decomposizione è stato utilizzato il metodo *STL* (in questo caso *MSTL*, in quanto le stagionalità sono multiple), applicato tramite la libreria *stats*, che risulta essere l'acronimo di "Seasonal and Trend decomposition using Loess". Il seguente metodo permette di manipolare qualsiasi tipo di stagionalità, comprese le stagionalità che variano nel tempo.

Osservando il risultato della decomposizione si può notare il trend che è in leggera crescita nel periodo precedente al Covid. In corrispondenza del periodo di chiusure, invece, si nota una netta inversione di tendenza. Tuttavia, si osserva una minima crescita del trend nei mesi successivi al periodo di lockdown, dovuta ad una progressiva riapertura delle attività. Considerando invece la prima stagionalità, quella settimanale, è piuttosto evidente la differenza di valori tra giorni di settimana e di weekend, dove le vendite già dal venerdì aumentano di molto rispetto ai giorni lavorativi. Passando invece alla stagionalità annuale si confermano le intuizioni avute in fase di esplorazione, dove si è rilevato un aumento delle vendite nella parte finale dell'anno e in corrispondenza di alcuni mesi come aprile.

Osservando infine i residui si può notare che in corrispondenza delle chiusure dovute al Covid-19 sono presenti i residui più importanti: questo si può spiegare dal modo improvviso e imprevedibile con cui è stato necessario applicare tali restrizioni.

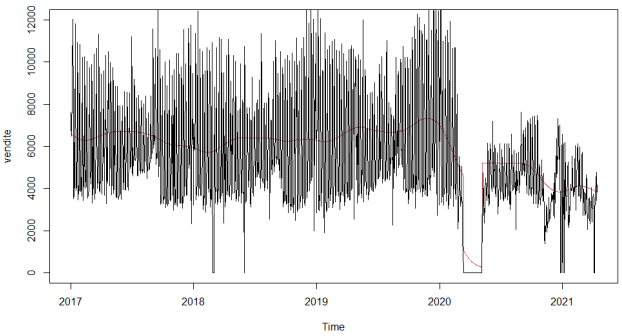

Figura 9. Trend vendite giornaliere ristorante 1

Dopo aver eseguito una prima decomposizione si è voluto studiare in maniera più approfondita il trend contenuto nella serie storica, in particolare nel periodo relativo al Covid-19.

Per effettuare ulteriori studi è stata utilizzata la libreria *KFAS* che ha permesso di creare un modello che avesse come regressori trend, stagionalità e residui. Oltre a ciò l'utilizzo di tale modello permette di prendere in considerazione alcune date di riferimento, in corrispondenza delle quali si sono verificati punti di shock: in particolare le date di chiusura e riapertura della prima ondata di Covid-19. Dopo aver effettuato il fit dei dati è stato generato un plot (Figura 9) che ha permesso di mostrare il trend delle vendite a granularità giornaliera.

Considerando la linea rossa che rappresenta il trend delle vendite è possibile notare che nonostante il periodo di chiusura della prima metà del 2020 dopo la riapertura dei locali si osserva un trend che riparte più o meno dallo stesso livello di quando i ristoranti sono stati costretti a chiudere. Inoltre il modello ha confermato ciò che già era stato osservato considerando la stagionalità, in particolar modo l'importanza del weekend dal punto di vista delle vendite. Concludendo si può osservare che pur proseguendo su un livello di vendite inferiore rispetto al periodo precedente al Covid-19 nel periodo successivo alle riaperture è diminuita la varianza, con le vendite che

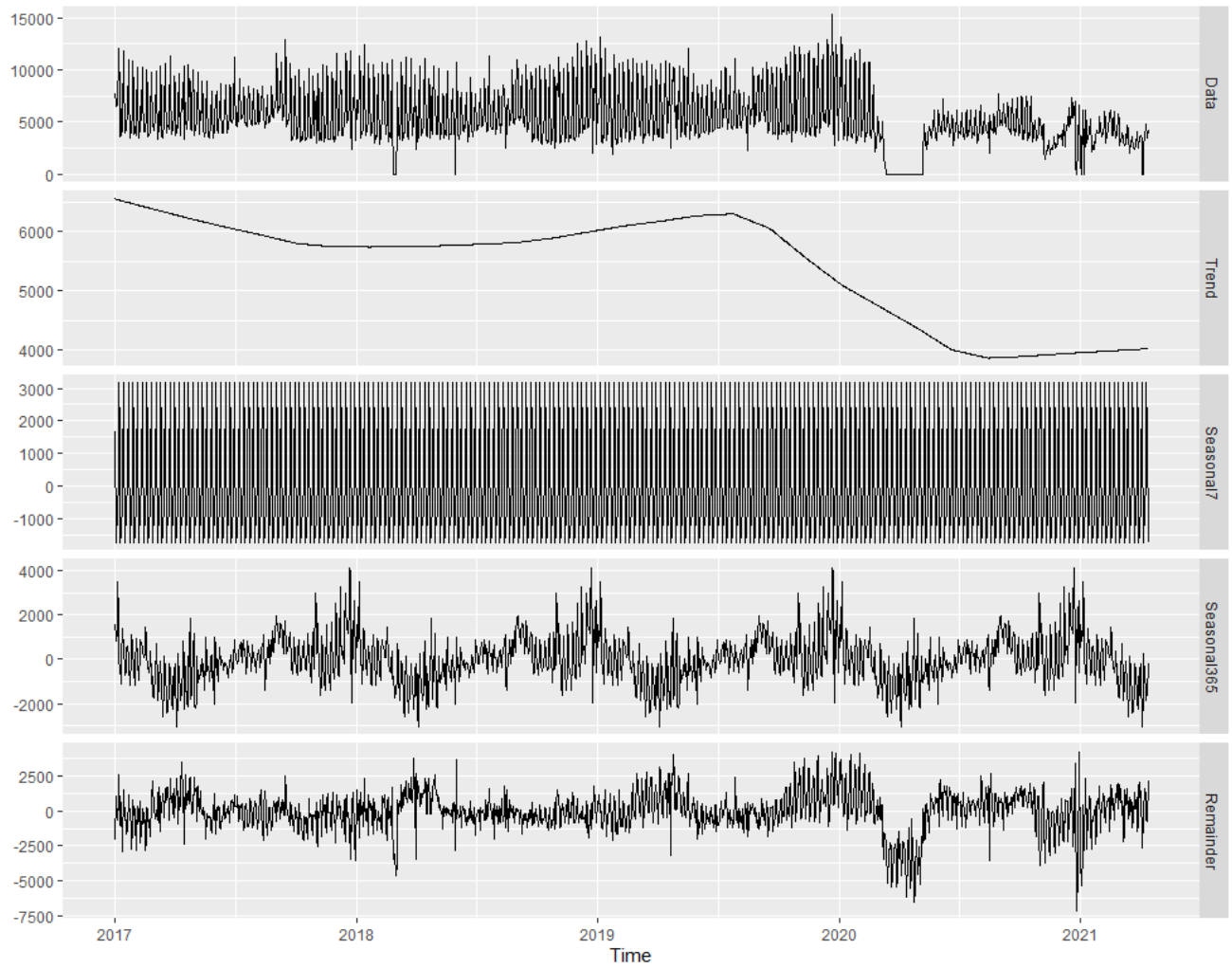

Figura 8. Decomposizione dati giornalieri ristorante 1

si concentrano tra i 2.000 euro e i 6.000 euro.

4.2 Altri ristoranti

Considerando anche i dati relativi ai cinque restanti esercizi di ristorazione è possibile osservare che le vendite sono calate per poi azzerarsi durante il periodo di lockdown dovuto alla situazione pandemica. Prendendo in analisi i singoli ristoranti si può constatare che le vendite per il terzo e il sesto ristorante non hanno come data di riferimento iniziale, come lo è per gli altri ristoranti, il 1 gennaio 2017. Quest'assenza di dati può essere motivata dalla mancata registrazione dei dati di vendita e scontrini o da un inizio di operatività dell'esercizio posticipato. Nonostante ciò, l'andamento dei sei ristoranti è simile. Si può osservare che le vendite per i ristoranti presi in considerazione negli anni a disposizione registrano chiusure in corrispondenza delle principali feste: Pasqua, Ferragosto e Natale. Si nota per i ristoranti in questione un calo del fatturato nei mesi di agosto mentre si registra quasi sempre un aumento nei mesi finali degli anni.
In seguito al periodo di chiusure, osservando i dati relativi alle vendite mensili medie, i ristoranti che hanno registrato la migliore ripesa sono stati il quarto e il sesto mentre quelli che hanno subito maggiormente il peso delle restrizioni in termini di fatturato sono stati il primo ristorante, il secondo e il quinto. Per quanto riguarda il terzo ristorante i dati a disposizione non risultano essere sufficienti per poter definire con una certa confidenza il tipo di ripresa. Tuttavia, osservando i dati con granularità settimanale, si potrebbe azzardare ad affermare che le vendite del terzo ristorante hanno risentito in maniera meno pesante delle chiusure rispetto ad altri ristoranti.

4.3 Confronto estati

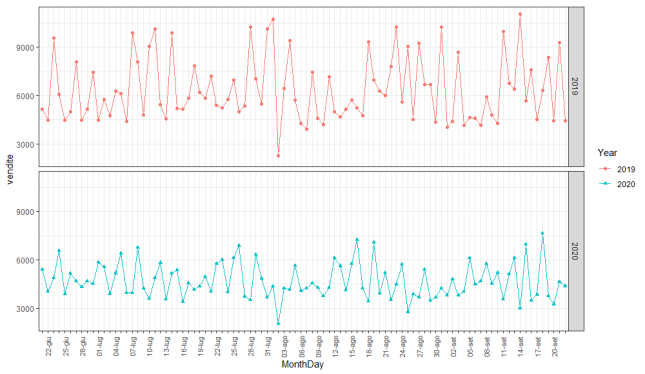

Figura 10. Confronto estate 2019 e 2020 ristorante 1

Il seguente paragrafo è destinato ad approfondire la seconda domanda di ricerca di questo studio, ovvero comprendere se ci siano state sostanziali differenze tra le vendite registrate nell'estate del 2020 e quelle dell'estate precedente. Il motivo per cui si è voluto indagare questo aspetto è da attribuirsi al fatto che l'estate del 2020 è stata la prima a risentire dell'effetto Covid: questo ha permesso di effettuare un confronto diretto tra la situazione pre e post pandemia, cercando di valutare i risultati ottenuti alla luce sia dell'allentamento delle misure restrittive che all'instaurarsi di un clima di incertezza nei consumatori che, inevitabilmente, ha comportato una contrazione delle vendite. In particolare, per il primo ristorante, è stato possibile osservare che per tutta la durata del periodo estivo (21 Giugno – 22 Settembre) la media delle vendite giornaliere registrate nel 2019 si aggirava intorno a 6.500 euro, con un solo calo registrato tra fine luglio ed inizio agosto, in cui le vendite sono scese sotto i 3.000 euro. Nell'estate 2020, come ci si aspettava, la media delle vendite risulta significativamente inferiore, pari a circa 4.500 euro. Inoltre, come è possibile osservare dal grafico riportato in Figura 10, nell'estate 2020 è stato raramente superato il livello dei 6.000 euro giornalieri, mentre nell'anno precedente si erano spesso registrate vendite oltre i 9.000 euro. Gli allentamenti delle misure restrittive, sebbene abbiano contribuito a risanare parzialmente i danni subiti, non sono quindi riusciti a ripristinare i volumi di vendita originari.

4.4 Trend scontrini

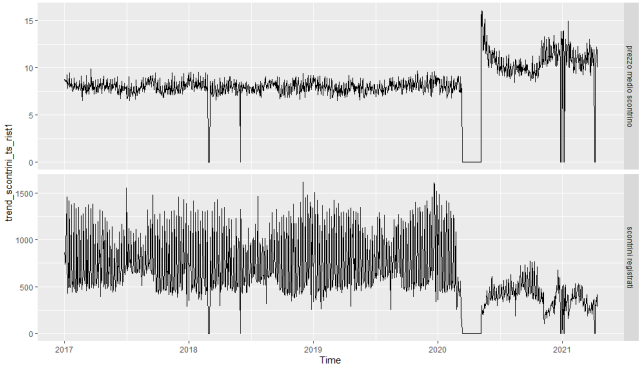

Figura 11. Focus scontrini ristorante 1

Prima di procedere con la creazione di modelli per poi dedicarsi alla parte relativa alle previsioni, sono stati analizzati i dati riguardanti gli scontrini. In particolare, è stato preso in esame il prezzo medio giornaliero dello scontrino e, a sostegno di tale analisi, è stata considerata la serie storica rappresentante la quantità di scontrini effettuati giornalmente.
Partendo dal periodo pre Covid-19 il prezzo medio del singolo scontrino è sempre stato intorno agli 8 euro in media. Tale osservazione potrebbe portare alla conclusione che il primo esercizio possa essere un bar, magari con servizio di tavola calda.
Un ulteriore osservazione riguardo il prezzo medio degli scontrini è che dopo le riaperture (eccetto alcuni giorni nei quali il fatturato è stato pari a zero a causa di ulteriori chiusure forzate) tale valore è aumentato, arrivando a superare i 10 euro. Si è provato a dare una spiegazione a questo fenomeno ipotizzando un aumento dei prezzi nel listino per far fronte alla crisi causata dalle chiusure o in alternativa pensando ad un maggior supporto dei clienti abituali, visto il periodo di difficoltà per l'attività.
Come si nota nella Figura 11 seppur il prezzo medio degli scontrini abbia registrato un incremento, le vendite (in termini di scontrini effettuati) sono diminuite con la pandemia. Questo ha portato dunque ad una perdita in termini di fatturato per il primo ristorante.
Si è scelto di studiare anche in questo caso il primo ristorante perché centrale in tutte le analisi effettuate; tuttavia, anche gli altri ristoranti si sono comportati in maniera simile.

4.5 Previsioni

La parte conclusiva e più corposa del progetto riguarda le previsioni. Si è deciso di implementare, prendendo

in considerazione il primo ristorante, diversi modelli destinati alla previsione per due differenti periodi:

- Il periodo Covid-19
- Il periodo successivo ad aprile 2021

Le tipologie di modelli utilizzati in questa sezione, come già accennato in precedenza, sono:

- *SARIMA*, *SARIMAX*
- *Random Forest*
- *Prophet*
- *TBATS*
- *HoltWinters*

4.5.1 Previsione periodo Covid-19

In questa sezione si cerca di rispondere ad una delle domande di ricerca: come sarebbero andate le vendite del primo ristorante se non ci fosse stata la pandemia che ha forzato le chiusure nell'ambito della ristorazione.
Per cercare di individuare delle risposte sono stati creati 4 diversi modelli su dati giornalieri e settimanali (sono stati utilizzati dati settimanali solo per i modelli della famiglia *SARIMA* in quanto quelli giornalieri presentavano una doppia stagionalità, non trattabile altrimenti con modelli di questa famiglia).
I dati utilizzati per addestrare i modelli sono quelli relativi alle vendite del primo ristorante: in particolare è stato considerato il periodo compreso tra l'1 gennaio 2017 e il 5 gennaio 2020. Tale scelta è stata motivata dal fatto di voler trattare dati che non fossero influenzati in maniera evidente dalla presenza del Covid. Inoltre, per avere un maggior riscontro delle performance dei modelli, si è proceduto dividendo il dataset in training (70%) e test (30%) set come raffigurato nella Figura 12. Le previsioni sui dati di vendita sono state calcolate fino al giorno 17 maggio 2020, giorno precedente a quello delle riaperture.

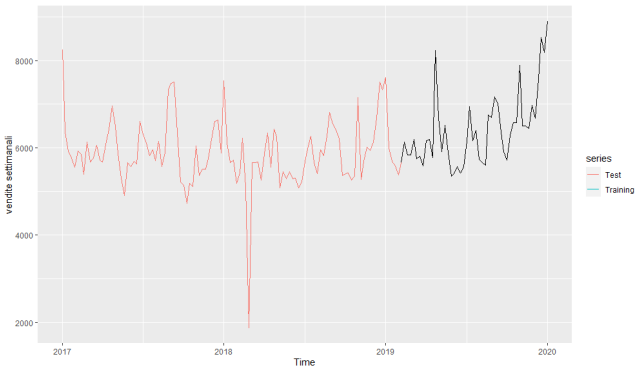

Figura 12. Divisione dataset in training e test

SARIMA I primi modelli utilizzati sono stati i modelli *SARIMA*. Dopo aver aggregato i dati giornalieri in dati settimanali e aver applicato su di essi l'operatore *mean*, ottenendo quindi il valore medio giornaliero delle vendite a livello settimanale, si è iniziato provando a ricavare manualmente i parametri del modello, consultando i grafici delle correlazioni *ACF* e *PACF*.
Il primo passo effettuato è stato quello di rendere la serie stazionaria e per farlo è necessario eliminare stagionalità e trend attraverso la destagionalizzazione e differenziazione. É stata dapprima rimossa la stagionalità con *STL* (nel dettaglio è stata usata *seasad*), che restituisce dati aggiustati stagionalmente, costruiti rimuovendo la componente stagionale. In seguito, sono state utilizzate le funzioni *ndiff* e *ndiffs* per stimare il numero di differenze richieste per rendere la serie stazionaria: *ndiff* ha restituito il valore 1. Si è quindi provveduto a rimuovere la presenza del trend con la differenziazione. Per verificare l'effettiva stazionarietà sono stati utilizzati due test statistici: il test *ADF* (Dickey-Fuller aumentato) e il test *KPSS*. Entrambi i test hanno confermato la stazionarietà della serie in questione. A seguire si è provveduto a dividere i dati in training e test set così da modellare i dati a disposizione e valutare la qualità del modello così ottenuto. Per quanto riguarda l'implementazione del modello, i valori delle componenti p, q e d sono risultate essere (2,0,0) mentre per quanto riguarda la componente stagionale sono stati ricavati i valori (0,0,1) con stagionalità di 52 settimane (annuale).

Successivamente si è provato ad utilizzare la funzione *auto.arima* in modo tale da delegare la ricerca dei parametri e l'operazione che porta la serie storica ad essere stazionaria. Inoltre questa funzione sceglie il modello migliore tra diversi modelli tramite la misura di *AIC*: il modello che verrà restituito sarà quello con *AIC*

minore. Il modello trovato anche in questo caso è stato un *SARIMA*, con stagionalità annuale e in particolare le componenti risultano essere (0,1,1) mentre le componenti stagionali (0,1,0). Il modello travato con *auto.arima* presenta un *AIC* minore rispetto al modello identificato manualmente precedentemente, che è stato dunque scartato. Inoltre anche la misura *MAPE* è risultata migliore nel secondo tentativo.

In aggiunta, per avere la certezza che il modello selezionato spieghi in buona misura i dati sono stati controllati i residui: tale operazione è dovuta al fatto che se nei residui è presente correlazione ciò è sintomo che dell'informazione è ancora presente nei residui e dunque il modello non è in grado di spiegare in modo corretto i dati. Guardando la Figura 13 si nota che la correlazione non è significativa per nessun lag, dunque i residui non sono correlati. Per un'ulteriore conferma si è proceduto ad eseguire il *Ljung-Box* test con risultato *p-value* = 0,352: tale risultato non permette di rifiutare l'ipotesi nulla di incorrelazione tra i dati. Nella stessa Figura 13 si può affermare osservando il grafico in basso a destra che i residui seppur abbiano una coda molto più allungata presentano una distribuzione normale.

Nella Figura 14 sono riportati i risultati della previsione: dal grafico si nota che se non ci fosse stata la pandemia e le complicazioni conseguenti il trend delle vendite sarebbe rimasto sullo stesso livello dell'anno precedente. Le vendite dei primi mesi dell'anno 2020 sarebbero rimaste intorno ai 6.000 euro di media giornaliera. Si può inoltre osservare che il picco negativo di vendite ad inizio marzo verificatosi a marzo 2018 si ripete anche nella previsione degli anni 2019 e 2020: questo perché le previsioni del modello *SARIMA* sono basate sui dati già esistenti che precedono il periodo. In questo caso specifico il "problema" nasce dalla mancanza di dati, in quanto utilizzando un periodo di circa due anni per la parte di training del modello e differenziando con stagionalità 52 si va a perdere tutto il primo anno. Il modello è quindi addestrato su un solo anno dove ad inizio marzo il ristorante è chiuso per motivi meteorologici.
Per risolvere questo problema si potrebbe pensare di aggiungere un regressore esterno, come ad esempio il meteo, oppure cercare un modello che permette di non dover differenziare e quindi di non dovere "perdere" un anno di dati.

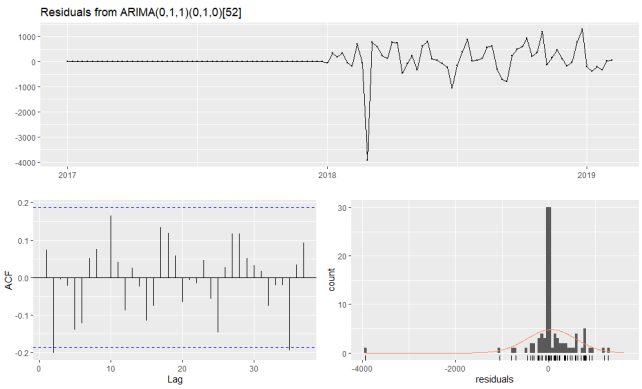

Figura 13. Residui modello *SARIMA(0,1,1)(0,1,0)[52]*

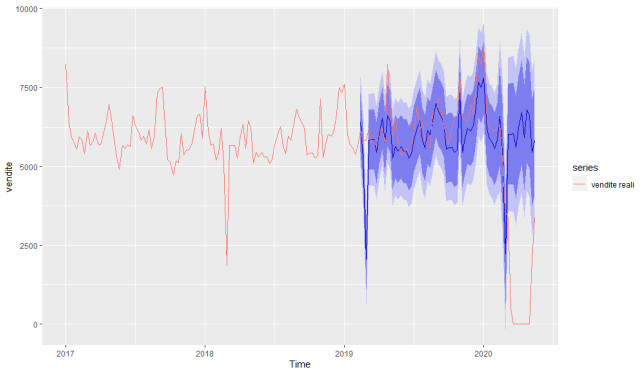

Figura 14. Previsione *SARIMA* periodo Covid

Il modello finale trovato dunque è un *SARIMA (0,1,1)(0,1,0)[52]* che presenta un *AIC* di 918. Come misure di accuratezza per appurare la bontà del modello sono state utilizzate il *RMSE* e il *MAPE*, con i rispettivi valori di 520,959 e 5,283 sul training set e 797,846 e 8,146 sul test set.

Random Forest Per effettuare una previsione a livello giornaliero è stato implementato il modello *Random Forest*. Dopo aver provveduto a dividere i dati in training e test set sono state selezionate le variabili più rilevanti attraverso la funzione *varImpPlot()* ai fini dell'implementazione del modello. In particolar modo, come è possibile osservare in Figura 15, è emerso che le variabili più importanti sono: *weekday*, *is_weekend*, *mese*, *is_holiday*, *stagione*. A seguire, utilizzando la funzione *randomForest* è stato implementato il modello utilizzando come regressori le variabili precedentemente indicate.

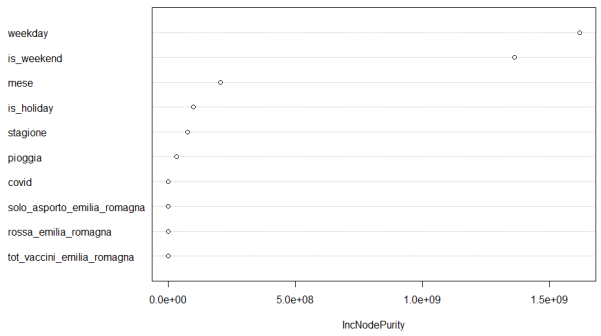

Figura 15. Variable Importance Plot

Le metriche scelte per valutare la qualità del modello ottenuto sono state il *RMSE* e il *MAPE*, che hanno assunto rispettivamente i valori 1.333,147 e 17,539. Osservando le previsioni ottenute in Figura 16 si può constatare che le vendite si attestano tra 5.000 euro e 10.000 euro, con una minore variabilità rispetto ai dati precedenti.

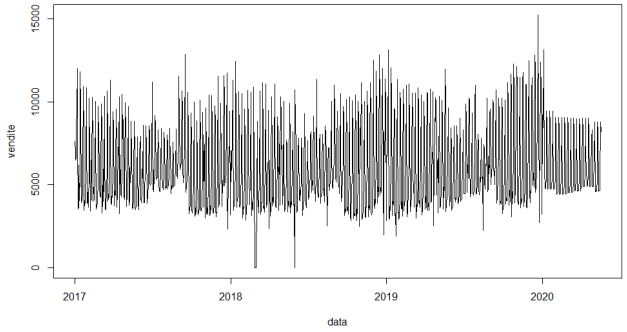

Figura 16. Previsione *Random Forest* periodo Covid

Prophet Dopo aver predisposto i dati relativi alle vendite giornaliere divise in training e test set viene utilizzata la funzione *prophet* per identificare il modello che meglio si adatta ai dati in questione. Per valutare l'adattamento del modello ai dati sono state utilizzate come metriche il *RMSE* e il *MAPE*. I valori ottenuti sono rispettivamente pari a 1.353,581 e 13,961. Successivamente, una volta implementato il modello, quest'ultimo viene utilizzato per generare le previsioni delle vendite che si sarebbero registrate durante il periodo di lockdown, registrando valori nel range 5.000euro-10.000euro, in termini di fatturato.

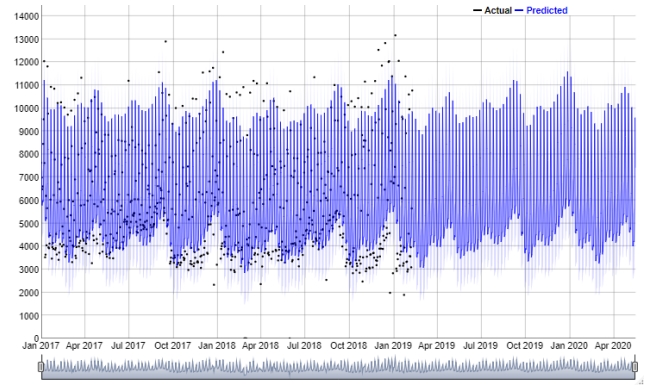

Figura 17. Previsione *Prophet* periodo Covid

TBATS Anche per il seguente modello i dati delle vendite vengono considerati con granularità giornaliera. Dopo aver diviso i dati a disposizione in training e test set viene implementato il modello utilizzando la funzione *tbats*. Per valutare la qualità del modello sono state considerate le metriche *RMSE* e *MAPE*, che hanno assunto rispettivamente i valori 1.431,843 e 14,835. Successivamente si è proceduto in modo analogo ai precedenti modelli e sono state generate le previsioni fino a circa metà maggio i cui valori, come osservabile in Figura 18, si aggirano mediamente attorno al valore di 5.000euro, registrando un'elevata variabilità.

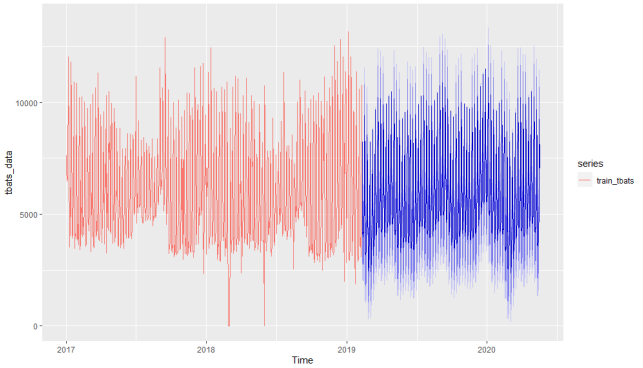

Figura 18. Previsione *TBATS* periodo Covid

4.5.2 Previsione futuro

L'ultima domanda di ricerca all'interno del progetto consiste nella previsione delle vendite nel periodo successivo ai dati disponibili. In particolare le previsioni sono state fatte nel periodo compreso tra il 12 aprile 2021 e il 12 agosto 2021.
Per l'implementazione dei modelli, in questo caso, è stato deciso di non adoperare la tradizionale distinzione del dataset in training e test set, addestrando il modello sull'intera serie storica a disposizione. Il motivo di ta-

le decisione è da attribuirsi alla volontà di catturare lo shock dovuto al Covid e ottenere in tal modo un miglior adattamento ai dati e una migliore accuratezza nelle previsioni.
Le tipologie dei modelli utilizzati per le previsioni sono i medesimi della sezione precedente, con due eccezioni: l'aggiunta di un modello a lisciamento esponenziale quale *HoltWinters* e l'utilizzo di *SARIMA* con dei regressori esterni che hanno permesso di spiegare in modo più accurato l'andamento relativo al Covid-19.

HoltWinters Il modello *HoltWinters* viene addestrato su tutti i dati con granularità settimanale a disposizione. Il modello in questione permette di catturare sia il trend che la stagionalità presenti nella serie storica. Come osservato nel paragrafo relativo agli aspetti metodologici il modello *HoltWinters* essendo una versione più efficiente rispetto alla decomposizione tramite media mobile procede nell'effettuare una trasformazione della media mobile in una media esponenziale con uno o più parametri di smorzamento, definiti come α, β e γ. Un valore α elevato favorisce la parte innovativa, in questo caso tale parametro risulta essere pari a 0.693. Il parametro β pari a zero, invece, è indice di un trend perfettamente conservativo, dunque lineare. Il parametro γ, relativo alla componente della stagionalità, è pari ad 1.
Osservando la Figura 19 si può dire che le previsioni si mantengono intorno al valore di 3.000euro, rispecchiando un livello leggermente inferiore rispetto all'anno precedente.

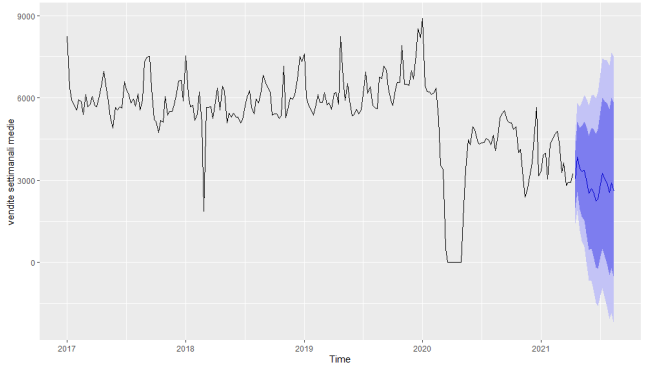

Figura 19. Previsione *HoltWinters* post aprile 2021

SARIMAX Anche per le previsioni riguardanti il periodo successivo ai dati disponibili si è deciso di implementare un modello *SARIMA* su dati a granularità settimanale. Tale scelta riguardo il tipo di granularità si basa sulle stesse considerazione fatte nella sezione relativa alle previsioni del periodo Covid-19. In questo caso però insieme al modello *SARIMA* sono stati considerati alcuni regressori esterni, in modo tale da modellare in modo migliore lo "shock" provocato dalle chiusure per il Covid-19. Si parla dunque di *SARIMAX*.
I regressori utilizzati sono stati due e di tipo dummy: in questo modo è stato possibile non aumentare in modo eccessivo la complessità del modello creato. Le variabili in questione sono:

- *week covid bin*: come detto in precedenza si tratta di una variabile dummy e segnala la presenza del Covid o meno nel corso della settimana. La variabile è stata ottenuta integrando i dati giornalieri: in particolare, se vi sono quattro o più giorni della settimana nei quali viene identificato il Covid il suo valore è 1, al contrario il valore è 0.
- *week rossa bin*: anche in questo caso la variabile è dummy e porta un'informazione aggiuntiva riguardo il colore delle zone. In particolare, è stata ottenuta integrando settimanalmente i dati giornalieri relativi alle colore delle zone. Se in una settimana vi sono quattro o più giorni di zona rossa la variabile è pari a 1, altrimenti è 0.

Questi regressori sono stati utilizzati dopo aver testato diversi modelli con diverse variabili.
Il modello è stato ottenuto tramite la funzione *auto.arima* e in particolare aggiungendo i regressori tramite il parametro *xreg*. Il modello risultante risulta essere un *SARIMAX (1,0,0) (0,0,1)[52]*. Successivamente è stata testata la significatività dei regressori considerati tramite *F-test* che ha riportato valori di *p-value* molto bassi in grado di far scartare l'ipotesi nulla di non significatività delle variabili.

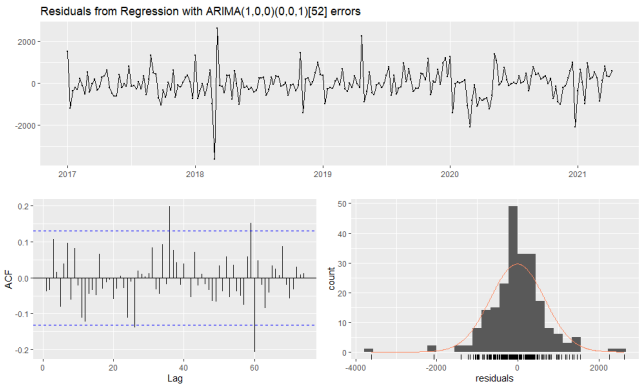

Figura 20. Check residuals SARIMAX

Non avendo utilizzato la suddivisione in training e test set come detto in precedenza, le uniche misure di performance si hanno sui dati di training utilizzati per l'addestramento del modello. Quest'ultime risultano essere pari a 667.428 per quanto riguarda il *RMSE* e *inf* per quanto riguarda il *MAPE*. Si è ottenuto un valore infinito per l'indice *MAPE* dal momento in cui sono presenti alcuni dati che hanno valore 0.
p-value pari a 0.149 che non permette dunque di rifiutare l'ipotesi nulla di incorrelazione dei residui. Il grafico in Figura 20 mostra come eccetto alcuni lag intorno al 30esimo non vi sia autocorrelazione tra i residui.

Successivamente si è proceduto analizzando l'autocorrelazione degli errori. Il risultato del test effettuato, in particolare il *Ljung-Box* test, ha restituito un Considerando invece i regressori, la presenza del Covid influenza il modello implementato con un valore di -2012.012 per quanto riguarda il valore delle vendite, mentre se il colore della zona durante la settimana è rosso il valore delle vendite scende di 1.522,909.
Dopo aver creato il modello si è proceduto alla parte riguardante la previsione che ha interessato il periodo che termina con la settimana comprendente il giorno 12 agosto 2021. Come si può notare dalla Figura 21, tra l'ultimo valore appartenente ai dati a disposizione e il primo appartenente alla previsione vi è una netta differenza: questa si può giustificare con il cambio di colore della zona infatti la prima settimana di previsione corrisponde con l'uscita dalla zona rossa per la regione dell'Emilia-Romagna. Inoltre, si può osservare che il livello previsto per l'estate 2021 registra un andamento simile a quello riscontrato durante l'estate precedente 2020.

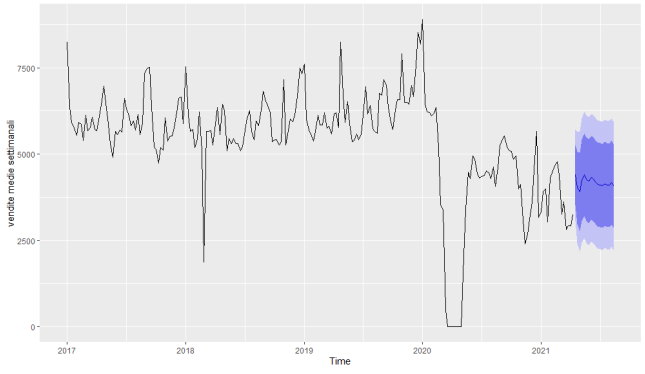

Figura 21. Previsione *SARIMAX* post aprile 2021

Per effettuare le previsioni è stato necessario considerare i regressori anche per le date future: in particolare sono stati utilizzati i dati reali ricavati dalle fonti precedentemente descritte nella Sezione 3, relativa ai dati.

Random Forest Avendo eseguito tramite i modelli *SARIMA* previsioni riguardanti una granularità settimanale sono stati scelti altri modelli che permettessero di svolgere previsioni su dati giornalieri.
Il primo modello utilizzato è il *Random Forest*. Dopo aver creato un modello con tutti i regressori a disposizione si è analizzato il relativo grafico *Variable Importance Plot* che ha permesso di selezionare i regressori più significativi, quali *weekday*, *is weekend*, *Covid*, *rossa emilia romagna* e *mese*.
Dopo aver utilizzato l'intero dataset riguardante le vendite del primo ristorante per effettuare la fase di training del modello si sono calcolate le previsioni fino al giorno 12 agosto 2021. Le previsioni, come nel caso precedente, seppur con granularità differente, indicano un livello di vendite pari a quelle dell'estate precedente, intorno ai 5.000 euro di vendite giornaliere.

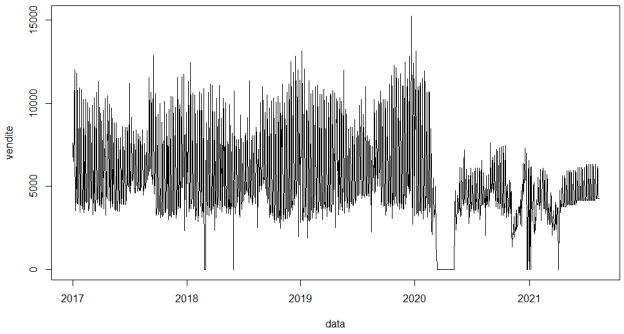

Figura 22. Previsione *Random Forest* post aprile 2021

Le misure di performance per questo modello considerando il *RMSE* e il *MAPE* risultano essere rispettivamente 1.537,093 e *Inf*.

Prophet Il modello in questione è stato creato su dati giornalieri e fa utilizzo di regressori. In particolare con la funzione *add_country_holidays* del package *Prophet* sono state considerate le vacanze. I restanti regressori risultano essere *rossa*, variabile dummy che indica se il colore della zona di quel determinato giorno è rosso o meno e *Covid*, anch'essa variabile dummy che indica se la data è successiva al 9 marzo 2020.
La Figura 23 riporta le previsioni ottenute, che cadono all'interno della stessa fascia dei dati riguardanti l'anno precedente, con valori medi di 6.000euro.

Le misure di performance *RMSE* e *MAPE* risultano essere rispettivamente 1.344,278 e *Inf*

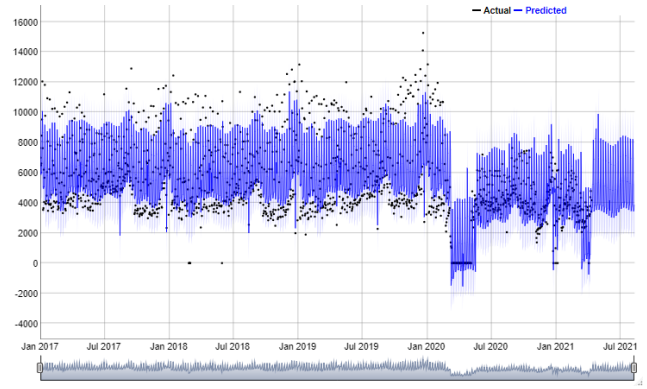

Figura 23. Previsione *Prophet* post aprile 2021

.

TBATS Il seguente modello si basa anch'esso su dati con granularità giornaliere e non viene effettuata alcuna divisione in training e test set. Il modello per quanto riguarda le performance ha un *RMSE* di 1.144,543, mentre il *MAPE* è *Inf*. Nella Figura 24 è possibile osservare le previsioni che sono più ottimistiche rispetto ai modelli precedenti, denotando un netto trend crescente con valori superiori ai 7.000 euro.

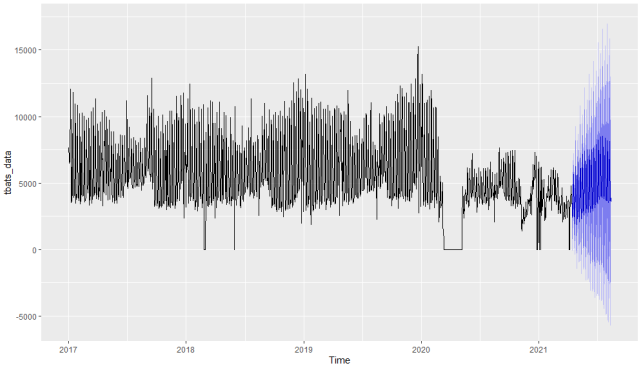

Figura 24. Previsione *TBATS* post aprile 2021

5. Risultati

Dopo aver implementato i diversi modelli sopra riportati, facendo variare opportuni parametri, si è proceduto ad una fase di valutazione al fine di trovare i modelli che meglio si sono adattati ai contesti presi in esame. Sono state confrontate le misure di performance ottenute per ogni modello, tenendo conto però della granularità dei dati su cui questi sono stati creati: da un lato i modelli della famiglia *SARIMA/SARIMAX* e *HoltWinters* basati su dati settimanali, dall'altro i restanti modelli con granularità giornaliera, *TBATS*, *Random Forest* e *Prophet*.

Le misure di performance utilizzate per il confronto sono quelle esposte nella Sezione 2. Si è deciso di considerare come metrica principale il *RMSE*: tale decisione è dovuta dalla presenza di molti valori pari a zero all'interno dei dati riguardanti le vendite che non hanno permesso di utilizzare l'indice *MAPE*.
Si riportano nella Sezione Appendice le tabelle con i relativi risultati ottenuti.

Considerando i modelli riguardanti la previsione effettuata sul periodo Covid i modelli che hanno registrato le migliori performance, rispetto all'indice *RMSE*, sono il *SARIMA* a livello settimanale ($RMSE = 797,846$) mentre a livello giornaliero il modello migliore è risultato essere *Random Forest* ($RMSE = 1.333,147$).
Osservando le previsioni ottenute con il modello *SARIMA* si nota che se non ci fosse stata la pandemia che ha costretto il ristorante alla chiusura, il livello delle vendite sarebbe rimasto intorno ai 6.000 euro: le vendite avrebbero avuto dunque un andamento simile ai mesi precedenti alla pandemia.
Considerando invece i modelli sviluppati sui dati a granularità giornaliera, si nota come il miglior risultato è ottenuto dal *Random Forest* seguito da *Prophet* che ottiene un valore leggermente peggiore. Osservando le previsioni generate dal modello *Random Forest* vengono confermati i risultati del modello *SARIMA* già esposti precedentemente: le vendite si attestano intorno al valore di 6.000 euro Inoltre si può osservare che la varianza della previsione è minore rispetto al periodo precedente.

Passando ai modelli previsionali per il periodo successivo ad aprile 2021, il miglior modello settimanale risulta essere il modello *SARIMAX* ($RMSE = 667,428$). Il modello, invece, che si è meglio comportato riguardo ai dati giornalieri è il *TBATS* ($RMSE = 1.144,543$).
Le previsioni del modello *SARIMAX* hanno evidenziato un iniziale cambio di livello nelle vendite dovuto al cambio di colore della zona, da rossa ad arancione. Successivamente le vendite si sono stabilizzate a valori inferiori a 5.000 euro, valori che rispecchiano il periodo iniziale delle riaperture dopo la lunga chiusura.
Dal punto di vista dei dati giornalieri, le previsioni ottenute dal modello *TBATS* sembrano indicare un aumento delle vendite, aspetto riscontrato anche nel *SARIMAX*, anche se in maniera più brusca.

6. Conclusione e possibili sviluppi

Le analisi svolte hanno permesso di rispondere in maniera esaustiva alle domande di ricerca poste inizialmente. In particolare, è stato possibile osservare che:

- *Nel corso degli anni, pandemia esclusa, le vendite dei sei ristoranti considerati hanno registrato chiusure in corrispondenza dei principali periodi festivi, quali Pasqua, Ferragosto e Natale. Durante il mese di agosto sono spesso stati registrati ingenti cali di fatturato, probabilmente dovuti alle partenze tipiche del periodo estivo. Per quanto riguarda i dati relativi al periodo di riaperture post-lockdown, invece, è stato possibile osservare alcune differenze tra i ristoranti considerati: il quarto e il sesto ristorante hanno infatti registrato una migliore ripresa, mentre il primo, il secondo e il quinto hanno subito cali maggiori di fatturato.*
- *Ci sono state sostanziali differenze tra la prima estate che ha risentito degli effetti del Covid-19, nonostante le riaperture concesse, e quella dell'anno precedente (2019). In particolare, per il primo ristorante la media totale delle vendite del periodo estivo è risultata inferiore di circa 2.000 euro, e i livelli di fatturato raggiunti in periodo pre-Covid, che facevano toccare anche i 9.000 euro giornalieri, si sono abbassati sensibilmente, facendo saltuariamente raggiungere un massimo di 6.000 euro al giorno. Per gli altri ristoranti si è registrato un andamento molto simile: gli allentamenti delle misure restrittive, sebbene abbiano contribuito a risanare parzialmente i danni subiti, non sono quindi riusciti a ripristinare i volumi di vendita originari*
- *Il prezzo medio di ogni scontrino del primo ristorante durante il periodo pre-Covid è sempre stato pari a circa 8 euro. Successivamente alle riaperture, invece, il prezzo medio è aumentato, arrivando a superare anche i 10 euro. Probabilmente questo fenomeno è da attribuirsi al tentativo di applicare dei rincari per far fronte alla crisi causata dal periodo di chiusure. Tuttavia, il numero di scontrini ha subito un decremento: probabilmente si tratta di una delle cause che hanno impedito di risanare completamente le perdite di fatturato. Anche in questo caso, il comportamento degli altri ristoranti è stato simile.*
- *Al fine di comprendere come sarebbero state le vendite durante il periodo Covid e come saranno le vendite durante il periodo successivo ad aprile 2021 sono stati applicati diversi modelli di previsione, i cui risultati sono stati esposti nella Sezione 5 relativa ai risultati.*

I possibili sviluppi futuri potrebbero mirare ad ottenere maggiori informazioni riguardo agli esercizi di ristorazione in questione, in quanto quelle a disposizione sono molto poche e generali. Maggiori e più precise informazioni, quali la precisa ubicazione e la tipologia specifica dell'esercizio, potrebbero infatti permettere di definire features più accurate e corrette, con il fine di aumentare la performance dei modelli.
Nella realizzazione del progetto l'attenzione è stata posta principalmente sul primo ristorante: si potrebbe perciò pensare di estendere le analisi di esplorazione e previsione in modo dettagliato anche per i restanti ristoranti.
Si potrebbe anche seppur avendo un numero esiguo di ristoranti effettuare una clustering analysis per cercare di raggruppare tra loro ristoranti con caratteristiche simili.

Riferimenti bibliografici

[1] Hyndman R.J. & Athanasopoulos G. *Forecasting: principles and practice, 2nd edition*. https://otexts.com/fpp2/, 2018.

[2] Fattore M. *Fundamentals of time series analysis, for the working data scientist*, 2020.

[3] Reddy B. *A Guide to Forecasting Demand in the Times of COVID-19*, 2020.

[4] Chatterjee S. *Time Series Analysis Using ARIMA Model In R*. https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/, 2018.

[5] Dalinina R. *Introduction to Forecasting with ARIMA in R*. https://blogs.oracle.com/ai-and-datascience/post/introduction-to-forecasting-with-arima-in-r, 2017.

[6] Singh Deepika. *Machine Learning for Time Series Data in R*. https://www.pluralsight.com/guides/machine-learning-for-time-series-data-in-r, 2020.

[7] Taylor SJ & Letham B. *Forecasting at scale*. https://peerj.com/preprints/3190/, 2017.

[8] Mattis B. *Time Series Forecasting in R with Holt-Winters*. https://towardsdatascience.c

om/time-series-forecasting-in-r-with-holt-winters-16ef9ebdb6c0, 2021.

[9] Hyndman R. J. *Time Series Forecasting in R with Holt-Winters*. https://robjhyndman.com/hyndsight/seasonal-periods/, 2014.

Appendice

Tabella 1. Periodo Covid-19

Modello	Frequenza dati	RMSE
SARIMA	settimanale	797,846
Random Forest	giornaliera	1.333,147
TBATS	giornaliera	1.431,843
Prophet	giornaliera	1.353,581

Tabella 2. Periodo successivo ad aprile 2021

Modello	Frequenza dati	RMSE
SARIMAX	settimanale	667,428
HoltWinters	settimanale	821,021
Random Forest	giornaliera	1.537,093
TBATS	giornaliera	1.144,543
Prophet	giornaliera	1.344,278