Covid e Ristorazione: un caso di studio

Gaetano Chiriaco,[1] Riccardo Porcedda,[1] Gianmarco Russo[1]

Sommario
Uno dei settori più colpiti dalla pandemia Covid-19 è stato sicuramente quello della ristorazione. A causa delle relative restrizioni infatti, i ristoratori hanno visto i propri introiti calare drasticamente. In tale periodo storico può risultare molto utile analizzare le serie storiche per cercare di studiare e prevedere quali saranno gli introiti della giornata o della settimana per regolare di conseguenza la fornitura delle materie prime. In questo elaborato è stato utilizzato dei modelli della famiglia ARIMA e il Cluster-Weighted Model con e senza cross-validation per la previsione delle serie storiche.

Parole chiave
Serie Storiche — Covid-19 — Ristorazione — Arima

[1] *Università degli Studi di Milano-Bicocca, Dipartimento di Informatica,Sistemistica e Comunicazione, CdLM Data Science*

Indice

1 Dati **2**
1.1 Dati sui ristoranti 2
1.2 Dati sui contagi ed il colore delle regioni 2
1.3 Dati meteorologici di Piacenza 2
2 Analisi esplorative e pre-processing **2**
3 Confronto dei trend post-lockdown **4**
4 Previsione del lordo totale **5**
4.1 Stima e previsioni con modello ARIMA 5
Suddivisione in Train/Test • Forecast Cross-Validation
4.2 Suddivisione in Feriali e Festivi 6
5 Uno sguardo sugli sviluppi futuri **6**
5.1 Cluster-Weighted Models 6
5.2 Inclusione del listino prezzi 7
6 Conclusioni **8**
7 Appendice **8**
7.1 Aspetti metodologici 8
Arima • Sarima • Funzione Vacanza • Cluster-Weighted Models

Introduzione

La ristorazione è stato senza dubbio uno dei settori economici più colpiti dall'emergenza sanitaria Covid-19. L'impossibilità di ogni contatto sociale ha portato alla chiusura totale o parziale dei principali luoghi di aggregazione come pub, bar e ristoranti. Ad oggi, grazie alla campagna vaccinale e all'utilizzo di DPI (dispositivi di protezione individuale) e al distanziamento sociale la maggior parte delle attività sono ripartite. Nonostante le riapertura, il modo di svolgere il proprio lavoro ha subito un processo di riadattamento continuo. Durante il secondo lockdown (DPCM del 6/11/2020) l'Italia ha deciso di regolare le restrizioni regionalmente, basandosi sui contagi locali e su codici colori (*bianco-giallo-arancione-rosso*). Ogni colore, assegnato sulla base di una serie di indicatori, specifica un insieme di regolamentazioni e restrizioni da seguire. Inoltre, vi sono stati aspetti legati alla protezione dal contagio che sono ricadute sui conti dei ristoranti, come mascherine, gel igienizzanti e sanificazione degli ambienti. Molti ristoratori sono stati costretti a ridurre il numero di posti a sedere per mantenere le distanze di sicurezza tra i tavoli, diminuendo effettivamente la capienza e il potenziale introito.

In questo studio, utilizzando i dati sugli incassi lordi di sei ristoranti della provincia di Piacenza ed i dati relativi alla diffusione del Covid-19 in Italia, è stata analizzata l'influenza dell'introduzione delle restrizioni per colore. Inoltre, utilizzando diverse metodologie, sono stati stimati modelli capaci di effettuare previsioni sugli incassi lordi, tenendo conto di variabili meteorologiche, delle restrizioni e dell'effetto delle vacanze e della stagionalità. Gli strumenti di previsione proposti permettono al proprietario del ristorante di avere un'idea di quelli che possono essere gli incassi attesi futuri ed organizzarsi di conseguenza.

Analisi proposte

Sono state portate a termine due differenti analisi studiando il rapporto tra il lordo totale, il numero di scontrini ed la situazione Covid-19 in Italia e nell'Emilia Romagna:

- Considerando solo il periodo post-lockdown (dopo il 2/6/2020), la serie storica riguardante il lordo dei sei ristoranti è stata decomposta. Una volta calcolati i trend, i vari ristoranti vengono confrontati per verificare se stanno vivendo un periodo di ripresa dopo la chiusura totale del 2020.

- Con un intento previsivo, sono stati stimati vari modelli capaci di cogliere l'andamento futuro degli incassi lordi dei ristoranti. Per ottenere previsioni più affidabili, i

dati a disposizione sono stati arricchiti con informazioni sul numero di nuovi contagi giornalieri nazionali e regionali e dati meteorologici. Le previsioni sono state effettuate solo sul Ristorante 3, situato a Piacenza. I risultati che seguono possono essere estesi anche agli altri ristoranti con l'opportuna modifica dei dati integrati relativi alla posizione di questi ultimi.

1. Dati

1.1 Dati sui ristoranti

I dati utilizzati nelle seguenti analisi fanno riferimento a sei ristoranti, tre dei quali situati a Piacenza (Ristorante 1, 2 e 3) e i rimanenti tre situati in province del nord-Italia (Ristorante 0 in provincia di Rimini e il Ristorante 4 e 5 in provincia di Pavia). La prima osservazione a disposizione è datata 1° Gennaio 2018, mentre l'ultimo dato è del 30 Aprile 2022. Le variabili a disposizione sono:

- **Id dei ristoranti**: numero identificativo del ristorante, da 0 a 5;
- **Data**: data a cui si riferiscono le informazioni (*formato: aaaa/mm/gg*);
- **Lordo totale**: incassi totali lordi nella giornata di riferimento;
- **Scontrini**: numero di scontrini effettuati nella giornata di riferimento;

Quindi, per ognuno dei sei ristoranti si hanno a disposizione due serie storiche giornaliere. Il numero totale di osservazioni pari a 8817.

1.2 Dati sui contagi ed il colore delle regioni

Per integrare nello studio le informazione relative la situazione Covid-19, i dati di partenza sono stati arricchiti con le informazioni sui nuovi contagi giornalieri ed i colori della regione di riferimento. Le variabili esplicative aggiunte sono state:

- **Contagi**: numero di nuovi contagi registrati il giorno precedente. Questa variabile esplicativa è stata provata in fase di previsione e stima ed utilizzata con un ritardo di un giorno, per verificare se le notizie sui contagi del giorno precedente influenzano il numero dei clienti nei ristoranti nei giorni successivi. è stato osservato in tutti i modelli analizzati che l'aggiunta di questa informazione aggiuntiva non ha migliorato le performance previsive dei modelli rispetto al solo utilizzo del colore delle regioni.
- **Colore**: colore della regione Emilia-Romagna per ogni giorno nel periodo analizzato (*bianco-giallo-arancione-rosso*). Nei giorni che precedono il 6 Novembre 2020 è stato assegnato il colore "bianco", siccome le regioni sono state divise in fasce di colori solo dopo il DPCM del 3 Novembre 2020.

1.3 Dati meteorologici di Piacenza

Per ottenere un modello più accurato, l'insieme dei regressori è stato integrato con i dati meteorologici della zona di Piacenza, dove sono situati il Ristorante 1, 2 e 3. I dati iniziali sono stati arricchiti con le variabili:

- **Heat Index**: misura utilizzata per esprimere la temperatura percepita, calcolata con la temperatura media e umidità:

$$\begin{aligned} HI = &-42.379 + 2.04901523T + 10.14333127RH \\ &-0.22475541TRH - 0.00683783T^2 \\ &-0.05481717RH^2 + 0.00122874T^2RH \\ &+0.00085282TRH^2 - 0.00000199T^2RH^2 \end{aligned}$$

  Dove T è la temperatura indicata in Fahrenheit e RH è l'umidità relativa espressa in percentuale.

- **Fenomeni atmosferici**: Variabile categoriale pari ad uno dei seguenti valori: *temporale*, *neve*, *nebbia*, *pioggia*, *sereno*.

I dati meteorologici sono disponibili per tutto l'intervallo di tempo considerato ed oltre, permettendo di effettuare previsioni future anche sulla base delle condizioni meteorologiche.

2. Analisi esplorative e pre-processing

Innanzitutto, è stata effettuata un'analisi esplorativa delle serie storiche del lordo totale e del numero di scontrini dei sei ristoranti a disposizione. Nella Figura 1 è rappresentato l'andamento giornaliero del lordo totale del Ristorante 3, dal Febbraio 2018 ad Aprile 2020.

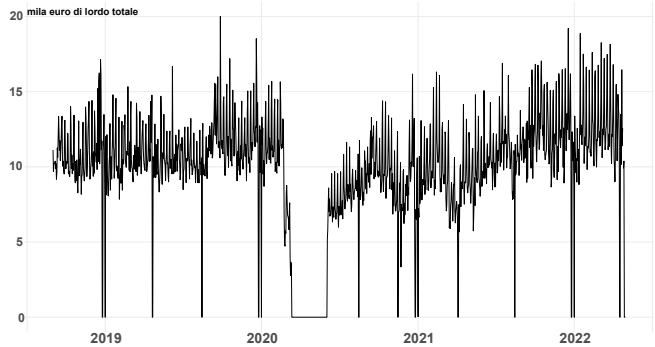

Figura 1. Serie storica del lordo totale del Ristorante 3

La prima evidente anomalia è il valore pari a 0 del lordo totale nel periodo che va da Marzo 2020 a Giugno 2020. Ciò è dato dall'emergenza Covid-19 ed è un aspetto comune a tutti i sei ristoranti. Per questo motivo, questa porzione di quattro mesi non è stata utilizzata nelle successive analisi. I dati sono stati quindi divisi in due parti:

- **Dati pre-Lockdown**, ovvero 487 osservazioni giornaliere dal Settembre 2018 al Dicembre 2019.

- **Dati post-Lockdown**, ovvero 695 osservazioni giornaliere dal Giugno 2020 ad Aprile 2022.

Nelle successive analisi esplorative sono state utilizzate entrambe le porzioni di dati. Nella fase predittiva dello studio, è stata studiata unicamente l'ultima porzione dei dati.

Un'ulteriore anomalia è stata riscontrata in tutti e sei i ristoranti: per alcuni giorni il lordo totale ed il numero di scontrini è pari a zero. Ciò può essere causato da una mancanza di clienti per quel giorno, da una mancata registrazione dei dati o dalla chiusura dei ristoranti. I valori nulli sono stati sostituiti con un valore mancante e successivamente imputati stimando un modello ARIMA, dato che la presenza di questi valori estremamente bassi non spiegati, seppur pochi, può distorcere fortemente le previsioni. Nel calcolo delle metriche di performance, i valori nulli nel test set non sono stati considerati, siccome è molto probabile che il ristoratore non abbia bisogno di effettuare previsioni per un determinato giorno se ha intenzione lasciar chiuso il ristorante.

Dal grafico della Figura 1 è inoltre possibile cogliere due aspetti fondamentali della serie storica:

- Prima del primo lockdown datato 9 Marzo 2020, la serie storica è caratterizzata da una stagionalità settimanale e dalla mancanza di un trend crescente o decrescente.
- Dopo il lockdown, la serie storica assume un comportamento più anomalo, con la presenza di alcuni giorni con lordo nullo. La stagionalità settimanale è ancora presente, ma vi è anche un trend crescente.

La stagionalità settimanale presente in entrambe le sezioni della serie storica è mostrata analizzando l'andamento delle serie storiche nelle Figure 2 e 4 e nella Tabella 1. Le quattro serie storiche sono state ottenute suddividendo i giorni in feriali e festivi, per poi calcolare le medie settimanali delle due categorie. I giorni festivi includono il week-end (Venerdì, Sabato e Domenica) e le festività come Natale, Capodanno ed altre. Tutti i rimanenti giorni sono stati considerati "feriali".

Tabella 1. Valore medio del lordo totale e numero di scontrini per giorno della settimana

Pre-Lockdown

Variabile	Lun	Mar	Mer	Gio	Ven	Sab	Dom
Lordo totale	10534	10195	10495	10585	12302	13316	10636
N. scontrini	772	752	760	764	841	893	657

Post-Lockdown

Variabile	Lun	Mar	Mer	Gio	Ven	Sab	Dom
Lordo totale	9730	9477	9761	9852	11552	13214	10257
N. scontrini	513	495	508	513	565	620	455

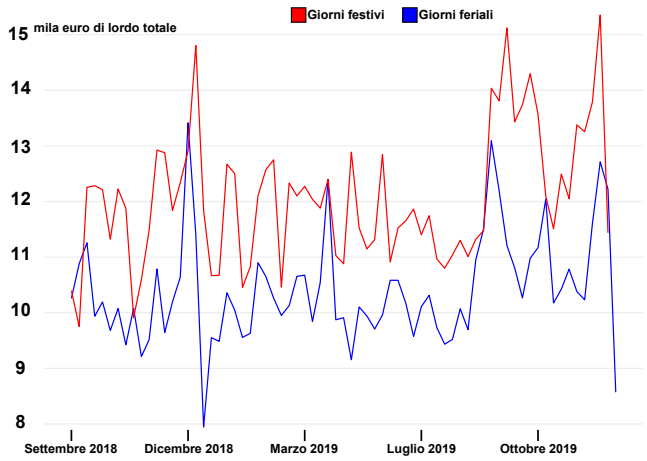

Figura 2. Serie storica del lordo totale nei giorni pre-covid

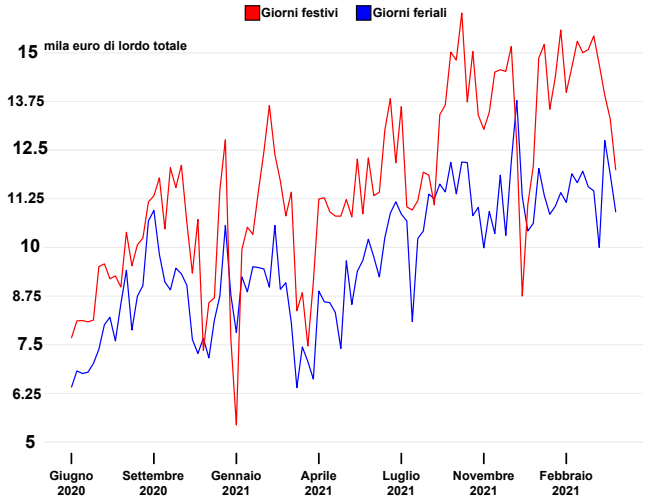

Figura 3. Serie storica del lordo totale nei giorni post-covid

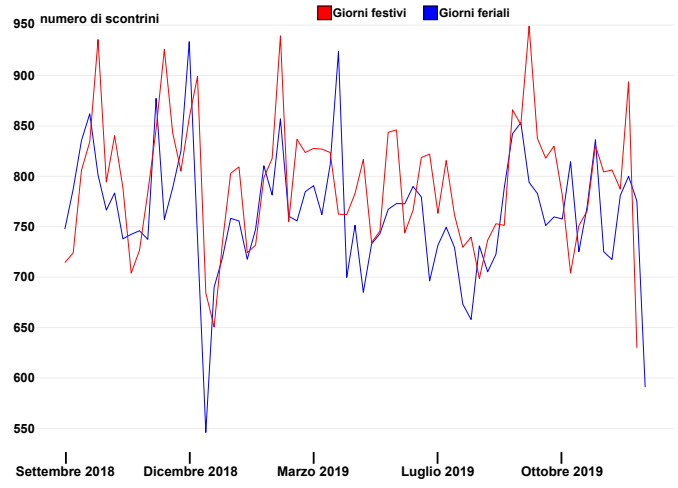

Figura 4. Serie storica del numero di scontrini nei giorni pre-covid

Tabella 2. Statistiche riassuntive per le variabili utilizzate

Pre-Lockdown

Variabile	Media	Dev. Std	Mediana
Lordo totale	11152.51	1741.89	10755.83
N. scontrini	777.38	104.07	769
Heat Index	84.90	21.09	80.27
Temperatura (F)	60.18	14.93	59
RH	77.26	19.48	78.5

Post-Lockdown

Variabile	Media	Dev. Std	Mediana
Lordo totale	10547.33	2540.77	10350.64
N. scontrini	524.64	98.34	526
Heat Index	88.36	20.38	83.02
Temperatura (F)	58.94	15.90	57.2
RH	77.65	17.85	77

Sia nelle settimane pre-lockdown che nelle settimane post-lockdown gli incassi dei ristoranti sono più elevati nei giorni festivi rispetto ai feriali, soprattutto il Sabato ed il Venerdì. Per tutto il periodo che va dal Settembre 2018 a Dicembre 2019, i valori del lordo totale del Ristorante 3 oscillano attorno ad un valore medio fisso sia nei week-end che durante la settimana. Inoltre, non è presente una stagionalità annuale nella serie storica del lordo totale, ma è possibile notare dei picchi negativi e positivi ciclici.

Dall'andamento della serie storica degli scontrini si nota che questa differenza marcata tra i giorni feriali e festivi non è presente, a causa del minor numero di scontrini effettuati in media nelle Domeniche. Nonostante ciò, sia nel periodo pre-lockdown che post-lockdown, la Domenica è il terzo giorno per maggiore lordo totale medio, indicando che l'incasso maggiore non è dovuto ad un aumento del numero di clienti, ma ad una propensione dei clienti di spendere di più di Domenica o all'aumentare nei week-end del numero medio di persone per tavolo. Questi risultati sono stati riscontrati anche nei dati riguardanti gli altri cinque ristoranti non mostrati nello studio.

Infine, nella tabella 2 sono riportate alcune statistiche riassuntive riguardanti le variabili utilizzate nello studio. Come già visto dai precedenti grafici, il lordo totale ed il numero di scontrini medio ha avuto un calo dopo l'arrivo della pandemia. Inoltre, vi è anche un aumento della varianza dovuto principalmente ai forti picchi negativi causati dall'introduzione di nuove restrizioni (come il DPCM del 6/11/2020).

3. Confronto dei trend post-lockdown

Come già anticipato in precedenza, l'introduzione del lockdown e l'arrivo della pandemia hanno avuto un forte impatto sugli incassi dei ristoranti, costretti a chiudere nei primi mesi di pandemia e ad operare non a pieno regime durante tutto il 2020 e gran parte del 2021.

L'introduzione molto discussa dei colori per le regioni ha interessato molte attività e, avendo a disposizione gli incassi di sei ristoranti del nord Italia, è stato possibile analizzare come il cambiamento dei colori ha impattato il lordo e se le attività hanno attraversato un periodo di ripresa. Per presentare i risultati è stato scelto un approccio grafico, ottenuto attraverso:

- Identificazione dei periodi in cui il colore della regione di riferimento ha assunto colori diversi dal "bianco". Siccome tutti i ristoranti a disposizione sono localizzati in Emilia-Romagna o in Lombardia, sono state rappresentate le date in cui è stato applicato un colore diverso alla regione in questione nei periodi che vanno da Giugno 2020 ad Aprile 2022.
- Decomposizione delle sei serie storiche del lordo dei ristoranti. Al fine di isolare i fattori esterni che non dipendono dal cambiamento del colore della regione, la parte stagionale e residua della serie storica è stata rimossa, mantenendo solo il trend. È stata utilizzata una decomposizione STL ("Seasonal and Trend decomposition using Loess") [3]. Questo metodo di decomposizione permette di utilizzare qualsiasi tipo di stagionalità ed è possibile regolare quanto "smooth" deve essere il trend risultante. Inoltre, regolando il valore del parametro λ, è stato possibile applicare la famiglia di trasformazioni di Box-Cox alle serie storiche prima di decomporle.

Il risultato ottenuto è rappresentato nella Figura 5, da cui è possibile dedurre che:

- I trend di tutti i ristoranti sono in leggera crescita. Il lordo medio ad Aprile 2022 è più elevato rispetto al 2021 ed al 2020.
- Dopo l'estate 2020 si è registrato un forte calo nel lordo dei ristoranti, dovuto al rapido aumento dei contagi che ha portato alla successiva introduzione del DPCM del Novembre 2020.
- Il passaggio alla zona rossa del Gennaio 2021 e dell'Aprile 2021, anche se breve, ha avuto un forte impatto sulle vendite dei sei ristoranti.
- Solo dopo Giugno 2021 per tutti i ristoranti è iniziata una fase di ripresa, dovuta sia all'alleggerimento delle restrizioni sia all'aumento del numero di persone vaccinate in Italia. Il 15 Giugno 2021 il numero di persone che avevano ricevuto almeno la prima dose di vaccino ha superato il 50%.
- Ad Ottobre 2021, il Ristorante 1 ha chiuso per più di tre settimane, ovviamente portando gli incassi a zero. In concomitanza, il Ristorante 2 ha subito un forte rialzo dei propri incassi nello stesso periodo. Poiché questi ristoranti, insieme al Ristorante 3, sono gli unici localizzati precisamente a Piacenza e non in provincia, si

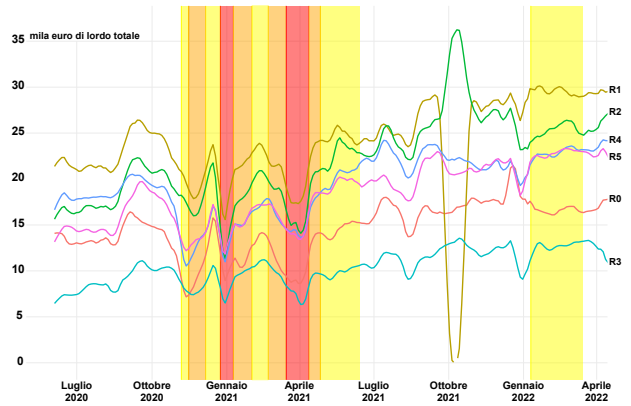

Figura 5. Confronto trend ristoranti post-lockdown

potrebbe avanzare l'ipotesi che la chiusura del Ristorante 1 abbia giovato al Ristorante 2 per via del cambio di flusso di clienti.

4. Previsione del lordo totale

Nella seguente sezione vengono esposti i risultati ottenuti utilizzando diverse metodologie di previsione per serie storiche. In particolare, nella Sezione 4.1.1 sono stati stimati quattro modelli ARIMA stagionali e non stagionali, effettuando previsioni e validazione delle performance con una semplice divisione delle osservazioni in train set e test set. Nella Sezione 4.1.2 è stato utilizzato il modello SARIMA che ha ottenuto le migliori performance ed è stato nuovamente effettuato un processo di stima e previsione utilizzando una strategia di campionamento nota come Time Series cross-validation, mostrata nella Figura 6. Dopodiché, nella sezione 4.2, le metodologie mostrate nella Sezione 4.1 sono state riapplicate sulle serie storiche a periodicità settimanale ottenute dividendo i giorni tra feriali e festivi e calcolando il lordo medio settimanale per il terzo ristorante. Infine, nella Sezione 5.1 è stata applicata un modello noto come Cluster-Weighted Models. Per semplicità di esposizione, in questa parte dell'analisi, si fa riferimento ad un solo ristorante, il numero 3 (si tratta del ristorante che presenta meno comportamenti anomali fra quelli sopra riportati e, inoltre, è l'unico situato a Piacenza, comune di cui possediamo i dati meteorologici).

4.1 Stima e previsioni con modello ARIMA

4.1.1 Suddivisione in Train/Test

Nel primo approccio i modelli sono stati addestrati su un train set composto dalle prime 569 osservazioni giornaliere. La valutazione delle performance previsive dei modelli è stata effettuata confrontando le previsioni del lordo totale dei successivi 126 giorni. Il train set va quindi dal 3 Giugno 2020 al 23 Dicembre 2021, mentre l'ultima osservazione del test set è datata 28 Aprile 2022. In particolare, sono stati utilizzati quattro modelli diversi:

- **Arima(1,1,1)**, in cui sono state utilizzate una variabile dicotomica per indicare i giorni di vacanza ed i termini di Fourier per indicare la stagionalità settimanale.
- **Arima(5,1,4)**, in cui è stata utilizzata unicamente la variabile dicotomica indicante le vacanze come esplicativa.
- **Sarima(2,0,1)(0,1,1)**, con la variabile dicotomica indicante le vacanze.
- **Sarima(2,0,1)(0,1,1)** con vari regressori, tra cui: heat index, il colore della regione, la presenza di nebbia ed altre.

Tabella 3. Metriche di performance per i quattro modelli

Modello	MAPE	MAE	RMSE
Arima(1,1,1)	22.06	2607.7	3095.2
Arima(5,1,4)	23.83	2727.7	3063.2
Sarima con dummy vacanze	11.89	1365.6	1863.1
Sarima con regressori	11.39	1329.5	1869.0

Dalla Tabella 3 si apprende che:

- Il **Sarima con regressori** risulta essere il modello con i migliori valori di *MAPE* e *MAE*. Le informazioni precedentemente citate riguardo il numero di contagiati nella provincia di Piacenza sono state escluse in quanto non significative. Il modello è più complesso rispetto agli altri tre, ma i risultati sono migliori.
- Il modello **Arima(5,1,4)** ha pessime performance, ciò è dovuto al fatto che è l'unico modello che non spiega in alcun modo la stagionalità della serie storica.

4.1.2 Forecast Cross-Validation

Utilizzando il modello Sarima con regressori, è stata adoperata una differente strategia per effettuare previsioni e dividere le osservazioni in train e test: la time-series cross-validation a finestra mobile. Questo tipo di suddivisione è basata sulla scelta dei valori di due iperparametri: la finestra di previsione p ed il numero osservazioni nel test set n_{test}. Le previsioni e la stima del modello, a differenza del metodo precedente, non vengono calcolate un'unica volta. Con la ts-CV la stima del modello viene effettuata un numero di volte pari a $\frac{n_{test}}{p}$. Ad esempio, utilizzando una finestra di previsione pari a 7 giorni ed un numero totale di osservazioni nel test set pari a 70, vengono effettuate un totale di 10 iterazioni. Per ogni iterazione il modello viene stimato sulle osservazioni appartenenti al train set e viene effettuata una previsione delle prime 7 osservazioni appartenenti al test set. Dopodiché il train set viene aggiornato, rimuovendo le 7 osservazioni meno recenti, ed aggiungendo in coda le ultime 7 osservazioni. Il processo viene ripetuto fino a quando non si ha a disposizione una previsione per tutte le 70 osservazioni del test set. Nella figura

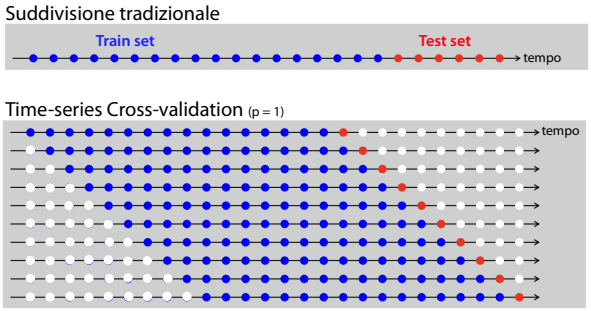

Figura 6. Differenza tra la suddivisione classica in train e test set e la time-series cross-validation

6 viene sinteticamente mostrato il funzionamento di questo metodo.

Nella tabella 4 sono riportati i risultati ottenuti applicando una strategia di cross-validation al modello Sarima con regressori. Il numero di osservazioni appartenenti al train set è fisso e pari a 569, mentre è stato scelto di applicare una finestra di previsione giornaliera (126 iterazioni), settimanale (18 iterazioni) e bisettimanale (9 iterazioni).

Tabella 4. Metriche di performance per le tre finestre di previsione

Modello	MAPE	MAE	RMSE
$p = 1$	9.19	1120.8	1570.5
$p = 7$	9.90	1212.4	1685.7
$p = 14$	10.62	1279.6	1811.9

Dalla Tabella 4 si ha che:

- Le performance previsionali ottenute sono migliori rispetto ad una semplice suddivisione tradizionale per ogni finestra di previsione scelta e per ogni metrica studiata.
- I risultati migliori per ogni metrica utilizzata si hanno con una finestra di previsione "giornaliera", che risulta essere anche l'approccio più computazionalmente costoso, dato l'elevato numero di modelli da stimare.

4.2 Suddivisione in Feriali e Festivi

Dalle analisi esplorative e i modelli visti in precedenza è emerso che vi è una netta differenza tra i week-end, i giorni di festa ed i giorni lavorativi. Per questo motivo, è stata testata una differente strategia di previsione. Sono stati stimati due modelli ARIMA differenti: uno per i giorni festivi ed uno per i giorni feriali. Una volta suddivise le osservazioni in questi due gruppi, la granularità delle osservazioni è stata ridotta, passando da una serie giornaliera ad una serie in cui ogni osservazione rappresenta il lordo totale medio di una settimana dell'anno. Sia per il gruppo di osservazioni "festivo" che per il gruppo "feriale" il modello ARIMA(0,1,1) è stato scelto autonomamente dal software come migliore. Infine, lo stesso modello è stato addestrato sia suddividendo le osservazioni in train e test set sia utilizzando la time-series cross-validation con finestra settimanale ($p = 1$), mensile ($p = 4$) e trimestrale ($p = 14$).

Tabella 5. Metriche di performance per i giorni feriali e festivi

Feriali

Finestra di previsione (settimane)	MAPE	MAE	RMSE
Train/Test	6.77	743.6	956.1
$p = 1$	6.86	779.1	985.8
$p = 4$	6.52	741.2	1021.9
$p = 14$	9.10	994.3	1225.6

Festivi

Finestra di previsione (settimane)	MAPE	MAE	RMSE
Train/Test	14.50	1787.4	2312.7
$p = 1$	9.49	1206.9	1646.3
$p = 4$	12.9	1661.2	2112.5
$p = 14$	16.50	2066.4	2707.0

Dalla Tabella 5 si deduce che:

- I valori ottenuti per le tre metriche sono sensibilmente più elevati per i giorni festivi. Ciò indica che la variabilità associata a questi giorni è fortemente più elevata, il che rende qualsiasi tipo di previsione più inaffidabile.
- Utilizzare modelli e previsioni differenti risulta essere una scelta sensata, siccome poca parte dell'errore percentuale delle previsioni è data dai giorni feriali.
- L'utilizzo della time-series cross-validation ha portato ad un miglioramento nelle performance nel caso dei giorni festivi. Effettuare delle previsioni settimanali con una finestra di previsione breve porta a risultati migliori, anche a causa dell'alta volatilità del lordo totale nelle vacanze e nei week-end. Nel caso dei giorni feriali l'utilizzo della ts-CV non ha portato ad un minore errore di previsione.

5. Uno sguardo sugli sviluppi futuri

5.1 Cluster-Weighted Models

Sulla scia dell'utilizzo di diversi modelli, come si è appena visto per la suddivisione in giorni feriali e festivi, ci si potrebbe domandare se gli stessi dati possiedono una propensione ad essere intrinsecamente spiegati da molteplici modelli.
Per poter approfondire questa domanda di ricerca, è stato

adottato l'approccio dei Cluster-Weighted Models [1] implementato tramite il pacchetto R **flexCWM** [5].
Dato che un numero troppo alto di cluster non fornirebbe insights apprezzabili sull'evoluzione dei ristoranti, si è deciso di eseguire la ricerca di misture di modelli ottimali sino ad un massimo di 5 cluster, tutti inizializzati tramite algoritmo di k-means.
I risultati ottenuti sembrano mostrare, dopo un'analisi grafica, la presenza di fasce distinte di guadagno lordo a seconda del giorno della settimana e del colore della regione in quella stessa settimana. In figura 7 si mostra quanto appena detto per il miglior modello secondo il criterio AICu, con 3 cluster.

Poiché questo approccio richiede uno studio molto più approfondito, non sono state eseguite delle previsioni su un dataset di test, ma si vogliono mostrare in figura 8 i risultati ottenuti sul training set.

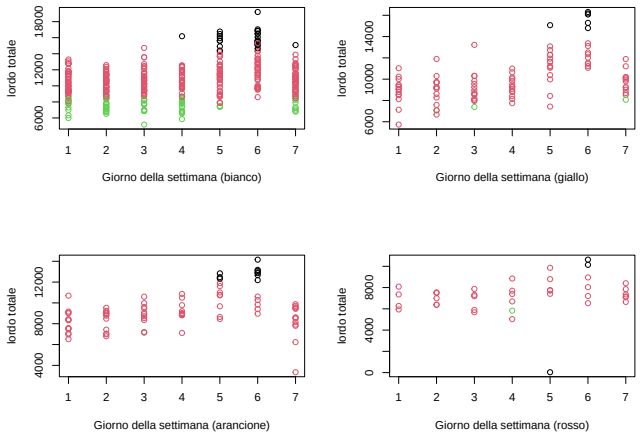

Figura 7. Cluster identificati col metodo CWM

Si può osservare che in tutti i casi è presente una intercetta significativa e di alto valore, ad indicare che gran parte del lordo totale non dipende da altri fattori esterni, ma, come già visto nelle precedenti analisi, da trend e stagionalità già presenti.
Altro dato degno di nota è il fatto che il regressore *holiday* non è significativo per due fasce di lordo totale, ma lo è per la più alta.

Notare i valori del RMSE (qui erroneamente indicati con *sigma* dal pacchetto flexCWM).

5.2 Inclusione del listino prezzi

Un altro studio interessante che però necessiterebbe di ulteriori dati come il listino prezzi dei ristoranti potrebbe essere quello analizzare come siano cambiate le spese pro capite dei clienti prima e dopo la pandemia, che però con i dati aggregati che si hanno a disposizione non risulta possibile.

```
------------------------------------
Best fitted model according to AICu
------------------------------------

Clustering table:
  1   2   3
 45 454  70

Prior: comp.1 = 0.084858, comp.2 = 0.728354, comp.3 = 0.186788

Distribution used for GLM: gaussian(identity). Parameters:

Component 1
               Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  11549.5340   146.4883  78.8427 < 2.2e-16 ***
HI              13.7103     1.3094  10.4710 < 2.2e-16 ***
pioggia       -164.1012   106.2312  -1.5448     0.123
neve         -2408.8115   257.8902  -9.3405 < 2.2e-16 ***
nebbia         503.0401   111.5117   4.5111 7.865e-06 ***
temporale    -4277.1797   149.6817 -28.5752 < 2.2e-16 ***
colorebianco  3232.0512    88.7445  36.4197 < 2.2e-16 ***
coloregiallo  2521.1696   117.0958  21.5308 < 2.2e-16 ***
colorerosso  -2571.9794   170.6561 -15.0711 < 2.2e-16 ***
holiday      -8172.4277   342.2223 -23.8805 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
sigma = 740.28

Component 2
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)   7927.1463   427.6834 18.5351 < 2.2e-16 ***
HI               7.7025     3.7201  2.0705   0.03886 *
pioggia       -233.2011   202.4733 -1.1518   0.24991
neve         -1621.2903   737.0307 -2.1998   0.02823 *
nebbia         974.8103   241.1507  4.0423 6.035e-05 ***
temporale     -539.6259   307.4834 -1.7550   0.07981 .
colorebianco  2283.3553   216.3357 10.5547 < 2.2e-16 ***
coloregiallo  1452.5691   252.0169  5.7638 1.361e-08 ***
colorerosso  -1365.6682   331.3988 -4.1209 4.345e-05 ***
holiday        570.3137   625.0696  0.9124   0.36195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
sigma = 1568.5

Component 3
               Estimate Std. Error  t value  Pr(>|t|)
(Intercept)   6527.7458   231.7401  28.1684 < 2.2e-16 ***
HI              26.8190     2.1018  12.7603 < 2.2e-16 ***
pioggia       -204.1556    95.0584  -2.1477 0.0321678 *
neve           547.5811   297.3505   1.8415 0.0660729 .
nebbia        -989.4735   112.5284  -8.7931 < 2.2e-16 ***
temporale     -520.3834   142.7338  -3.6458 0.0002914 ***
colorebianco  -879.4917   103.2652  -8.5168 < 2.2e-16 ***
coloregiallo   317.1440   112.3808   2.8220 0.0049416 **
colorerosso  -1807.0782   138.0595 -13.0891 < 2.2e-16 ***
holiday        341.3662   262.4603   1.3006 0.1939179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
sigma = 706.73
```

Figura 8. Risultati delle misture di modelli di regressione lineare

6. Conclusioni

Partendo da un'analisi descrittiva dei dati a disposizione, sono stati implementati modelli predittivi in grado di effettuare previsioni di carattere economico con una discreta accuratezza. Questo studio fornisce una base gestionale per intraprendere alcune decisioni strategiche di investimento delle risorse degli imprenditori, in particolare la gestione del magazzino(e quindi delle materie prime) e del personale. Riguardo lo strumento previsionale, la separazione dei giorni in due gruppi differenti (feriali e festivi) e la stima di un modello differente per ogni gruppo si è rivelata una scelta vincente. I risultati ottenuti nei giorni feriali sono soddisfacenti, con modelli poco complessi che riescono ad ottenere un errore percentuale minore del 7%. Gli incassi nei giorni festivi, caratterizzati da una maggiore volatilità, risultano essere più difficili da prevedere, con un errore percentuale intorno al 10%. La strategia consigliata ai ristoratori è quindi la seguente:

- nei giorni feriali è possibile affidarsi alle previsioni del lordo futuro e di conseguenza affidarsi ad esse per decidere la quantità di materie prime da pre-ordinare ed il numero di tavoli da aspettarsi
- Nei giorni festivi è comunque consigliato utilizzare le previsioni per avere una linea guida, ma l'utilizzo di un sistema di prenotazioni con conseguente penale in caso di assenza di un cliente potrebbe rivelarsi utile nel gestire la volatilità degli incassi.

7. Appendice

7.1 Aspetti metodologici

Per effettuare le previsioni sono stati utilizzati tre strumenti: Arima, Sarima e CWM.

7.1.1 Arima

Modello utilizzato per analizzare quali siano le variabili che aiutano a prevedere l'incasso futuro. [4] è un punto di riferimento della modellazione delle serie storiche. E' composto da 3 componenti:

- **AR**: Modellazione autoregressiva: $X_t = c + \sum_1^p \phi_i X_{t-i} + \varepsilon_t$
- **I**: Integrazione della serie storica, utilizzata per serie non stazionarie;
- **MA**: Modellazione a media mobile(moving average): $X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q + \varepsilon_{t-q}$

Il modello ARIMA deriva dal modello ARMA a cui sono state applicate le differenze di ordine d per renderlo stazionario. I parametri del modello sono 3:

- **p**: numero di lag della componente autoregressiva;
- **d**: numero di integrazioni;
- **q**: numero di lag della componente moving average.

7.1.2 Sarima

Seasonal ARIMA [2], necessario in quanto i modelli arima non considerano le stagionalità all'interno delle serie storiche. Una stagionalità è definita come la ricorrenza ciclica di pattern all'interno della serie storica. I modelli arima si aspettano dati senza stagionalità o con stagionalità rimossa con metodi come la differenziazione stagionale. Sarima è perciò una estensione di arima e oltre ai parametri già descritti si aggiungono anche parametri puramente stagionali:

- **P**: numero di stagioni della componente autoregressiva;
- **D**: numero di integrazioni nelle stagioni;
- **Q**: numero di stagioni della componente moving average;
- **m**: numero di punti componenti la stagione.

7.1.3 Funzione Vacanza

La funzione holiday consente di tenere conto delle festività all'interno del modello che sono possono influenzare in maniera considerevole l'incasso nei suddetti giorni. Le vacanze considerate sono quelle delle festività italiane:

- 1 gennaio e 6 gennaio;
- Pasqua;
- Ferragosto;
- Natale e Santo Stefano.

7.1.4 Cluster-Weighted Models

Lo scopo dei CWM è di identificare dei cluster dei dati su cui effettuare diverse stime dei parametri di un modello statistico e ottenere un fitting ottimale (una mistura di modelli di regressione lineare).
La distribuzione dei dati (X,Y) può essere scritta come

$$p(x,y;\theta) = \sum_{j=1}^{k} \pi_j p(y|x;\beta_j,\gamma_j) p(x;\alpha_j)$$

dove, per la j-esima componente, π_j è la proporzione di mixing, con $\pi_j > 0$ e $\sum_{j=1}^k \pi_j = 1$, $p(y|x;\beta_j,\gamma_j)$ è la distribuzione parametrica (rispetto a β_j,γ_j) di $Y|X = x$, e $p(x;\alpha_j)$ è la distribuzione parametrica di X rispetto a α_j.

Il fitting dei parametri viene fatti tramite l'algoritmo EM (expectation-maximization).

Riferimenti bibliografici

[1] Paolo Berta et al. "Multilevel cluster-weighted models for the evaluation of hospitals". In: *Metron* 74.3 (2016), pp. 275–292.

[2] Duanyang Liu Wei Jiang Bin Ma Chen Aichen Niu. "Time series forecasting of temperatures using SARIMA: an example from Nanjing". In: (2018).

[3] Robert B. Cleveland et al. “STL: A Seasonal-Trend Decomposition Procedure Based on Loess (with Discussion)”. In: *Journal of Official Statistics* 6 (1990), pp. 3–73.

[4] Aman El Moussami Lachhab Fattah Ezzine. “Forecasting of demand using ARIMA model”. In: (giu. 2018).

[5] Angelo Mazza, Antonio Punzo e Salvatore Ingrassia. “flexCWM: A Flexible Framework for Cluster-Weighted Models”. In: *Journal of Statistical Software* 86 (2018), 1–30. DOI: 10.18637/jss.v086.i02. URL: https://www.jstatsoft.org/index.php/jss/article/view/v086i02.