Quanto ci è mancato andare al ristorante?

Calvi Alice a.calvi10@campus.unimib.it
La Franca Marta m.lafranca@campus.unimib.it
Maggipinto Alessandra a.maggipinto@campus.unimib.it

1 Sinossi

La pandemia ha colpito fortemente il settore della ristorazione e questo ha motivato principalmente il nostro studio, che si è basato sui dati delle vendite e degli scontrini emessi dal 2017 al 2021 di 6 ristoranti sul territorio tra Emilia Romagna e Lombardia.

L'analisi si è concentrata su due domande: come abbia influito il Covid 19 sulle vendite e sugli scontrini emessi tenendo conto del comportamento delle persone negli anni passati e riuscire ad effettuare una previsione su quanto le persone siano disposte a spendere.

La prima di tali domande ha fatto emergere che lo "shock" portato dalla pandemia, ed in particolare il periodo di chiusura totale, ha comportato un aumento notevole delle vendite nel mese di Maggio 2020 (cioè fine lockdown) rispetto al periodo pre-Covid, per tutti i ristoranti considerati.

Inoltre, si è cercato di arricchire la previsione sullo scontrino medio considerando fattori esterni, e i dati a disposizione di tutti i ristoranti, ottenendo un buon risultato di previsione.

2 Parole chiave

- **Time Series Decomposition;**
- **Sarima model;**
- **Var Model;**
- **Forecasting;**
- **Exogenous variable.**

3 Introduzione

Riuscire ad elaborare modelli per stimare lo scontrino medio di un cliente è di importanza fondamentale per un ristoratore.

In questo lavoro si utilizza un dataset contenente le vendite giornaliere e il numero di scontrini di sei esercizi di ristorazione. Si sono elaborati dei modelli di previsione sullo scontrino medio di un ristorante, dopo aver effettuato un'analisi sulle caratteristiche delle sei serie temporali presenti nel dataset.

Si è deciso di utilizzare come variabile dipendente lo scontrino medio piuttosto che le vendite perché è una buona sintesi dei dati a disposizione e una misura meno distorta o influenzata da possibili outliers.

Si è inoltre effettuata un'analisi di più ristoranti attraverso un modello che riuscisse a considerare più serie temporali, per riuscire ad effettuare un'analisi e una previsione più generale che tenga conto di più fattori.

4 Obiettivo/problema affrontato

L'obiettivo del lavoro svolto è quello di provare a rispondere alle domande:

- Quando le persone preferiscono andare al ristorante? E quanto spendono? C'è un trend di comportamento?
- Quale situazione si è profilata al momento delle riaperture e come ha influito lo shock pandemico sulle vendite?
- Conoscere dati di più ristoranti può aiutare a eseguire la previsione di uno?

Si è ritenuto molto importante cercare di capire come la popolazione abbia reagito dopo il Covid: i ristoranti

hanno subito gravi danni per via della pandemia, molti hanno chiuso, molti sono riusciti a reinventarsi e un modo per riuscire a ritornare competitivi è studiare il comportamento delle persone.

5 Aspetti metodologici

Durante il lavoro sono stati utilizzati i seguenti modelli:

- Modello SARIMA (Seasonal Autoregressive Integrated Moving Average) su un ristorante;
- Modello SARIMA con variabili esterne;
- Modello VAR (vettori autoregressivi) su più ristoranti;

Il modello ARIMA (Autoregressive Integrated Moving Average) è uno dei metodi di previsione più utilizzati per la previsione di dati di serie temporali univariate.

Sebbene il metodo possa gestire i dati con una trend, non supporta le serie temporali con una componente stagionale, per questo motivo si è deciso di utilizzare un'estensione di ARIMA che supporti la modellazione diretta della componente stagionale della serie, ossia il modello SARIMA.

Si è inoltre deciso di implementare tale modello utilizzando nuove features create durante la fase di Preprocessing.

Infine è stato utilizzato un modello VAR (Vector Autoregression), un algoritmo di previsione capace di trattare più serie temporali congiuntamente.

6 Dati

Il dataset riporta i dati di vendita di sei esercizi nel campo della ristorazione, del Nord Italia.

Il dataset in esame è composto da 1563 righe e da 13 colonne, le quali descrivono il fatturato giornaliero in euro e il numero di scontrini per ogni esercizio.

I dati si presentano così composti:

- date;
- Vendite_1;
- Scontrini_1;
- Vendite_2;
- Scontrini_2;
- Vendite_3;
- Scontrini_3;
- Vendite_4;
- Scontrini_4;
- Vendite_5;
- Scontrini_5;
- Vendite_6;
- Scontrini_6.

I dati partono da gennaio 2017, tuttavia alcuni esercizi sono stati aperti in un periodo successivo e per questo le serie storiche cominciano successivamente.

Tutti gli esercizi presentano valori nulli relativi al periodo di chiusura del lockdown avvenuto da marzo 2020 ad aprile 2020.

7 Analisi/Processo di trattamento dei dati

7.1 Preprocessing: Feature Creation

Sono state create nuove variabili ritenute utili ai fini del lavoro svolto utilizzando il linguaggio R, in particolare:

- Variabile *month*: è stata estratta l'informazione relativa al mese.
- Variabile *weekday*: è stata estratta l'informazione relativa al giorno della settimana. Sono state create tali colonne al fine di poter trasformare la colonna relativa alla data secondo il formato standard %y-%m-%d.
- Variabile *is_weekend*: variabile booleana che assume valori 0,1 a seconda che il giorno considerato sia un giorno appartenente al weekend o meno.
- Variabile *holiday*: variabile booleana che assume valori 0,1 a seconda che il giorno considerato sia un giorno festivo o meno. Sono stati considerati i giorni festivi corrispondenti ad ogni anno:
  - 1 gennaio

- 6 gennaio;
- San Valentino;
- Domenica delle palme;
- Pasqua;
- Lunedì dell'Angelo;
- 25 aprile;
- 1 maggio;
- 2 giugno;
- 15 agosto;
- 1 novembre;
- 8 dicembre;
- 24 dicembre;
- 25 dicembre;
- 26 dicembre;
- 31 dicembre.

- Variabile *scontr_medio*: variabile float ricavata per ogni ristorante, indica il valore medio degli scontrini emessi dando un'idea di quanto spende in media ogni cliente. Si ottiene dividendo il fatturato totale per il numero di scontrini emessi.

- Variabile *temp_media*: variabile float, ottenuta integrando il dataset d'origine con un dataset presente sul sito "*https://www.ilmeteo.it/*" relativo alle temperature medie giornaliere dell'area a sud di Lodi, in quanto area circostante il dataset considerato.

7.2 Visualizzazione serie

In un primo momento di analisi si è scelto di porre l'attenzione su un singolo ristorante, suddividendo così il dataset e prendendo in considerazione il primo ristorante in ordine di presentazione. Si è inoltre deciso di escludere i valori nulli al fine di evitare alterazioni.

E' stata analizzata la serie temporale avente variabile dipendente relativa lo scontrino medio, per considerare la maggior quantità di informazioni possibili.

Attraverso l'uso della libreria *matplotlib* si è effettuata una prima visualizzazione della serie in figura 1: Da tale rappresentazione è possibile osservare come lo

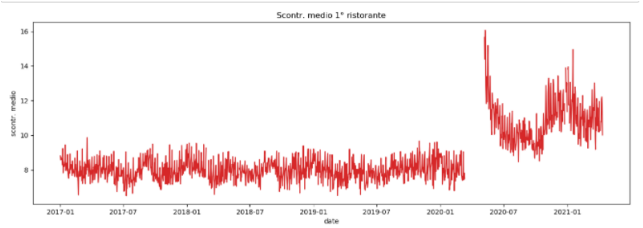

Figura 1: Overview sulla serie temporale.

scontrino medio nel periodo successivo al primo lockdown abbia subito un rialzo che ha iniziato subito a decrescere soprattutto durante l'inverno del 2020, probabilmente dovuto alla seconda ondata pandemica e la conseguente chiusura delle attività in loco ad eccezione dell'asporto. Sul finire dell'anno 2020 la curva ha ricominciato a risalire con alti e bassi ma mantenendosi piuttosto stabile.

Si è proseguito effettuando dei boxplot in *Figura 2* per considerare il trend e la stagionalità.

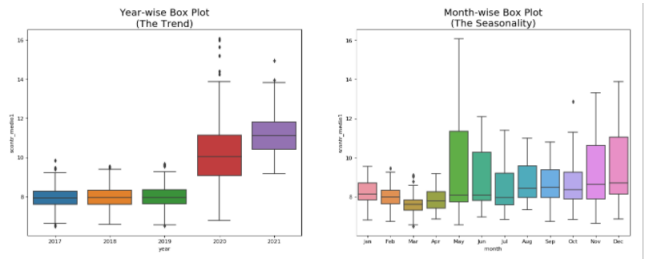

Figura 2: Boxplot prima serie temporale.

É possibile osservare come per il primo ristorante nel 2020 e nel 2021 vi sia un aumento dello scontrino medio, dovuto allo shock causato dalle chiusure e dalle conseguenti riaperture.

Il mese di maggio e il mese di giugno presentano un aumento dei valori, supponiamo che sia presente una stagionalità.

Successivamente si è proceduto effettuando tali analisi anche per quanto riguarda gli altri esercizi.

In *Figura 3* sono presentati i boxplot ottenuti analizzando il 2° ristorante in ordine di presentazione nel dataset, dalla cui analisi è possibile osservare comportamenti molto simili alla prima serie temporale.

In *Figura 4* osservando i boxplot ottenuti analizzando il 3° ristorante, in ordine di presentazione nel dataset, si osserva come tale, nonostante abbia aperto solo successivamente rispetto gli altri esercizi, presenti comun-

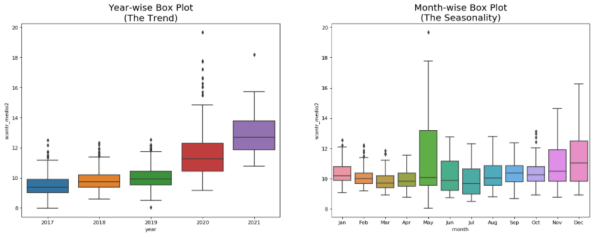

Figura 3: Boxplot seconda serie temporale.

que un andamento simile ai precedenti: un aumento dello scontrino medio nel mese di maggio e giugno e un aumento nel 2020 in seguito allo shock causato dal Covid 19.

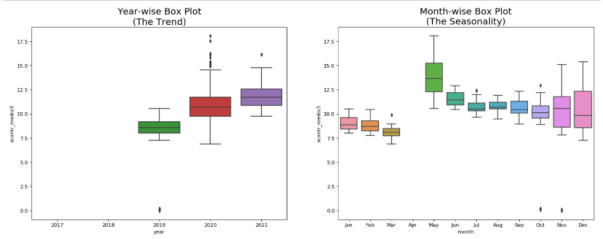

Figura 4: Boxplot terza serie temporale.

Osservando i boxplot riferiti al 4° ristorante in *Figura 5* si osserva un comportamento leggermente diverso, il picco è presente nel mese di giugno, probabilmente dovuto a fattori esterni dovuti alla natura e alle caratteristiche dell'esercizio in analisi.

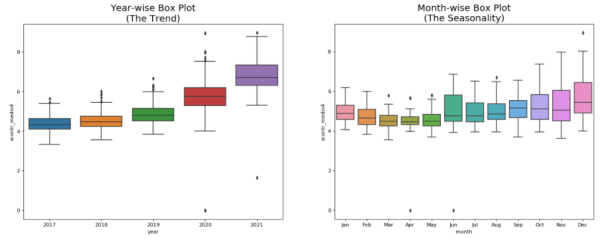

Figura 5: Boxplot quarta serie temporale.

Osservando invece in *Figura 6* i boxplot riferiti al 5° ristorante, anche in questo caso si osserva un comportamento simile ai restanti servizi.

I boxplot del 6° ristorante presenti in *Figura 7* sono anch'essi simili al comportamento dei precedenti servizi.

7.3 Modello SARIMA

Al fine di costruire un modello che riuscisse a definire tutte le componenti della serie temporale si è scelto di

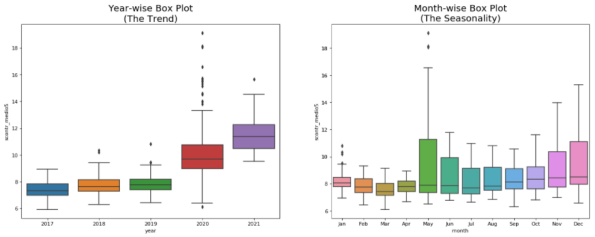

Figura 6: Boxplot quinta serie temporale.

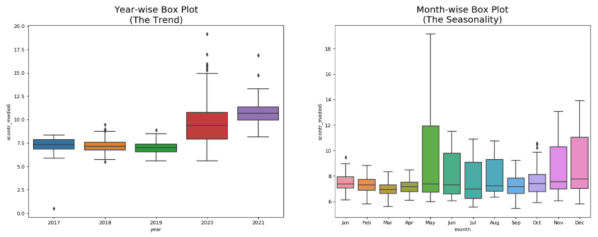

Figura 7: Boxplot sesta serie temporale.

utilizzare un modello SARIMA. Poichè, come visto precedentemente, i sei esercizi sembrano avere un andamento molto simile, si è deciso di utilizzare solo il primo ristorante.

Si è deciso di utilizzare come variabile dipendente della serie lo scontrino medio piuttosto che le vendite perché è una buona sintesi dei dati a disposizione e una misura meno distorta o influenzata da possibili outliers.

Si è dunque deciso di considerare il subset relativo al primo ristorante attraverso l'uso della libreria Pandas, e suddividere così il dataset in training e test set, al fine di sviluppare una previsione dei successivi 4 mesi.

7.3.1 Decomposizione serie temporale

Attraverso l'uso delle librerie:

- *matplotlib.pyplot*;
- *statsmodels.tsa.seasonal*;

si è inizialmente effettuata la decomposizione della serie temporale così costituita, di cui è possibile analizzare i risultati in *Figura 8*:

Dall'analisi di tali risultati è possibile infatti osservare la presenza di una stagionalità e un trend che presenta un pattern crescente in relazione allo shock causato dalle riaperture dopo il primo lockdown dovuto al Covid 19.

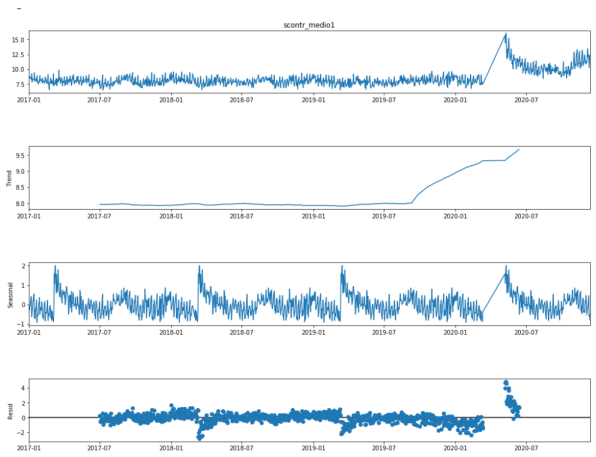

Figura 8: Decomposizione serie temporale.

7.3.2 Stazionarietà

Prima di procedere alla formulazione del modello, è necessario verificare se la serie sia stazionaria.

Per effettuare tale analisi è stata utilizzata la funzione *"adfuller"* della libreria *statsmodels.tsa.stattools*, implementando in particolare una funzione che effettuasse il test di Dickey-Fuller e determinasse la rolling statistics rappresentata in *Figura 9*.

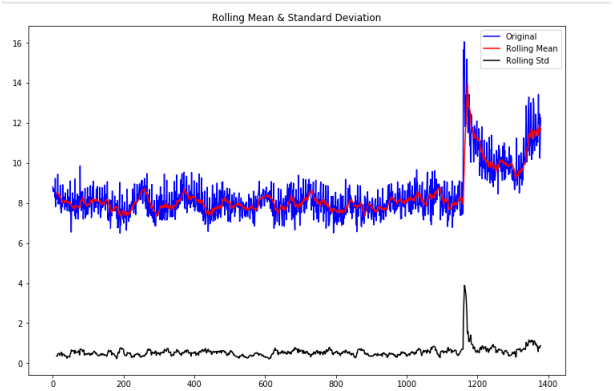

Figura 9: Media Mobile e Deviazione Standard.

La serie temporale d'origine presenta un p-value pari a 0.6273 con un valore della statistica test pari a -1.304175, per questo si considera la serie non stazionaria.

Viene effettuata una differenziazione di primo ordine, riconsiderando il test precedentemente fatto, la serie così definita presenta un p-value inferiore alla soglia scelta di 0.05 e un valore della statistica test pari a -1.114067e+01, così la serie differenziata viene ritenuta stazionaria.

In *Figura 10* è presente il plot della media mobile e della deviazione standard.

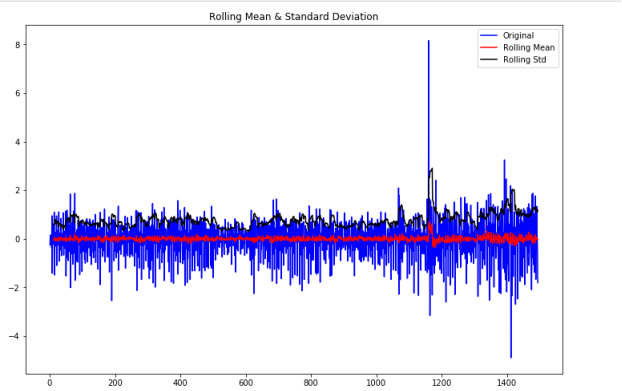

Figura 10: Media Mobile e Deviazione Standard.

7.3.3 Definizione del modello

Attraverso l'uso della libreria *statsmodel*, vengono effettuati i grafici ACF e PACF per la serie d'origine, di cui non stupiscono i risultati presentati in *Figura 11* e in *Figura 12* dopo i test appena effettuati.

Infatti nel grafico PACF sono presenti picchi positivi fino a 7, successivamente 14, 21 e proseguendo. Il grafico ACF presenta invece una situazione chiaramente non stazionaria.

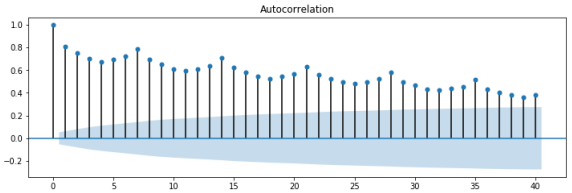

Figura 11: ACF serie temporale d'origine.

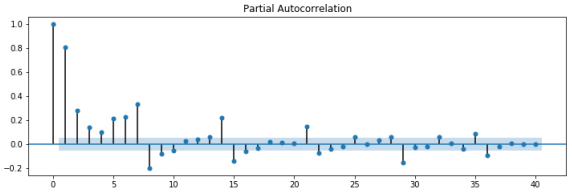

Figura 12: PACF serie temporale d'origine.

I grafici per la serie differenziata in *Figura 13* e *Figura 14* presentano picchi positivi significativi nel grafico ACF nei lag 1, 7, 14, 21. Nel grafico PACF ci sono picchi significativi nei lag fino a 7 proseguendo con 14, indicando la presenza di una stagionalità settimanale.

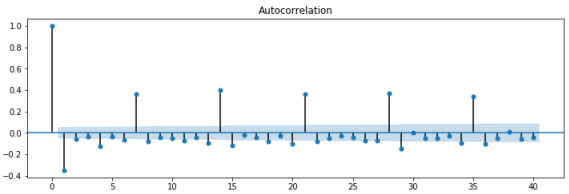

Figura 13: ACF serie temporale differenziata.

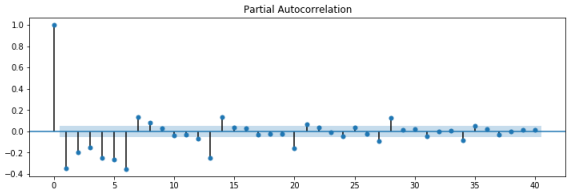

Figura 14: PACF serie temporale differenziata.

Congiuntamente all'analisi dei grafici ACF e PACF, per selezionare i parametri ottimali viene utilizzata la funzione *itertools* implementando il metodo del *gridsearch*, in cui attraverso l'analisi di tutte le combo possibili dei parametri, viene selezionata la combinazione che presenta il valore AIC minore.

7.3.4 Diagnostica del modello

Viene effettuata la diagnostica del modello al fine di valutarlo e attraverso l'analisi dei grafici in *Figura 15* si osserva come dal Q-Q plot, è possibile osservare una linea quasi dritta, solo alcuni punti si discostano dovuti probabilmente alla presenza di outliers, quindi non suggerisce un allontanamento sistematico dalla normalità.

Inoltre il correlogramma in basso a destra suggerisce che non vi sia una autocorrelazione dei residui.

L'istogramma mostra simmetria.

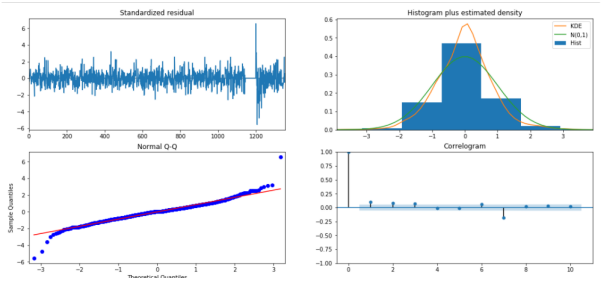

Figura 15: Diagnostica modello.

7.3.5 Previsione

Si procede dunque effettuando la previsione utilizzando la libreria *statsmodels* di cui è disponibile in *Figura 16* il grafico congiuntamente ai valori effettivi presenti nel test set.

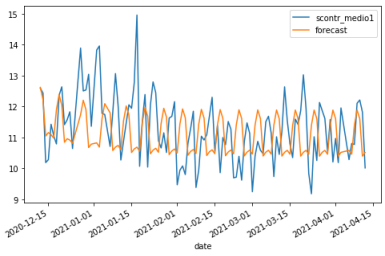

Figura 16: Forecast vs actual.

Attraverso la libreria *numpy*, viene definita una funzione che calcoli il valore MAPE (Mean Absolute Percentage Error) e SMAPE come misura per l'accuratezza previsionale.

Si osserva un valore MAPE pari circa al 15% indice di una buona previsione.

7.4 Exogenous Variables

Per tentare di migliorare il modello precedentemente costruito si è deciso di inserire delle variabili esterne all'interno del suddetto modello come definito precedentemente.

Il fine ultimo è stato quello di valutare quali variabili risultino influenti nella costruzione del modello.

Vengono considerate le variabili inserite all'interno del dataset durante la sezione di Data Preprocessing, ed in particolare le colonne:

- *'is_weekend';*
- *'is_holiday';*
- *'temp_media'.*

7.4.1 Variabile temperatura media

In prima istanza si è costruito un modello SARIMA utilizzando la sola variabile relativa alla temperatura media della regione circostante al primo ristorante considerato.

Dall'analisi del modello in *Figura 17* è possibile osservare che la variabile temp_media presenta un p-

value pari a 0.904, dunque risulta non statisticamente significativa.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.017e-06	1.95e+04	-5.21e-11	1.000	-3.83e+04	3.83e+04
temp_media	0.0007	0.006	0.121	0.904	-0.011	0.013
ar.L1	0.9793	0.037	26.710	0.000	0.907	1.051
ar.L2	-0.0545	0.027	-1.983	0.047	-0.108	-0.001
ma.L1	-0.5691	0.032	-18.037	0.000	-0.631	-0.507
ar.S.L7	0.0930	0.026	3.630	0.000	0.043	0.143
ar.S.L14	0.0624	0.030	2.072	0.038	0.003	0.121
ma.S.L7	-0.8809	0.019	-45.766	0.000	-0.919	-0.843
sigma2	0.3275	0.007	49.497	0.000	0.315	0.340

Figura 17: Modello 1 con la variabile temp_media.

Si è deciso di effettuare tale analisi costruendo un modello SARIMA che modellasse una serie temporale basata questa volta sullo scontrino medio del secondo ristorante. Il modello così definito presenta la variabile temp_media con un p-value pari a 0.787.

Effettuando tale analisi su tutti gli altri ristoranti si riassumono così i risultati trovati:

- *'temp_media' nel 1° modello presenta un p-value pari a 0.904;*
- *'temp_media' nel 2° modello presenta un p-value pari a 0.787;*
- *'temp_media' nel 3° modello presenta un p-value pari a 0.104;*
- *'temp_media' nel 4° modello presenta un p-value minore della soglia;*
- *'temp_media' nel 5° modello presenta un p-value pari a 0.582;*
- *'temp_media' nel 6° modello presenta un p-value pari a 0.069.*

E' possibile dunque affermare che la variabile temperatura media non risulta significativa, considerando un livello di significatività $\alpha = 5\%$.

Per questo motivo si è proceduto alla formulazione di un modello per che non contenga tale variabile.

7.4.2 Variabili *is_weekend* e *is_holiday*

Sono state così considerate le variabili booleane che indicano se i giorni considerati siano giorni appartenenti al weekend, o a giorni festivi secondo il calendario italiano.

Si è creato un subset che contenesse tali variabili per poterlo inserire come input nel modello, dalla cui analisi si osserva in *Figura 18* che entrambe le variabili presentano un p-value inferiore alla soglia di 0.05, risultando dunque statisticamente significative.

	coef	std err	z	P>|z|	[0.025	0.975]
is_weekend	4.8415	0.082	58.765	0.000	4.680	5.003
is_holiday	-0.5344	0.075	-7.164	0.000	-0.681	-0.388
ar.L1	1.0113	0.036	27.890	0.000	0.940	1.082
ar.L2	-0.0109	0.036	-0.300	0.764	-0.082	0.060
ma.L1	-0.5758	0.037	-15.650	0.000	-0.648	-0.504
ma.L2	-0.4189	0.035	-11.941	0.000	-0.488	-0.350

Figura 18: Modello con la variabile is_weekend, is_holiday

Rieffettuando tali analisi anche sugli altri ristoranti si giunge a questi risultati:

- *'is_weekend' nel 1° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_holiday' nel 1° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_weekend' nel 2° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_holiday' nel 2° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_weekend' nel 3° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_holiday' nel 3° modello presenta un p-value pari a 0.209;*
- *'is_weekend' nel 4° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_holiday' nel 4° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_weekend' nel 5° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_holiday' nel 5° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_weekend' nel 6° modello presenta un p-value inferiore alla soglia di 0.05;*
- *'is_holiday' nel 6° modello presenta un p-value inferiore alla soglia di 0.05;*

Possiamo dunque affermare che una variabile statisticamente non significativa in uno dei sei servizi commerciali tende a non essere statisticamente significativa anche per gli altri.

7.5 Modello VAR (Vector Autoregression)

In questa sezione si è scelto di utilizzare un modello VAR (Vector Autoregression) al fine di prevedere congiuntamente lo scontrino medio dei singoli ristoranti. Nel modello la serie temporale è modellata come una combinazione lineare dei suoi lag, cioè i valori passati della serie vengono utilizzati per prevedere il presente e il futuro.

In questa sezione sono state utilizzate le librerie Python:

- Pandas;
- Numpy;
- matplotlib.python;
- Statsmodel.

di cui in particolare sono state utilizzate le funzioni:

- VAR;
- adfuller;
- rmse;
- aic.

E' stato considerato il sottoinsieme del dataset originale costituito dagli scontrini medi relativi ad ogni ristorante; anche in questo caso si è deciso di utilizzare il valore scontrino medio come variabile dipendente.

Poiché alcuni ristoranti presentano una data d'apertura successiva rispetto ad altri, al fine di evitare difformità si è deciso di considerare la serie temporale che abbia come primo data point il giorno in cui ogni ristorante risulta essere aperto.

Si è deciso di utilizzare come indice dei dati la colonna *'date'*.

In *Figura 19* è possibile osservare le prime cinque colonne del dataset così costituito:

date	scontr_medio1	scontr_medio2	scontr_medio3	scontr_medio4	scontr_medio5	scontr_medio6
2019-10-16	8.144181	9.853279	0.230000	4.432695	7.998927	6.685150
2019-10-17	7.445434	10.150817	0.076667	5.016006	7.804907	6.471310
2019-10-18	8.552256	9.866581	0.230000	5.267949	8.523818	6.850280
2019-11-04	7.949529	9.521661	0.000000	4.795881	7.421092	6.360165
2019-11-06	7.545577	9.638777	0.115000	4.974365	7.809680	6.975851

Figura 19: Prime 5 righe del dataset.

Si è effettuata una iniziale visualizzazione delle serie temporali in *Figura 20*.

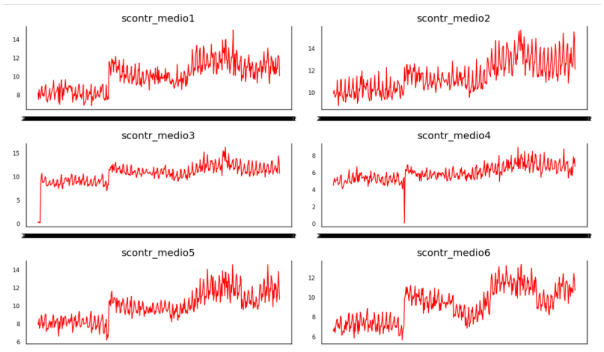

Figura 20: Visualizzazione delle 6 serie temporali.

Dalla figura è possibile osservare come ogni serie presenti andamenti piuttosto simili, ad eccezione del quarto ristorante in cui è presente una caduta dello scontrino medio in relazione alle prime aperture dopo il lockdown verificatosi nel 2020.

7.5.1 Test causalità di Granger

Successivamente è stato utilizzato il test di causalità di Granger.

Il test di causalità di Granger permette di testare la relazione presente tra le varie serie temporali prima di procedere alla costruzione del modello stesso.

L'ipotesi nulla considerata è che i coefficienti dei valori passati nell'equazione di regressione siano zero, ossia i valori passati di una serie temporale X non sono dipendenti dai valori di una serie Y, per cui si decide di considerare la soglia di significatività del 5%.

Si è dunque importata la funzione *grangercausalitytests* della libreria statsmodels in cui vengono testate tutte le possibili combinazioni della matrice di output.

	scontr_medio1_x	scontr_medio2_x	scontr_medio3_x	scontr_medio4_x	scontr_medio5_x	scontr_medio6_x
scontr_medio1_y	1.0000	0.0	0.0000	0.0001	0.0	0.0
scontr_medio2_y	0.0000	1.0	0.0000	0.0000	0.0	0.0
scontr_medio3_y	0.0000	0.0	1.0000	0.0000	0.0	0.0
scontr_medio4_y	0.0000	0.0	0.0000	1.0000	0.0	0.0
scontr_medio5_y	0.0161	0.0	0.0001	0.0000	1.0	0.0
scontr_medio6_y	0.0001	0.0	0.0000	0.0000	0.0	1.0

Figura 21: Matrice Granger causality test.

In *Figura 21* è possibile osservare come le variabili non presentino una forte correlazione tra loro, tuttavia si è deciso comunque di proseguire con il modello per riuscire ad effettuare una predizione con più informazioni possibili.

7.5.2 Cointegration test

Successivamente viene svolto il Cointegration test, che aiuta a stabilire se esista una connessione statisticamente significativa tra due o più serie temporali.

In particolare, se esiste una combinazione lineare delle serie temporali con un ordine di integrazione (d) inferiore a quello delle singole serie, allora l'insieme delle serie risulta cointegrato.

Per svolgere il test è sufficiente implementare la libreria statsmodels, definendo una funzione chiamata "cointegration_test" di cui è possibile analizzare i risultati in *Figura 22*

Per definire la funzione è stata utilizzata una procedura sviluppata da Soren Johanssen (Estimation and Hypothesis testing of cointegration vectors in gaussian vector autoregressive Models, 1991).

```
Name   ::  Test Stat > C(95%)    =>   Signif
 ----------------------------------------
scontr_medio1 ::  249.41    > 83.9383   =>   True
scontr_medio2 ::  120.2     > 60.0627   =>   True
scontr_medio3 ::  61.78     > 40.1749   =>   True
scontr_medio4 ::  32.65     > 24.2761   =>   True
scontr_medio5 ::  10.08     > 12.3212   =>   False
scontr_medio6 ::  0.28      > 4.1296    =>   False
```

Figura 22: Cointegration test.

Il modello VAR che si procederà a costruire sarà costituito solo dai primi 4 ristoranti escludendo il 5 e 6 poiché dal test risultano non significativi. Il modello sarà applicato su un training set e successivamente utilizzato per effettuare una previsione delle successive 15 osservazioni.

Tali previsioni saranno poi comparate utilizzando le effettive previsioni presenti in un test set attraverso alcune metriche per l'accuratezza che saranno successivamente sviluppate.

7.5.3 Stazionarietà

Si osservi che il modello VAR richiede che le serie temporali di cui si intende svolgere una previsione siano stazionarie, si procede a verificare se le serie temporali di interesse siano stazionarie o meno.

Qualora una serie temporale non dovesse essere stazionaria, si procede differenziando una volta la serie e si ripete di nuovo il test finché tale serie non diventa stazionaria.

Effettuando la differenziazione si riduce la lunghezza della serie di 1; poiché nel modello considerato è necessario che tutte le serie abbiano la stessa lunghezza, nel caso in cui una serie risultasse non stazionaria, si procederà differenziando tutte le serie temporali.

Per svolgere il test di stazionarietà viene definita la funzione *"adfuller_test()"* in cui vengono scritti i risultati del test ADF per ogni serie temporale data, così da poter richiamare la funzione per ciascuna serie.

Il test ADF conferma che nessuna serie risulta stazionaria, si procede differenziando ciascuna di esse così da richiamare nuovamente il test.

Dopo la prima differenziazione ogni serie risulta stazionaria.

7.5.4 Definizione modello

A questo punto è necessario selezionare il giusto ordine (P) del modello VAR, vengono fittati iterativamente ordini crescenti di modelli VAR per scegliere l'ordine che fornisca un modello con il minor AIC possibile.

Viene dunque richiamata la funzione VAR, e attraverso i risultati presenti in *Figura 23* è possibile osservare come l'AIC presenta un minimo al lag 8.

Si decide di procedere con il modello lag 8.

Si presenta la summary dei risultati della regressione in *Figura 24*

7.5.5 Test di Durbin Watson

Dopo aver definito il modello, si verifica l'eventuale presenza di correlazione tra i residui utilizzando la statistica di Durbin Watson.

Il controllo della correlazione viene effettuato per garantire che il modello sia sufficientemente in grado di spiegare le varianze e i modelli delle serie temporali.

In *Figura 25* è possibile osservare che i valori della statistica risultano appartenere ad un intorno di 2, valori che sono vicini al valore due infatti implicano che non vi sia alcuna correlazione.

7.5.6 Previsione

Si procede allora con la previsione utilizzando la libreria *exstatsmodels*.

Al fine di effettuare la previsione viene considerato il numero di lag order, i termini del modello VAR sono

```
Lag Order = 1
AIC :  -2.3072630785299473
BIC :  -2.113480133325079
FPE :  0.09953375481867134
HQIC:  -2.2306418625571856

Lag Order = 2
AIC :  -2.614411514643268
BIC :  -2.2649704863822917
FPE :  0.07321285138718413
HQIC:  -2.476230245110117

Lag Order = 3
AIC :  -2.6949341578376176
BIC :  -2.1892698326228435
FPE :  0.06755240961618486
HQIC:  -2.494957431482596

Lag Order = 4
AIC :  -2.8637678039675762
BIC :  -2.201311541688278
FPE :  0.05706401135277492
HQIC:  -2.601758770565719

Lag Order = 5
AIC :  -3.4882432709401128
BIC :  -2.6684229764125393
FPE :  0.030565309513105512
HQIC:  -3.163963621121432

Lag Order = 6
AIC :  -3.881009035589089
BIC :  -2.9032491293856255
FPE :  0.02064241075759221
HQIC:  -3.4942189885270762

Lag Order = 7
AIC :  -4.072443905900488
BIC :  -2.93616529486388
FPE :  0.01705183574630946
HQIC:  -3.6229021968689663

Lag Order = 8
AIC :  -4.093208992601859
BIC :  -2.7978290400403676
FPE :  0.016709083079081397
HQIC:  -3.5806728603984994

Lag Order = 9
AIC :  -4.087836599080188
```

Figura 23: AIC.

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Sun, 27, Jun, 2021
Time:                     20:15:53
--------------------------------------------------------------------
No. of Equations:         4.00000    BIC:                   -2.79783
Nobs:                     409.000    HQIC:                  -3.58067
Log likelihood:          -1352.32    FPE:                  0.0167091
AIC:                     -4.09321    Det(Omega_mle):       0.0122506
```

Figura 24: Summary del risultato.

```
scontr_medio1 : 1.99
scontr_medio2 : 2.02
scontr_medio3 : 2.01
scontr_medio4 : 2.04
```

Figura 25: Test di Durbin-Watson.

essenzialmente i lag delle varie serie temporali nel dataset, quindi è necessario fornire come input tanti valori precedenti quanti indicati dall'ordine dei lag usato nel modello.

La forecast così generata presenta la stessa scala del training set usato dal modello, dunque è necessario de-differenziare tante volte quanto sono stati differenziati i dati di input.

Nel caso considerato è necessario de-differenziare una volta.

Avendo riportato le previsioni alla loro scala originale, si procede plottando in *Figura26* e *27* le previsioni rispetto ai valori attuali presenti nel test set.

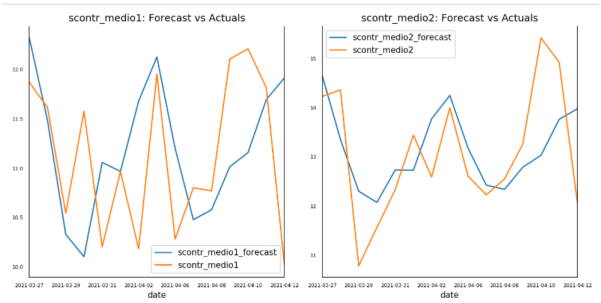

Figura 26: Forecast vs Actual.

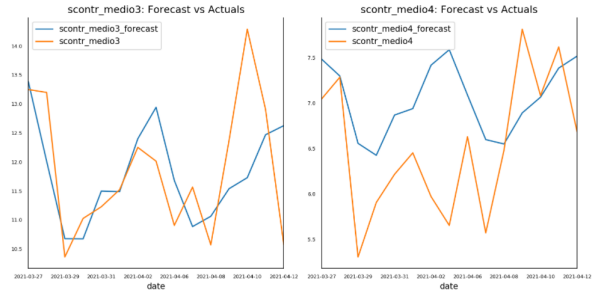

Figura 27: Forecast vs Actual.

7.5.7 Valutazione previsione

Per valutare le previsioni, vengono calcolate un insieme di metriche, in particolare MAPE, ME, MAE, MPE, RMSE, corr e minmax, di cui vengono presentati i risultati in *Figura 28*.

Analizzando i risultati ottenuti, è dunque possibile osservare come la previsione ottenuta sia più che buona.

```
Forecast Accuracy of: scontr.1
mape :  0.0637
me :  0.0799
mae :  0.6922
mpe :  0.0113
rmse :  0.9065
corr :  0.1698
minmax :  0.0594

Forecast Accuracy of: scontr.2
mape :  0.0658
me :  0.0677
mae :  0.8604
mpe :  0.0114
rmse :  1.0713
corr :  0.5239
minmax :  0.0622

Forecast Accuracy of: scontr.3
mape :  0.062
me :  -0.0643
mae :  0.7465
mpe :  0.0001
rmse :  1.018
corr :  0.4828
minmax :  0.0589

Forecast Accuracy of: scontr.4
mape :  0.1127
me :  0.5301
mae :  0.6863
mpe :  0.0926
rmse :  0.8709
corr :  0.3604
minmax :  0.0959
```

Figura 28: Metriche per l'accuratezza.

8 Risultati

Come si vede dalla figura il modello Sarima ottenuto *Figura 29* ha un valore MAPE pari al 15% indice di una buona previsione.

```
MAPE: 15.82 %
SMAPE: 16.90 %
```

Good MAPE

Figura 29: MAPE forecast con SARIMA.

Per quanto riguarda lo studio della significatività delle variabili esogene aggiunte "is_weekend" "is_holiday" e "temp_media", solo le prime due sono risultate significative con un p-value pari a 0.000 e un livello di significatività dello 0.01 nell figura 18.

Inoltre per lo studio della dipendenza e la costruzione del modello VAR, come si vede nella figura 21, il test di causalità di Granger evidenzia una debole correlazione tra le variabili.

Attraverso il Cointegration test si osserva come solo i primi 4 ristoranti siano cointegrati, per questo motivo si procede alla formulazione di un modello che consideri solamente questi ristoranti.

Le previsioni risultanti ottenute con il modello VAR, considerando come misura di accuratezza MAPE risultano buone, come si vede nelle figure 27, 28.

Dall'analisi fatta si è notato che il comportamento delle persone nell'andare al ristorante ha un forte carattere stagionale, infatti ciascuna serie temporale presenta dei picchi nei weekend, indipendentemente dai mesi anche se prediligono il periodo di Maggio, Giugno e Dicembre.

Non si notano invece picchi in estate inoltrata, ovvero nei mesi di Luglio e Agosto, dovuto probabilmente alle vacanze.

Inoltre, si è visto che la scelta delle persone di recarsi o meno al ristorante non risente del meteo, evidentemente i ristoranti presi in considerazione hanno ampi spazi interni e quindi non sono condizionati dalle condizioni atmosferiche.

Si è anche constatato che dopo il periodo di chiusura totale, vi è stata una forte ripresa: le persone dopo essere state costrette a restare a casa per un periodo

così lungo, hanno sentito la mancanza di uscire e evidentemente anche di andare a mangiare al ristorante. Come seconda domanda di ricerca, è stato importante poter confrontare il comportamento del singolo ristorante con gli altri perchè si è dimostrato come lo scontrino medio non risultasse eccessivamente influenzato dalla concorrenza ed inoltre, considerare un modello che utilizzasse i dati di più i ristoranti ha permesso di ottenere buone previsioni, come si evince dai risultati della misura di accuratezza MAPE.

9 Conclusione e possibili sviluppi

In questo report si sono ricercati gli effetti che la chiusura totale delle attività iniziata a Marzo 2020 ha avuto sul settore della ristorazione, analizzando quello che è stato il comportamento delle persone nel periodo in esame e le variabili che possono influenzare l'aumento delle vendite ed infine, indagare l'esistenza circa la dipendenza tra i ristoranti considerati e la possibilità di effettuare una previsione di ciascun scontrino medio considerando i restanti dati a disposizione.

Una futura ricerca potrebbe concentrarsi sul periodo post lockdown, in particolare, comprendere se la forte ripresa sia stata dovuta ai continui periodi di chiusura e quindi le vendite e gli scontrini siano ritornati ai valori del periodo che ha preceduto il lockdown di Marzo 2020, oppure se la situazione si è stabilizzata sui risultati.

Inoltre, si potrebbe formalizzare una futura ricerca studiando i possibili motivi che rendono i ristoranti più o meno influenzati da altri esercizi.

Infine, avendo ulteriori informazioni sulle tipologie di esercizi si potrebbero inserire nei modelli di previsione altre variabili così da avere risultati più accurati.

Riferimenti bibliografici

[1] Fattore M.(2020) Fundamentals of time series analysis, for the working data scientist (DRAFT)

[2] https://ichi.pro/it/come-prevedere-le-vendite-con-python-utilizzando-il-modello-sarima-204921640254077 02/06/2021

[3] https://towardsdatascience.com/time-series-forecasting-with-a-sarima-model-db051b7ae459 28/05/2021

[4] https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/ 28/05/2021

[5] https://www.ilmeteo.it/meteo/Piacenza 25/05/2021

[6] https://www.kaggle.com/poiupoiu/how-to-use-sarima 29/05/2021

[7] https://www.kaggle.com/sumi25/understand-arima-and-tune-p-d-q 04/06/2021

[8] Johanssen S. (1991) Estimation and Hypothesis testing of cointegration vectors in gaussian vector autoregressive Models