Analisi e previsioni su serie storiche: serie storica riguardante le vendite di un e-commerce.

Annamaria Teodori (matr. 889903)

Michele Gazzola (matr. 825267)

Giuseppe Napoli (matr. 888270)

Corso di Laurea Magistrale in Data Science, Dipartimento di Informatica, Sistemistica e Comunicazione DISCo. Università degli Studi di Milano Bicocca. Viale Sarca, 336 - 20126, Milano

Data consegna: 02/01/2023

1 Introduzione.

L'oggetto della nostra analisi coinvolge una serie storica riguardante il fatturato relativo a diverse categorie di vendita di un e-commerce. Il dataset per sua natura impone di lavorare su due piani: il piano del valore di fatturato globale ed il piano del fatturato relativo alle singole categorie che compongono la serie storica.
Abbiamo in primo luogo definito una seria storica temporale a frequenza mensile attraverso l'operatore somma. I dati disponibili sono infatti relativi al singolo giorno, tuttavia all'interno di ciascun mese non è presente lo stesso range temporale (alcuni mesi registrano il fatturato per soli 10 giorni piuttosto che 30 giorni) per cui per poter definire una frequenza mensile, è stato aggregato il dato giornaliero.
Successivamente, una volta indagate le componenti della serie storica e impostate le dovute trasformazioni a fronte di quanto emerso, abbiamo operato sul delta atteso tra t+1 e t, selezionando i parametri ottimi di un modello ARMA. Abbiamo inoltre esaminato la serie storica in termini di volatilità, selezionando in questo caso i parametri ottimi di un modello GARCH, in modo da cogliere i raggruppamenti (o cluster) di volatilità osservati ed effettuare una previsione futura.
Da ultimo abbiamo definito tre serie storiche per tre diverse categorie, calcio casual e fitness, selezionate in base alla frequenza di osservazione in quanto hanno numerosità del fatturato pressoché uguale (il calcio ha 2.956 osservazioni, il settore casual 2.901 e il fitness 2.834). Su queste tre serie storiche abbiamo provato a modellare, tramite le copule, la struttura e la tipologia di dipendenza, in quanto è potenzialmente interessante per un e-commerce osservare la relazione che sussiste tra il fatturato registrato da diverse categorie.

2 Descrizione del dataset.

La versione del dataset è composta da 3 colonne e 25.261 righe contenenti per ognuna delle 3 colonne: data di rilevazione, fatturato e categoria, nel periodo compreso tra l' 1/02/2013 e l' 8/04/2022. Il dataset è composto da 30 diverse categorie che hanno la seguente numerica totale di osservazioni nel periodo considerato:

Categoria	Numerica osservazioni
Arciera	7
Arti marziali	139
Bambino	1.250
Baseball	60
Basket	922
Buoni / acconti	7
Calcio	2.956
Casual	2.901
Ciclismo	759
Danza	53
Fitness	2.834
Freccette	10
Golf	31
Intimo	4
Mare	1.009
Nuoto	949
Padel	191
Pattini	154
Pesca	2.978
Ping-pong	105
Rugby	73
Running	2.271
Sci	1.209
Skateboard	94
Snowboard	1.372
Soft air	1
Subacquea	162
Tennis	972
Trekking	1.514
Volley	274

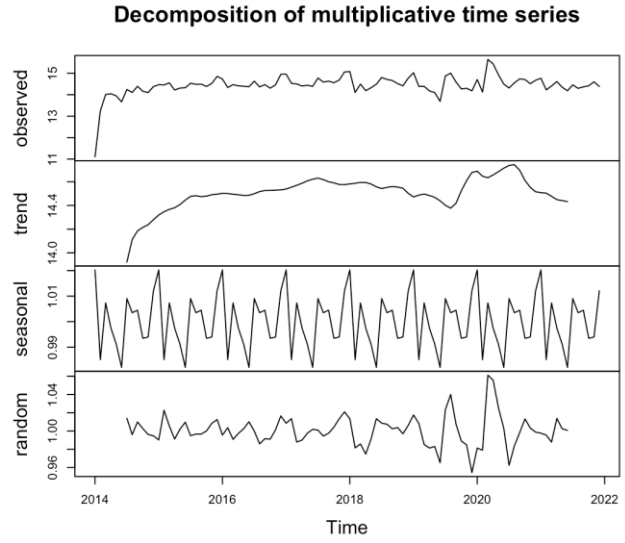
3 Data cleaning.

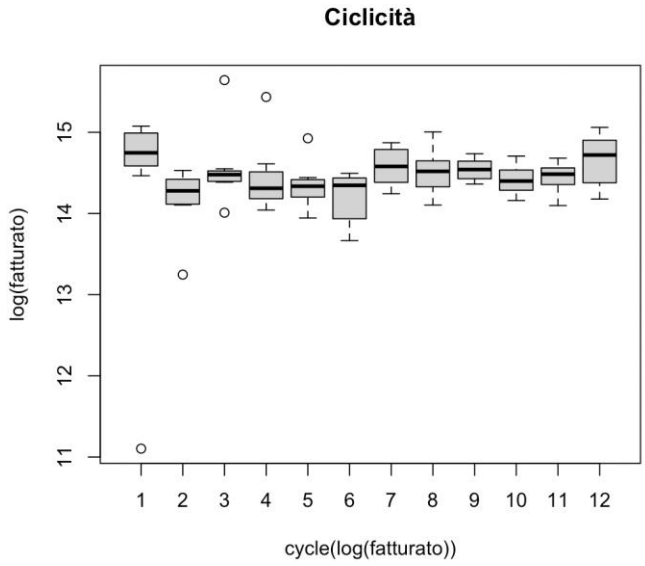
In fase preliminare si è proceduto a modellare il fatturato secondo un oggetto “time series” a frequenza mensile. Per far ciò è stata effettuata la somma delle osservazioni giornaliere per ogni mese di ogni anno, in modo da ottenere il fatturato totale mensile per ogni anno. Si è scelto inoltre di non coinvolgere nell’analisi l’anno 2022 e 2013 in quanto presentano dati giornalieri relativi a solo 4 mesi per il 2022 e 8 mesi per il 2013.
Successivamente a tali step è stato possibile ottenere una serie storica temporale con inizio al

01/2014 e fine al 12/2021, con frequenza di 12 mesi.

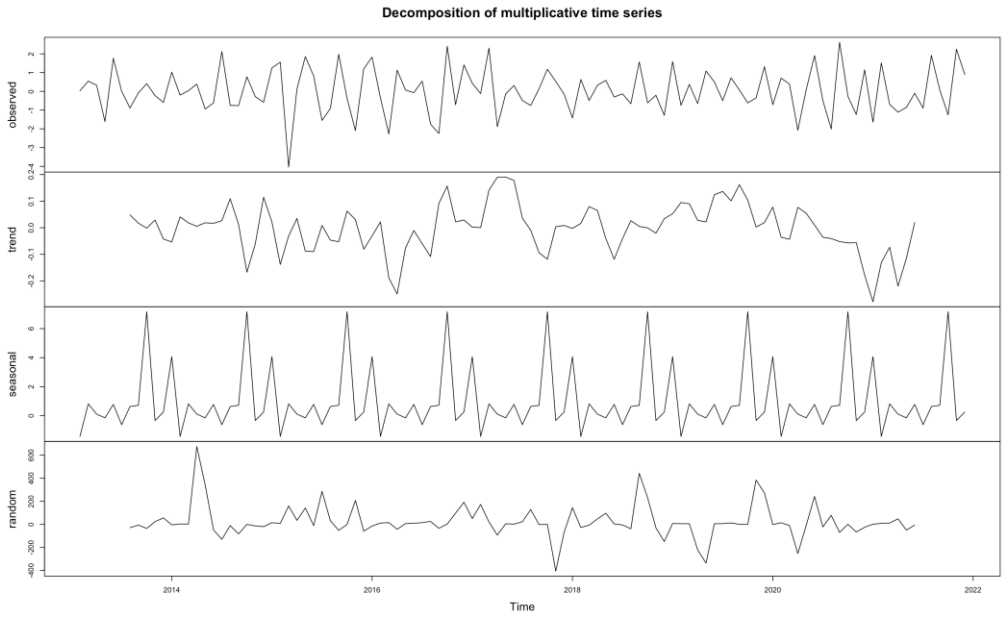
4 Data exploration.

Uno degli scopi fondamentali dell'analisi classica delle serie temporali è quello di scomporre la serie nelle sue componenti: trend, stagionalità, ciclicità e componente residua o erratica.
Esse possono essere legate tra loro in modo additivo **Yt=Tt + Ct + St + Et ,** oppure in modo moltiplicativo **Yt=Tt* Ct * St * Et**. Un modello di tipo moltiplicativo può essere facilmente trasformato in un modello additivo usando l'operatore logaritmo:
log(Yt)=log(Tt)+ log(Ct)+ log(St)+ log(Et).
Si è scelto di indagare le componenti della serie storica imputando un modello moltiplicativo, utilizzando l'operatore logaritmo.

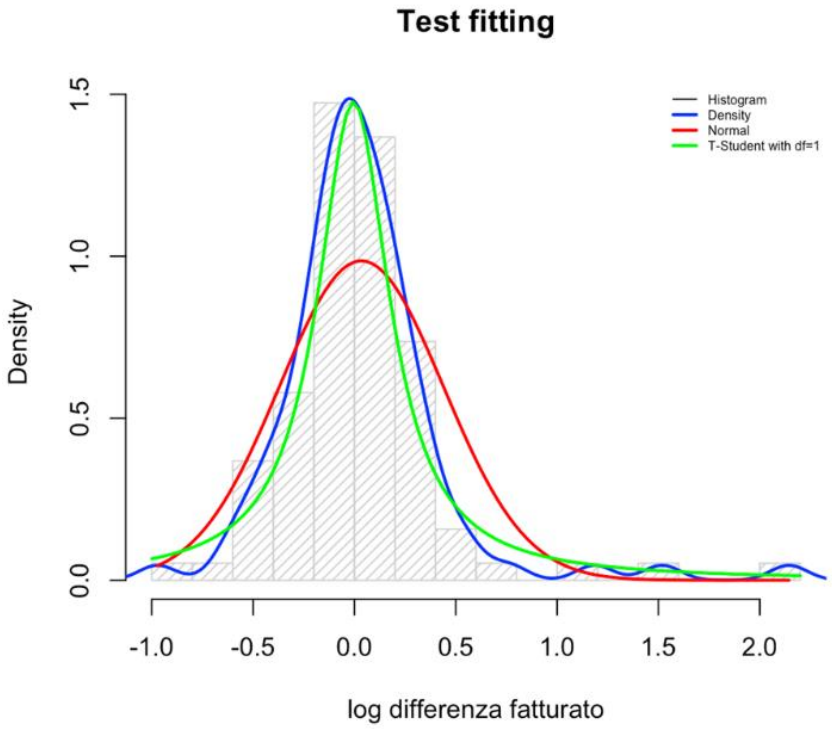




Dai grafici di cui sopra si osserva che la serie storica in questione non è stazionaria, in quanto presenta un trend crescente tra il 2014 e 2019, decrescente tra il 2019 e 2020, nuovamente crescente tra il 2020 e 2021 e decrescente tra il 2021 e 2022. È possibile, inoltre, notare una componente di stagionalità piuttosto marcata (come da box plot) ed un clustering di volatilità, ovvero la componente erratica che presenta momenti caratterizzati da volatilità maggiore seguiti da momenti di volatilità inferiore.
Un metodo per eliminare il trend, è quello di operare sulle differenze tra i termini (o i logaritmi dei termini in caso di modello moltiplicativo) della serie storica: le differenze del primo ordine rimuovono un trend lineare, quelle del secondo ordine un trend parabolico, quelle di ordine k rimuovono un trend polinomiale di grado k: **log(Δ t) =log(Yt+1) - log(Yt)**
In generale le differenze logaritmiche sono particolarmente utilizzate nell'ambito delle serie storiche, poiché sono legate al concetto di tasso medio di crescita.
Il tasso medio di crescita si ottiene partendo dalla media geometrica dei fattori di crescita.
Si è quindi deciso di operare sulle differenze logaritmiche del primo ordine proprio in modo da studiare il comportamento tra t+1 e t.
Di seguito le caratteristiche della serie storica secondo la trasformazione effettuata:



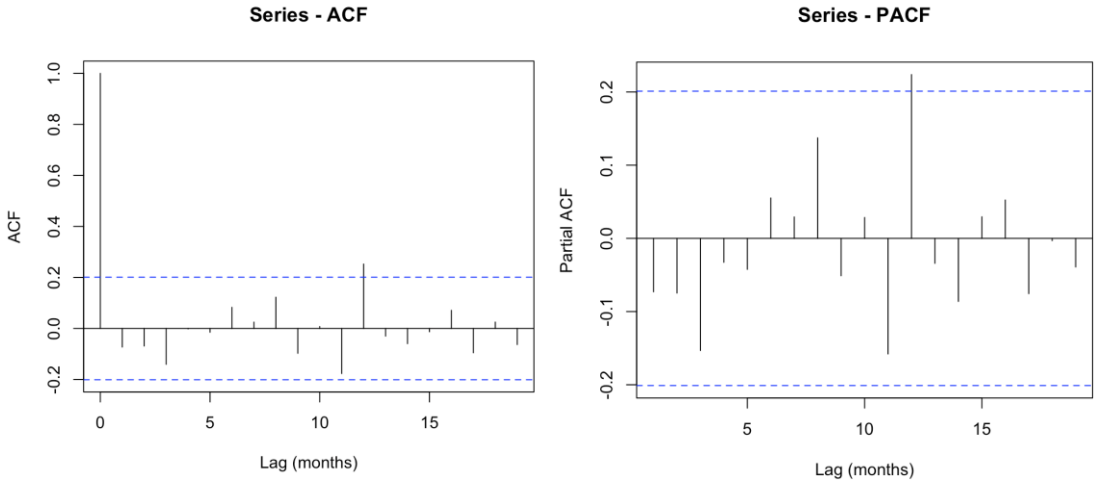
In questa decomposizione, si coglie maggiormente il clustering di volatilità presente ed è possibile notare come sia stato eliminato il trend. Una volta indagate le caratteristiche della serie storica, si analizza la distribuzione delle differenze logaritmiche, in modo da cogliere il comportamento delle code ovvero la volatilità e possibile eteroschedasticità.
Si testa il fitting della distribuzione kernel, T student e Normale:



Dai risultati ottenuti è possibile cogliere come la distribuzione T student, approssimi meglio le code più spesse e evidenzi una forte violazione dell'assunzione di normalità che traspare dagli indici SK (2.020222) e KU (11.56405), in particolare generata dalle code spesse e da un'asimmetria a sinistra. La distribuzione normale sottostima la frequenza di eventi estremi. In generale non esiste una distribuzione migliore di un'altra, ma la scelta varia in base alla finalità dell'analisi. Per una finalità di tipo descrittivo il grafico suggerirebbe la scelta di una T student in quanto approssima bene le code più spesse, se invece si vuole enfatizzare maggiormente il dato campionario può avere maggior senso supporre una distribuzione kernel.

5 Analisi della funzione di autocorrelazione e previsione con modello ARMA.

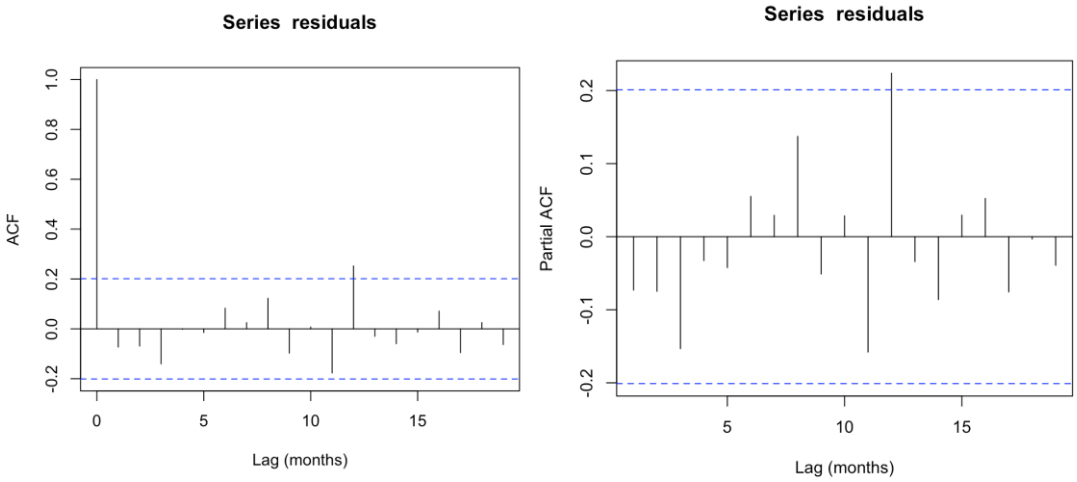
Analizzando le funzioni di autocorrelazione delle differenze logaritmiche di primo ordine del fatturato, si riscontra l'assenza di correlazione con il proprio passato con qualche eccezione (ritardo 12 e 0), il modello ACF mostra come ritardo anche lo 0 per il semplice fatto che, a differenza del modello PACF, il modello ACF parte dallo zero e non da 1 per cui è normale che lo zero venga considerato sempre come ritardo visto che la correlazione della serie storica con sé stessa è 1.



La serie storica mostra spesso periodi con alta o bassa concentrazione di volatilità. Per questa tipologia di serie storiche, una volatilità che cambia nel tempo è molto più frequente di una volatilità costante, motivo per cui la previsione con un modello ARMA generalmente ha poca significatività rispetto a delle stime ottenute con un modello GARCH.
Si è proceduto successivamente, in ogni caso, a stimare un modello ARMA selezionando il numero delle componenti auto-regressive e di media mobile più opportune tramite il criterio di informazione BIC. L'analisi effettuata ha condotto alla scelta del modello ARMA (0,0).
Il modello, quindi, non contiene una componente auto regressiva, ovvero le differenze logaritmiche non sono correlate al proprio passato né ad una media mobile. Un modello ARMA (0,0) è l'equivalente di un modello MA(0) e AR(0) ed è rappresentato dalla seguente equazione **xt=c+et**.
Quindi è composto da due parti: la costante e il termine di errore.
Di seguito si riporta il test sulla significatività delle stime dei coefficienti ottenute e i correlogrammi degli errori:

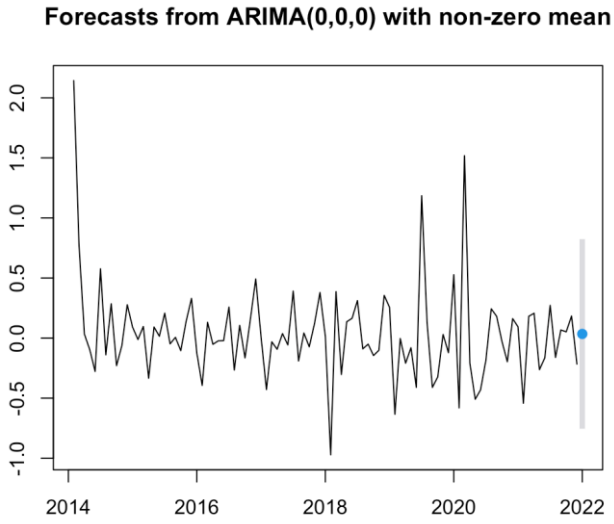
```
z test of coefficients:

          Estimate Std. Error z value Pr(>|z|)
intercept 0.034510   0.041295  0.8357   0.4033
```



Nonostante il criterio di selezione dei parametri BIC individui come modello migliore l'ARMA (0,0), la stima dell'intercetta non risulta significativa, mentre per quando riguarda i residui del modello non risulta essere presente correlazione con il proprio passato se non per qualche eccezione. Per quanto riguarda la previsione 1 passo avanti ovvero per gennaio 2022 si ottiene la seguente stima con i relativi intervalli di confidenza:

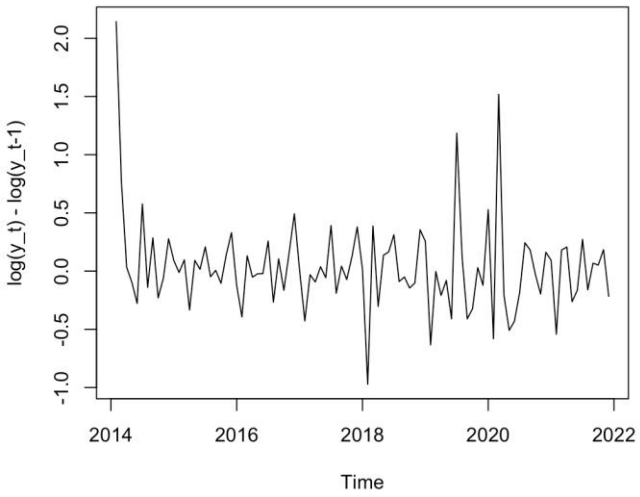
	Point Forecast	Lo 95	Hi 95
Jan 2022	0.03451038	-0.7543524	0.8233731



La stima trovata rientra perfettamente nel rispettivo intervallo di confidenza stimato al 95% e segnala una crescita nulla del fatturato per gennaio 2022 rispetto al mese precedente. Non si prevede quindi né un aumento né una diminuzione del fatturato rispetto a quanto registrato l'ultimo periodo di osservazione, ma piuttosto un andamento costante rispetto a quest'ultimo. In generale il modello ARMA è un metodo per modellare linearmente i dati e la larghezza della previsione rimane costante perché il modello non riflette le modifiche recenti o incorpora nuove informazioni.

6 Analisi della volatilità con un modello GARCH.

Si è analizzato il grafico della serie storica della differenza dei logaritmi del fatturato osservando quanto detto in precedenza, ovvero a momenti caratterizzati da volatilità inferiore seguono momenti di volatilità maggiore:



Al fine di modellare la volatilità della serie storica si è proceduto alla stima di un modello GARCH scegliendo come distribuzione condizionata per i termini di errore la T student (si è infatti verificato in precedenza che i termini di errore non seguono una distribuzione normale). Per trovare i parametri più appropriati del modello GARCH è stata utilizzata la funzione "autoGARCH", per la precisione questa funzione permette di ottenere i parametri più appropriati **m,n,p,q** di un modello ARMA(m,n)-GARCH(p,q). Il modello più adeguato sarà quello che rappresenterà il valore più basso dell'AIC (Akaike information criterion). La funzione suggerisce come il miglior modello sia l'ARMA(5,1)-GARCH(1,1). Le specifiche del modello da ottimizzare vengono poi definite con la funzione uGARCHspec al quale forniamo i parametri ottimi trovati e un modello eGARCH ovvero un GARCH esponenziale. Il modello eGARCH, infatti, offre stime dei parametri significative rispetto al modello GARCH standard (sgarch). Di seguito sono riportati i parametri ottimali stimati:

Optimal Parameters

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.000964	0.000001	1802.7	0
ar1	0.398820	0.000080	4954.7	0
ar2	-0.177803	0.000042	-4243.0	0
ar3	-0.140087	0.000035	-4029.3	0
ar4	0.070939	0.000017	4191.1	0
ar5	-0.008096	0.000003	-3201.8	0
ma1	-0.821998	0.000131	-6284.9	0
omega	-0.135937	0.000025	-5426.0	0
alpha1	-0.181791	0.000042	-4364.6	0
beta1	0.956075	0.000134	7146.8	0
gamma1	-0.503826	0.000094	-5364.9	0
shape	3.331875	0.000616	5409.7	0

L'analisi dei t values e delle probabilità associate ci permette di concludere che i parametri stimati sono tutti significativi oltre ogni ragionevole intervallo di confidenza. Va notato inoltre che beta1 è pari a 0,95 e questo implica una forte persistenza dei raggruppamenti di volatilità.

In base ai risultati ottenuti si arriva alla specificazione del seguente modello: eGARCH con 5 parametri auto-Regressivi, GARCH esponenziale o logartimo di sigma:

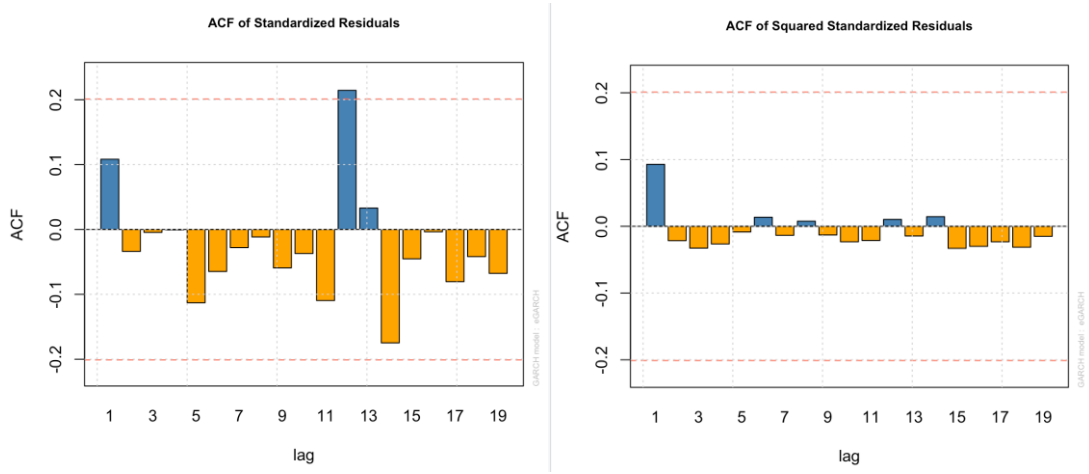
$$log\ \sigma_t^2 = w_t + \sum_{k=1}^{\bowtie} \beta_k\, g(Z_{t-k})$$

Successivamente vengono analizzati i residui del modello. I residui standardizzati si ottengono dividendo i residui ordinari per la loro deviazione standard condizionata stimata e sono quelli che vengono utilizzati per la verifica del modello, in quanto se questo è buono, né i residui stessi né i loro quadrati devono mostrare autocorrelazione.
Di seguito l'output relativo ai residui standardizzati e ai residui standardizzati al quadrato:

```
Weighted Ljung-Box Test on Standardized Residuals
------------------------------------
                         statistic p-value
Lag[1]                       1.075  0.2997
Lag[2*(p+q)+(p+q)-1][17]     4.828  1.0000
Lag[4*(p+q)+(p+q)-1][29]     9.130  0.9897
d.o.f=6
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals
------------------------------------
                         statistic p-value
Lag[1]                       1.315  0.2515
Lag[2*(p+q)+(p+q)-1][5]      1.380  0.7695
Lag[4*(p+q)+(p+q)-1][9]      1.433  0.9609
d.o.f=2
```

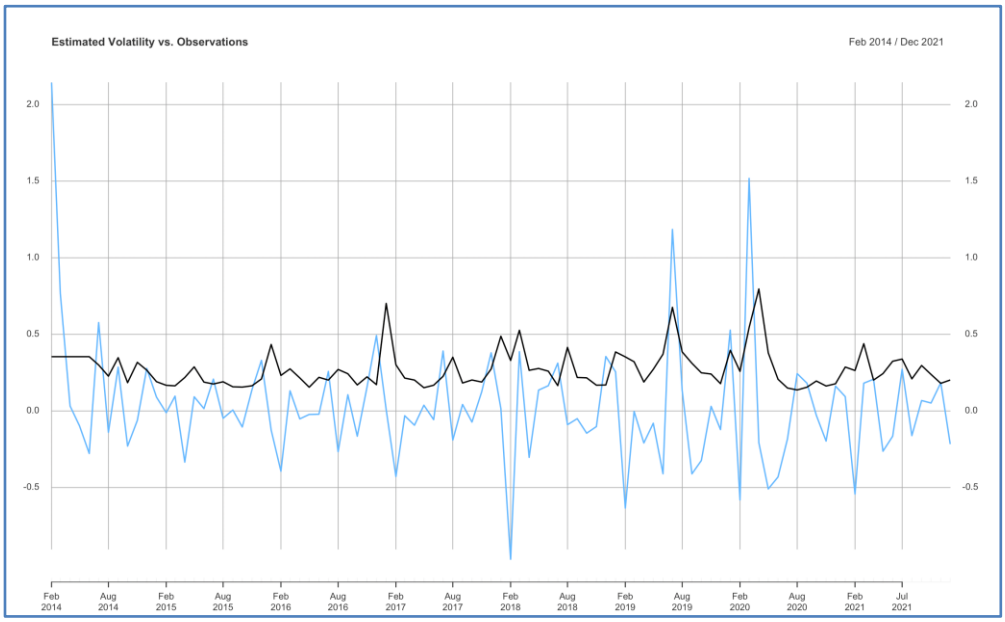
Per entrambi, più sono alti i p-values e minore è la probabilità di presenza di correlazione seriale (almeno per i ritardi testati). Nel nostro caso i p-values sono estremamente elevati, il che ci lascia pensare che non dovremmo riscontrare un'autocorrelazione significativa. L'analisi dei correlogrammi lo conferma:



Le bande superiori e inferiori dei correlogrammi fino al 19-esimo lag non vengono mai raggiunte. Si analizza dunque quanto ottenuto dal modello, il quale stima la seguente volatilità e i seguenti valori predetti per i successivi cinque mesi:

```
0-roll forecast [T0=Dec 2021]:
      Series  Sigma
T+1  0.09610 0.2725
T+2  0.00779 0.2582
T+3  0.04671 0.2534
T+4  0.02143 0.2517
T+5 -0.01783 0.2511
```

Di seguito si riporta invece la volatilità stimata versus le osservazioni:



Dai risultati ottenuti, un modello garch per i dati a disposizione sembrerebbe essere adeguato alla stima della volatilità della nostra serie storica temporale.

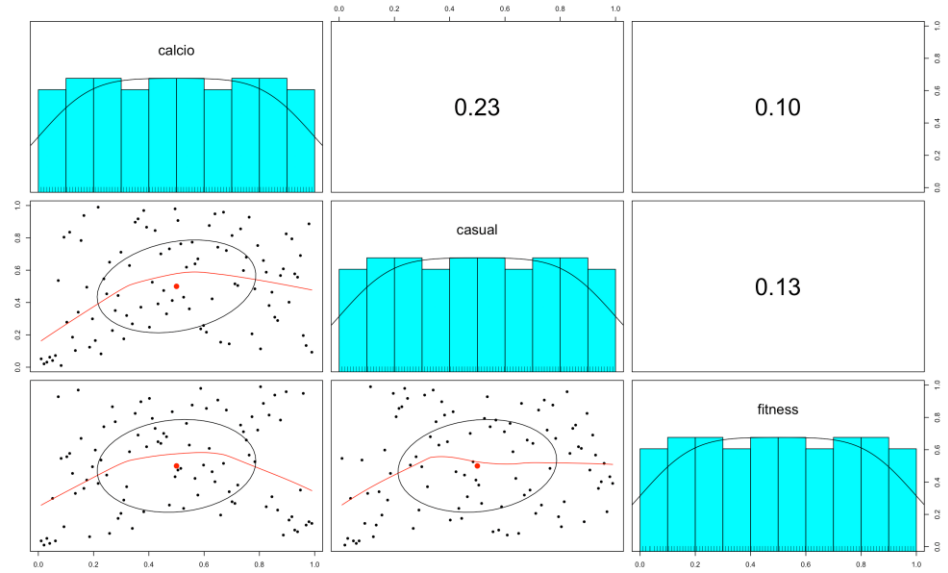
7 Studio della dipendenza tra serie storiche (di specifiche categorie) utilizzando il metodo generale di formulazione per una distribuzione multivariata della COPULA.

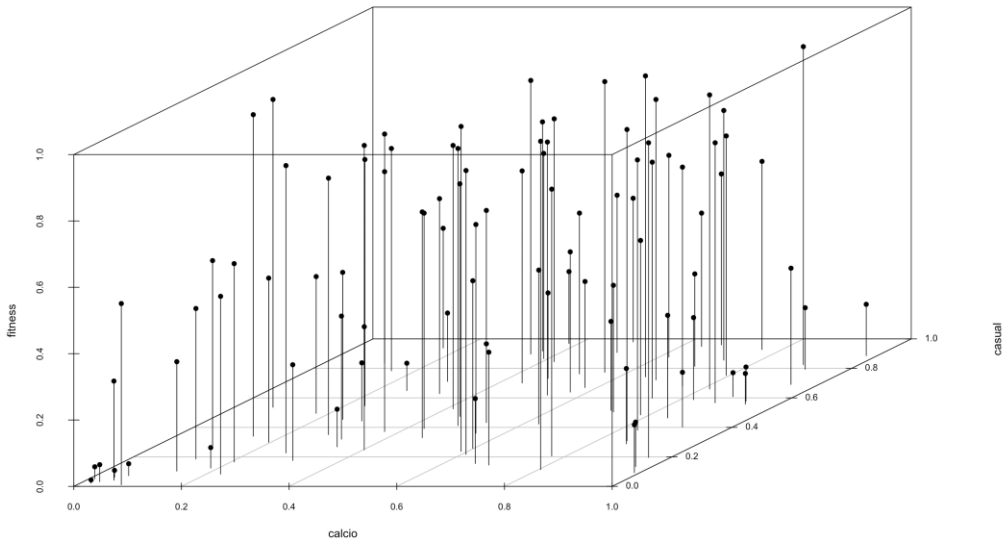
Da ultimo si è provato a descrivere il tipo di struttura di dipendenza tra le serie storiche relative al settore calcio, casual e fitness usando il metodo della copula.
Calcio, casual e fitness sono state selezionate come categorie per questa analisi perché tra le più alte e molto conformi tra loro in termini di frequenza di osservazioni:

Calcio	2.956
Casual	2.901
Fitness	2.834

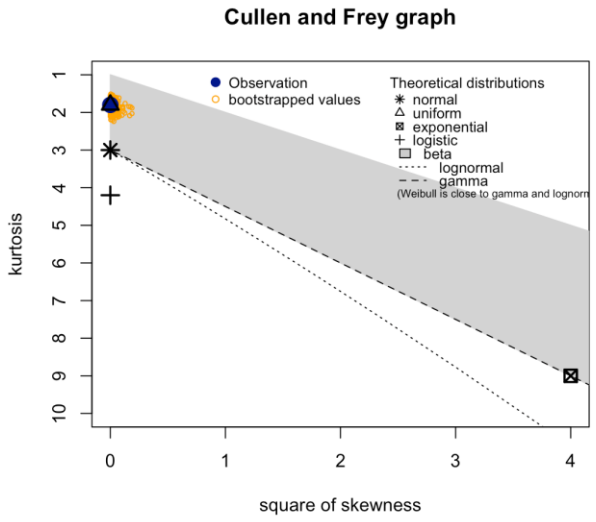
Si sono prima definite le serie temporali con inizio 01/2014 e fine 12/2021 e frequenza di 12 mesi

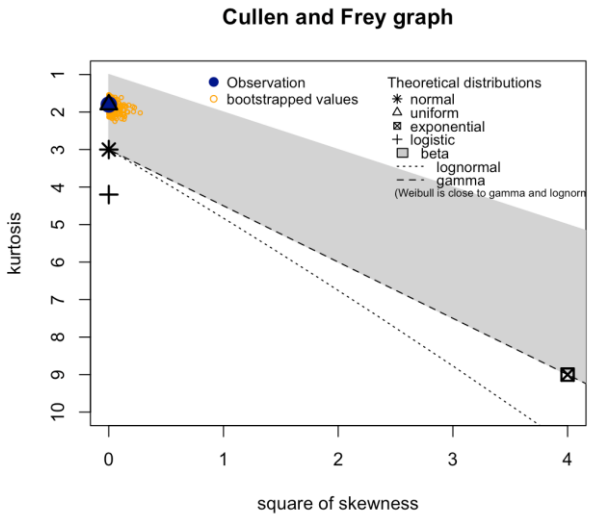
per le tre categorie, in quanto condividono un numero di osservazioni molto vicino relativamente a questo lasso temporale.
Si studiano in primo luogo la matrice di correlazione delle tre serie storiche e la loro relazione su un grafico tridimensionale delle distribuzioni marginali, in modo tale da testare la famiglia di copule più adatta in base alle relazioni marginali che sussistono tra loro:

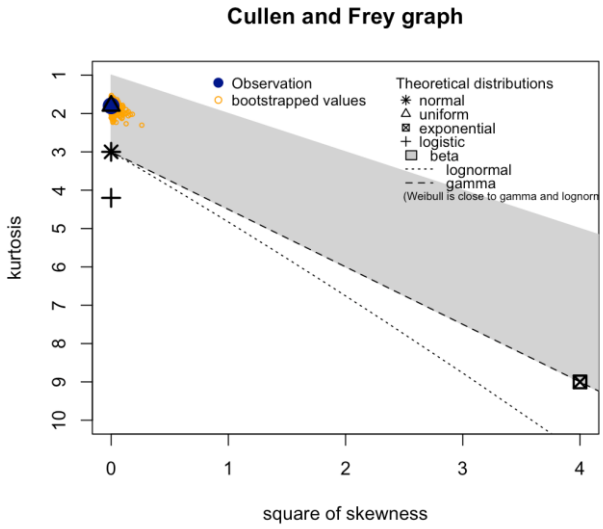




Dai grafici di cui sopra è possibile intravedere una correlazione positiva lieve tra le tre serie storiche e intuire, in via empirica, che il tipo di distribuzione marginale per tutte e tre le variabili è uniforme.
Si studia con maggior rigorosità la distribuzione marginale delle tre variabili testando il fitting di diverse distribuzioni attraverso il grafico di Cullen and Frey:



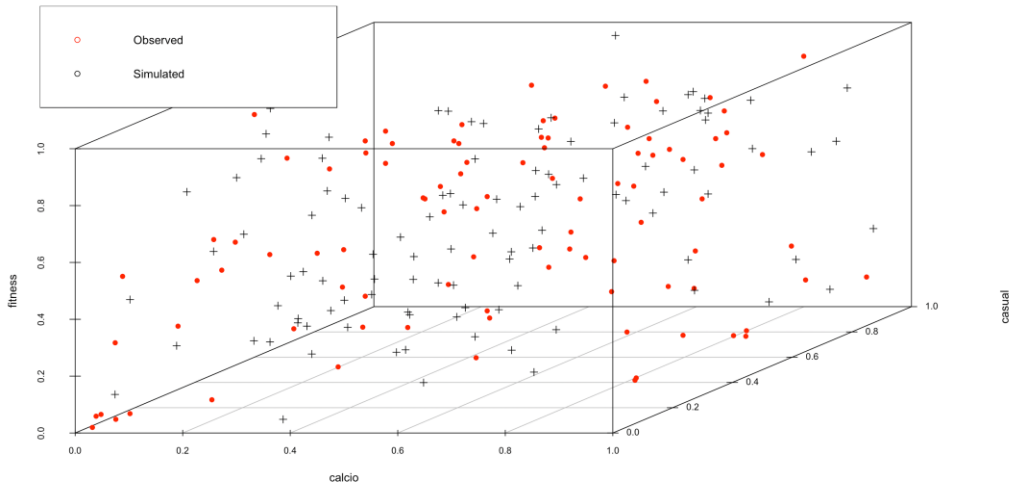




Dai risultati ottenuti si evince quanto già riscontrato in precedenza, le tre serie temporali hanno distribuzione uniforme. Una volta trovate le distribuzioni marginali per la copula si procede a testare la famiglia di copule più adatta per i dati a disposizione, ovvero la struttura di dipendenza esistente.

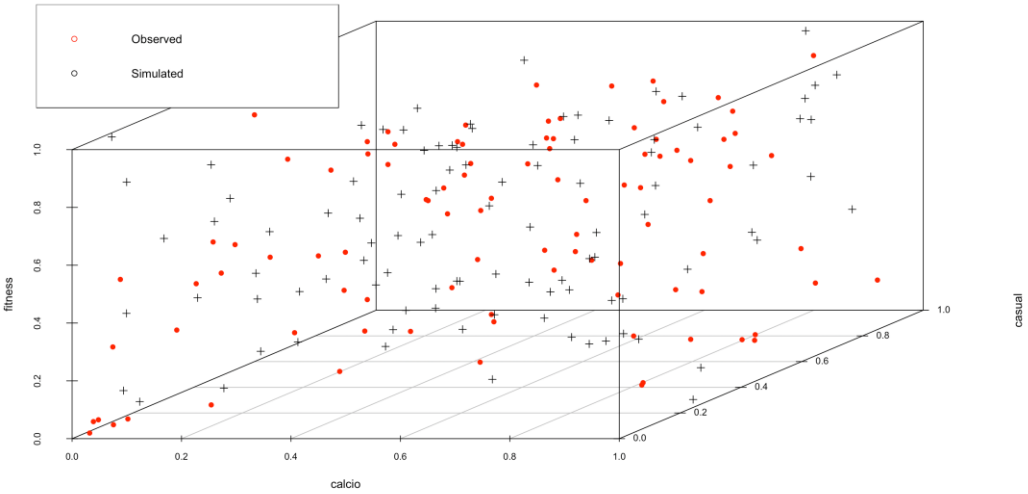
Sono state testate due diverse strutture di dipendenza: una basata sulla distribuzione Normale con distribuzioni marginali uniformi e una basata sulla T student con distribuzioni marginali uniformi.

NORMALE:




```
           [,1]       [,2]       [,3]
[1,] 1.0000000 0.13815150 0.20893770
[2,] 0.1381515 1.00000000 0.04700103
[3,] 0.2089377 0.04700103 1.00000000
```

T STUDENT:




```
           [,1]       [,2]       [,3]
[1,] 1.0000000 0.16434443 0.11956796
[2,] 0.1643444 1.00000000 0.04358836
[3,] 0.1195680 0.04358836 1.00000000
```

Dai grafici si osserva che tra una struttura di dipendenza Normale e una T Student, è preferibile la T student; infatti, la matrice di correlazione è più fedele a quella ottenuta per i dati osservati.

La t copula quindi porta a risultati vicini alle osservazioni reali, tuttavia, un fattore da non trascurare è che la struttura delle dipendenze potrebbe non essere fissata, ma piuttosto variare nel tempo. Questo comportamento ovviamente non è incluso nella copula ma dovrebbe essere preso in considerazione durante le simulazioni.

8 Conclusioni

Dall'analisi effettuata è emerso che la previsione tramite un modello ARMA perde di significatività per via delle caratteristiche della nostra serie storica. Per una tipologia di serie storica come quella in oggetto, che mostra spesso periodi con alta e bassa concentrazione di volatilità, una volatilità che cambia nel tempo è molto più frequente di una volatilità costante. Per questo motivo un modello ARMA, in questo caso, ha poca significatività rispetto alle stime ottenute con un modello GARCH. Di conseguenza abbiamo modellato la volatilità tramite un modello GARCH esponenziale, ottenendo stime dei parametri di ottimalità significativi, dei correlogrammi relativi ai residui standardizzati e dei residui standardizzati al quadrato senza la presenza di correlazione seriale. Abbiamo quindi dedotto un modello, che sulla base dei risultati ottenuti, è risultato statisticamente significativo per la stima della volatilità della serie temporale in oggetto. Da ultimo abbiamo scoperto tramite le copule una struttura di dipendenza tra le tre serie storiche prese in considerazione, individuando una struttura di dipendenza di tipo T student.

Bibliografia

- ✓ "Time Series and Forecasting", Robert I. Kabacoff, Ph.D. , 2017 (https://www.statmethods.net/advstats/timeseries.html#:~:text=Creating%20a%20time%20series,%3Dmonthly%2C%20etc.)
- ✓ "Stock Market Analysis with R programming language", Nikhil Adithyan, Jul 13, 2020 (https://medium.com/codex/stock-market-analysis-with-r-programming-language-c3ab502eb3e7)
- ✓ "Hands-On Time Series Analysis with R", Imran Arif (https://sites.google.com/site/imranlds80/teaching/forecasting-and-time-series-models-in-r?pli=1)
- ✓ "Hands-On Time Series Analysis with R: Perform time series analysis and forecasting using R", Rami Krispin, 2019
- ✓ "Modelling Dependence with Copulas in R", Michy Alice, Apr 28, 2017 (https://datascienceplus.com/modelling-dependence-with-copulas/)
- ✓ "Analisi delle serie storiche con R", Vito Ricci, 21 febbraio 2005 (https://cran.r-project.org/doc/contrib/Ricci-ts-italian.pdf)
- ✓ "Analisi delle serie storiche e previsioni di serie temporali in R", Paolo Gironi, 2022 (https://www.gironi.it/blog/analisi-delle-serie-storiche0-e-previsioni-di-serie-temporali-in-r-con-il-metodo-holt-winters/)