Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in

Statistica per l'Economia e l'Impresa

RELAZIONE FINALE

Modelli mSARIMA per serie storiche multistagionali: analisi e applicazioni

Relatore
Prof. Lisi Francesco

Laureando
Bortolato Alberto
Matricola n° 2038903

ANNO ACCADEMICO 2023-2024

Sommario

Nell'analisi delle serie storiche reali, specialmente quelle con elevata frequenza di campionamento, è possibile che i dati siano influenzati da più di una componente stagionale. I modelli mSARIMA sono un'estensione dei modelli SARIMA e consentono di cogliere la presenza di multistagionalità.

Il lavoro consiste nel definire la classe di modelli mSARIMA, verificarne il corretto funzionamento tramite uno studio di simulazione, confrontarne l'accuratezza previsionale con uno dei più conosciuti modelli in letteratura e applicarla allo studio del fenomeno delle maree a Venezia.

Indice

1 Introduzione **1**

2 Modelli mSARIMA **7**

2.1 Definizioni 7

2.1.1 Il modello ARIMA 7

2.1.2 Il modello SARIMA 8

2.1.3 Il modello mSARIMA 10

2.2 La procedura di Box-Jenkis 12

2.2.1 Fase 1: Analisi preliminari 12

2.2.2 Fase 2: Identificazione 13

2.2.3 Fase 3: Stima dei parametri 14

2.2.4 Fase 4: Diagnostica 19

3 Studio di simulazione **23**

3.1 Introduzione 23

3.2 Piano di simulazioni 24

3.2.1 Modello 1: mSARIMA$(1,0,0)(1,0,0)_{S_1}(1,0,0)_{S_2}$ 26

3.2.2 Modello 8: mSARIMA$(1,1,0)(1,1,0)_{S_1}(1,1,0)_{S_2}$ 29

3.2.3 Valutazioni 32

4 Previsioni **33**

4.1 Introduzione 33

4.2 Analisi effettuate 36

4.2.1 Generatore deterministico con trend 41

4.2.2 Generatore deterministico senza trend 42

4.2.3 Generatore stocastico con trend 43

4.2.4 Generatore stocastico senza trend 44

4.2.5 Modello 1: mSARIMA$(1,0,0)(1,0,0)_4(1,0,0)_7$ 45

4.2.6 Modello 8: mSARIMA$(1,1,0)(1,1,0)_4(1,1,0)_7$ 46

4.2.7 Valutazioni 47

5 Le maree a Venezia **49**
5.1 Contesto 49
5.2 Analisi 53
5.2.1 Il dataset 53
5.2.2 Il modello mSARIMA in-sample 57
5.2.3 Le previsioni di marea 59

6 Conclusioni **63**

Bibliografia e sitografia **65**

A Studio di simulazione **67**
A.0.1 Modello 2: mSARIMA$(1,0,0)(0,0,1)_{S_1}(0,0,1)_{S_2}$ 68
A.0.2 Modello 3: mSARIMA$(0,0,0)(1,0,1)_{S_1}(1,0,1)_{S_2}$ 71
A.0.3 Modello 4: mSARIMA$(1,0,1)(0,1,0)_{S_1}(0,1,0)_{S_2}$ 74
A.0.4 Modello 5: mSARIMA$(0,0,0)(1,1,0)_{S_1}(1,1,0)_{S_2}$ 77
A.0.5 Modello 6: mSARIMA$(0,0,0)(0,1,1)_{S_1}(0,1,1)_{S_2}$ 80
A.0.6 Modello 7: mSARIMA$(0,1,1)(0,1,1)_{S_1}(0,1,1)_{S_2}$ 83

B Previsioni **87**
B.0.1 Modello 2: mSARIMA$(1,0,0)(0,0,1)_4(0,0,1)_7$ 88
B.0.2 Modello 3: mSARIMA$(0,0,0)(1,0,1)_4(1,0,1)_7$ 89
B.0.3 Modello 4: mSARIMA$(1,0,1)(0,1,0)_4(0,1,0)_7$ 90
B.0.4 Modello 5: mSARIMA$(0,0,0)(1,1,0)_4(1,1,0)_7$ 91
B.0.5 Modello 6: mSARIMA$(0,0,0)(0,1,1)_4(0,1,1)_7$ 92
B.0.6 Modello 7: mSARIMA$(0,1,1)(0,1,1)_4(0,1,1)_7$ 93

Capitolo 1

Introduzione

Nell'analisi delle serie storiche, per spiegare i dati e per poter fare delle previsioni accurate, può essere importante tenere conto delle componenti di trend, ciclo e stagionalità che influenzano il fenomeno osservato.

Il trend è la tendenza di fondo del fenomeno, riferita ad un lungo periodo di tempo. Presenta quindi una dinamica regolare, legata all'evoluzione strutturale del sistema. Il ciclo rappresenta le fluttuazioni di periodo ampio, generalmente qualche anno. La stagionalità invece è una componente periodica i cui effetti si esauriscono nell'arco di un determinato periodo. Ad esempio è comune notare, nelle serie storiche reali, che fattori climatici, sociali o economici possono ripetersi con regolarità tutti gli anni.

Quando i dati sono raccolti con bassa frequenza, ad esempio come avviene per dati mensili o trimestrali, solitamente non è evidente la presenza di più di un ciclo stagionale (quello annuale). Quando però i dati sono raccolti con frequenza più elevata, ad esempio ogni giorno, ogni ora o ogni minuto, potremmo osservare la presenza di più componenti periodiche. Se i dati fossero raccolti ogni ora, ad esempio, potremmo notare che un fenomeno presenta caratteristiche simili alla stessa ora di ogni giorno, nello stesso giorno di ogni settimana, e nello stesso periodo di ogni anno. Questo sarebbe indice della presenza di tre cicli stagionali che esistono e influenzano i dati contemporaneamente. L'importanza di utilizzare dei modelli che riescano a considerare tutte le componenti periodiche necessarie è aumentata negli ultimi anni, perché grazie allo sviluppo tecnologico sono sempre più gli strumenti, ad esempio sensori o rilevatori, che registrano dati automaticamente e a frequenze di campionamento molto elevate.

Di seguito sono riportati alcuni esempi tratti da dataset reali in cui è evidente la presenza di multistagionalità.

Domanda energetica Le previsioni riguardanti la domanda di elettricità sono necessarie per la pianificazione dell'ammontare di energia da offrire nel mercato e per la programmazione delle manutenzioni. Errori di previsione in questo campo hanno un impatto significativo: l'energia generata dovrebbe sempre uguagliare quella consumata dagli utenti, quindi sovrastime della quantità richiesta portano ad uno spreco di energia, invece sottostime portano ad un sovraccarico della rete, che influenzerà il verificarsi di interruzioni di corrente.

Viene presentata la serie storica della domanda di energia elettrica, espressa in MW, in Brasile. Il dataset originale [1] contiene dati raccolti ogni ora dal 01/01/2000 al 31/12/2022, mentre nel grafico sono presentati quelli del mese di gennaio del 2000. È evidente la presenza di due periodi stagionali: un ciclo giornaliero, di periodo 24; un ciclo settimanale, di periodo 168. È facilmente intuibile, infatti, che la richiesta di energia elettrica sia simile in determinate ore del giorno, ma anche in determinati giorni della settimana.

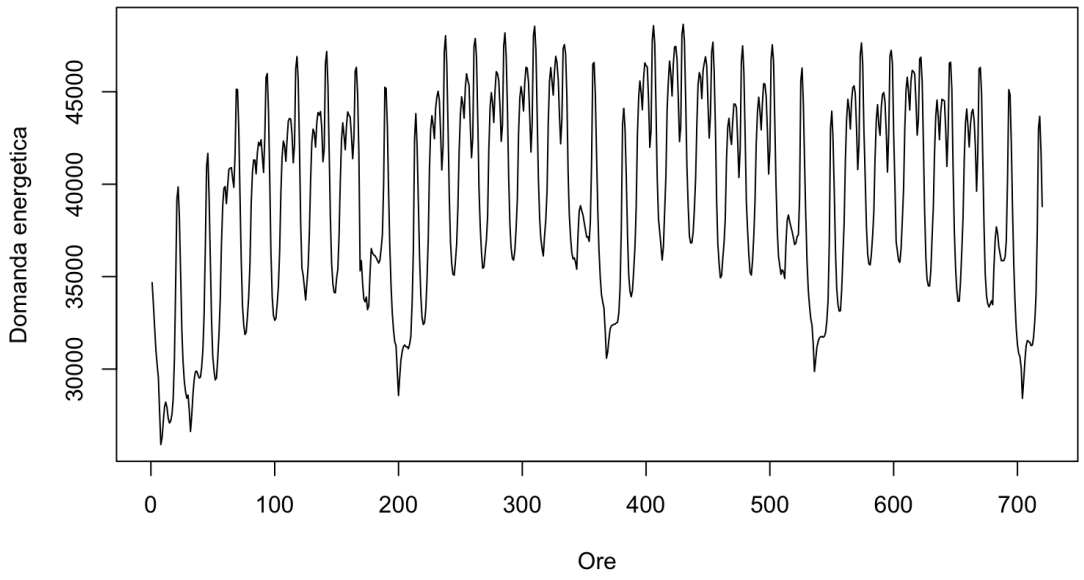


Figura 1.1: Domanda energetica oraria (in MW), Brasile, gennaio 2000.

Flussi di traffico Studiare e prevedere i flussi di traffico può essere utile per diversi fattori, come la gestione delle infrastrutture stradali, la sicurezza stradale, l'efficienza dei trasporti e molti altri.

Questo dataset [2], che comprende dati sul traffico, è stato raccolto per una ricerca intitolata "Twitter-informed Prediction for Urban Traffic Flow Using Machine

Learning". I dati sono stati raccolti tramite il California Performance Measurement System (PeMS), negli Stati Uniti. Comprende i dati sul traffico, tra cui le informazioni sulla velocità e sul flusso, per le corsie in direzione est della Ventura Highway a Los Angeles, coprendo il periodo dal 1° febbraio al 31 maggio 2020.

Nel grafico in figura (1.2) è riportata la serie storica del flusso di traffico dal 1° aprile al 31 maggio 2020. I dati hanno una frequenza di campionamento di cinque minuti. Si nota una componente stagionale giornaliera, di periodo 288, ma è possibile intravedere anche una stagionalità settimanale, di periodo 2016.

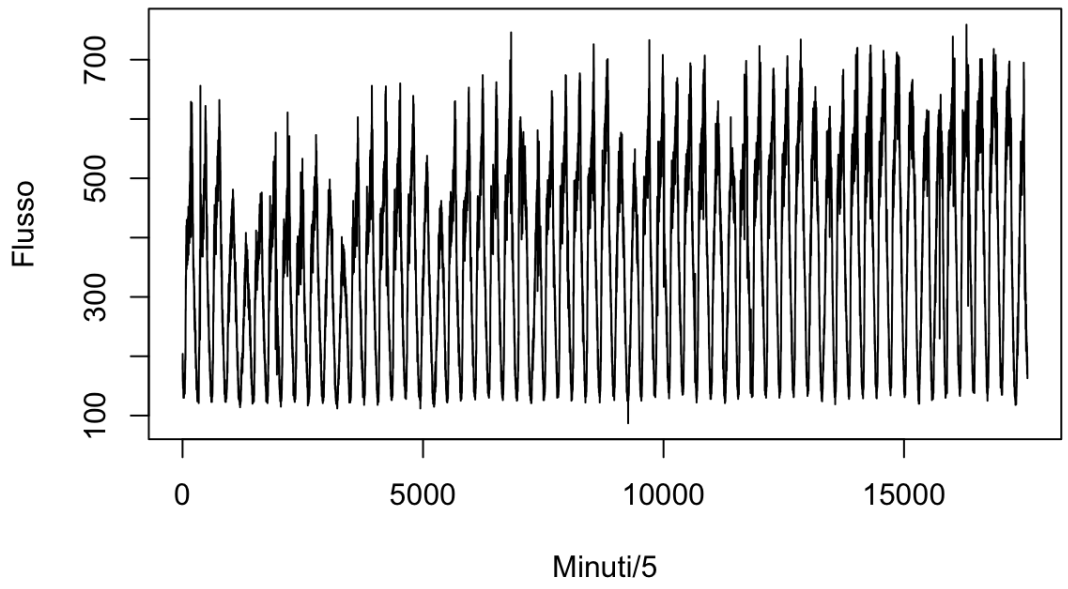


Figura 1.2: Flusso del traffico ogni 5 minuti, Ventura Highway, Los Angeles, aprile e maggio 2020

Maree a Venezia Prevedere le maree a Venezia è di fondamentale importanza per diverse ragioni. Venezia è particolarmente vulnerabile all'alta marea, che quando supera i 110cm è nota come 'acqua alta'. Previsioni accurate consentono di attivare tempestivamente le misure di protezione, come le barriere del MOSE. Mettere in funzione questo modulo è molto costoso, è quindi necessario azionarlo esclusivamente quando serve. Le previsioni, in caso di emergenze, portano ad attivare anche un sistema di sirene ed allarmi diversificati in base all'altezza prevista dell'acqua. Tutto ciò è volto a tutelare i residenti e i visitatori, riducendo il rischio di incidenti e

prevenendo la possibilità di danni più gravi a edifici di rilevanza storica, negozi e abitazioni.

Viene riportata la serie storica del livello dell'acqua (in cm), misurato a Punta della Salute a Venezia nei mesi di gennaio e febbraio 2022 [3]. Le maree sono influenzate dall'attrazione gravitazionale che coinvolge Terra, Luna e Sole. A causa di fenomeni spiegati nel dettaglio nel Capitolo 5, le principali stagionalità presenti in questa serie storica sono di circa 24 ore, di circa 29.5 giorni e di circa un anno. Dal grafico in Figura (1.3) è possibile notare in particolare l'interazione tra la ciclicità giornaliera e quella di 29.5 giorni.

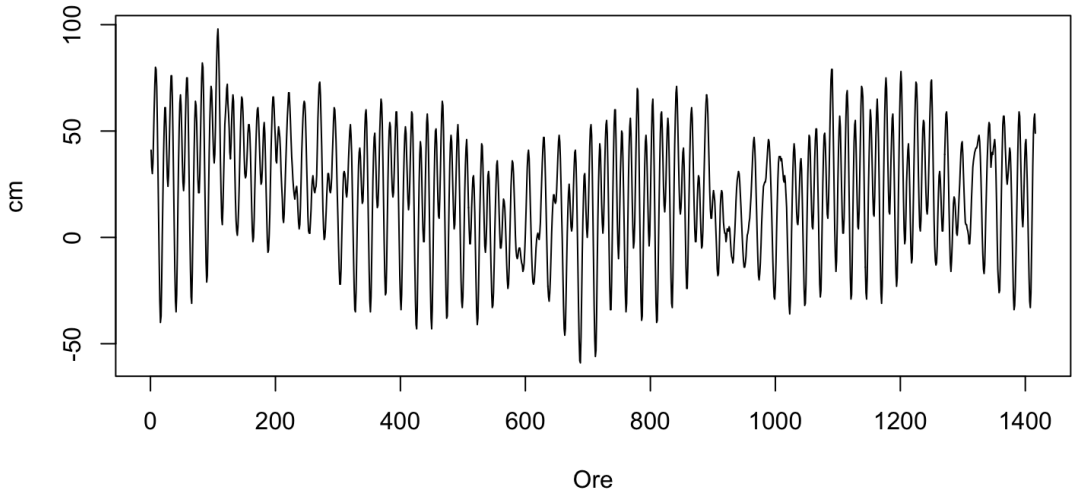


Figura 1.3: Livello dell'acqua in cm a Punta della Salute, Venezia, gennaio e febbraio 2022.

L'approccio classico dell'analisi delle serie storiche considera le componenti di trend, ciclo e stagionalità come elementi deterministici in funzione del tempo: il trend viene incluso nel modello attraverso la regressione della variabile tempo, o di sue particolari trasformazioni per includere trend di tipo non lineare; le stagionalità vengono incluse attraverso l'utilizzo di variabili dummy in funzione del tempo. Essendo un semplice modello di regressione lineare, per gestire molteplici stagionalità è sufficiente inserire nel modello le rispettive variabili dummy.

Nell'approccio moderno, invece, la serie storica osservata è ipotizzata come la realizzazione di un processo stocastico a componenti correlate, chiamato processo generatore dei dati. Ogni istante della serie storica è la realizzazione di una variabile casuale Y_t, che essendo caratterizzata da parametri diversi al variare di $t = 1, \ldots, n$,

fa assumere alla serie storica valori diversi ad ogni istante secondo determinate leggi di probabilità. Ciò che caratterizza queste variabili casuali, ed è oggetto di studio per la descrizione del modello, sono le loro correlazioni. Le componenti di trend, ciclo e stagionalità non sono più decomposte e trattate distintamente, e l'attenzione si sposta nel descrivere la natura e l'entità delle autocorrelazioni della serie storica. Ad esempio, se tutti i valori della serie storica distanti 12 osservazioni sono molto correlati tra di loro, potremmo essere in presenza di una stagionalità di periodo 12, che va gestita oppurtunamente tramite specifiche tecniche statistiche sulla base della natura delle correlazioni. Da questo modo di gestire le componenti di trend e stagionalità, risulta più complesso estendere il modello a casi in cui siano presenti più periodi stagionali contemporaneamente.

Per approcciare dinamiche in cui sia presente stagionalità i metodi generalmente più utilizzati sono l'utilizzo di modelli SARIMA o di tecniche di lisciamento esponenziale. Questi metodi però consentono tipicamente di gestire e modellare solo una componente periodica, limitando la capacità di spiegare i dati e di fare previsioni accurate.

Per cogliere la presenza di più componenti periodiche, di recente in letteratura sono stati presentati alcuni modelli, tra cui il TBATS (Trigonometric Exponential Smoothing State Space model) [4], o i modelli MSTL (Multiple Seasonal Trend decomposition using Loess), mentre i modelli SARIMA sono stati piuttosto trascurati. Un'alternativa, che però è limitante, è l'utilizzo di modelli REG-SARIMA o SARIMAX, in cui solo una componente stagionale è trattata come stocastica e le altre sono descritte in maniera deterministica utilizzando variabili dummy, funzioni trigonometriche e funzioni spline.

In questo lavoro suggeriamo l'estensione dei modelli SARIMA a quelli che abbiamo definito modelli mSARIMA (multiple Seasonal Autoregressive Integrated Moving Average), capaci di tenere conto di stagionalità multipla in maniera del tutto stocastica. Questa classe di modelli è semplicemente un'estensione della classe di modelli SARIMA, da cui, di conseguenza, eredita tutte le proprietà.

Nel secondo capitolo vengono definiti rigorosamente i modelli mSARIMA e la relativa procedura di Box-Jenkis. Nel terzo capitolo viene effettuato uno studio di simulazione per dimostrare il funzionamento di questi modelli, verificandone il comportamento asintotico. Nel quarto capitolo il modello mSARIMA viene confrontato con uno dei principali modelli presenti in letteratura, per capire qual è il più adatto ad effettuare previsioni, anche in base al processo di generazione (deterministico o stocastico) della serie storica. Infine nel quinto capitolo il modello mSARIMA viene

applicato allo studio del fenomeno delle maree a Venezia e ne vengono valutate le prestazioni previsive.

Capitolo 2

Modelli mSARIMA

2.1 Definizioni

In generale, il valore di una serie storica stazionaria in un certo istante può essere espresso tramite una combinazione lineare di valori passati e di errori passati, relativamente a quell'istante. La modellazione di tipo ARIMA [5], che comprende anche le sue estensioni SARIMA e mSARIMA, permette di includere in una funzione matematica (l'equazione del modello): componenti autoregressive, che descrivono e quantificano l'intensità delle relazioni tra valori della serie e valori passati; componenti a media mobile, che descrivono e quantificano l'intensità delle relazioni tra errori del modello ed errori passati; componenti integrate, che descrivono le operazioni necessarie a rendere stazionaria la serie storica.

2.1.1 Il modello ARIMA

Definizione Sia $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, con $\sigma_\varepsilon^2 < \infty$, allora il processo Y_t si dice $ARIMA(p, d, q)$ se può essere espresso come

$$\phi(B)(1 - B)^d Y_t = \theta(B)\varepsilon_t,$$

dove:

- B è l'operatore ritardo (dall'inglese Backward shift operator), definito come

$$B^j Y_t = Y_{t-j} \quad j = 0,1,2,\ldots$$

- $\phi(B)$ è il polinomio caratteristico della componente autoregressiva $AR(p)$, definito come
$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p,$$
con ϕ_i $(i = 1, \ldots, p)$ parametri costanti.

- $(1 - B)^d$ è il polinomio riferito alla componente integrata, dove d indica il numero di differenziazioni

- $\theta(B)$ è il polinomio caratteristico della componente a media mobile $MA(q)$, definito come
$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q,$$
con θ_i $(i = 1, \ldots, q)$ parametri costanti.

Esempio Se poniamo p=1, d=0, q=0, allora siamo in presenza di un $ARIMA$(1,0,0). Se consideriamo $\phi = 0.5$, il modello può essere riscritto come:

$$\begin{aligned} (1 - 0.5B)Y_t &= \varepsilon_t \\ Y_t &= 0.5Y_{t-1} + \varepsilon_t. \end{aligned} \tag{2.1}$$

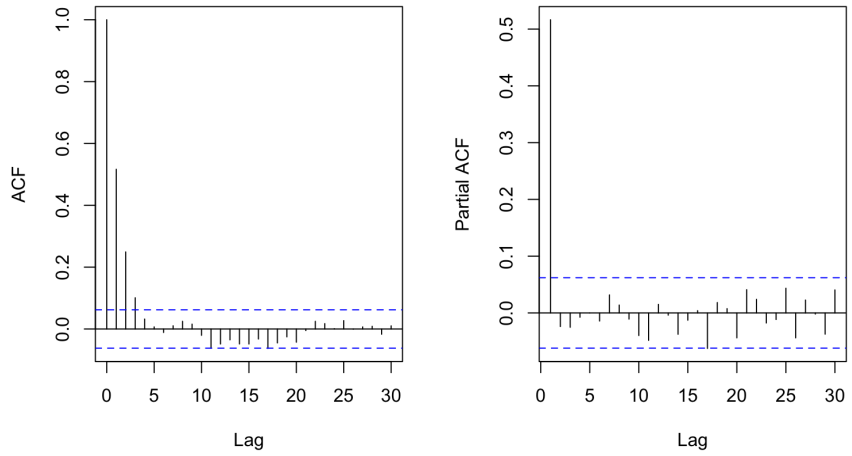


Figura 2.1: ACF e PACF di un processo (2.1) simulato

2.1.2 Il modello SARIMA

Nella pratica, molte delle serie storiche reali contengono una componente periodica stagionale, che si ripete ogni S osservazioni. Nasce quindi la necessità di estendere

i modelli $ARIMA$ a dei modelli più generici, i modelli $SARIMA$, che gestiscono la stagionalità in maniera moltiplicativa.

Definizione Sia $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, con $\sigma_\varepsilon^2 < \infty$, il processo Y_t si dice $SARIMA(p,d,q)(P,D,Q)_S$ se può essere espresso come

$$\phi(B)\Phi(B^S)(1-B)^d(1-B^S)^D Y_t = \theta(B)\Theta(B^S)\varepsilon_t,$$

dove:

- $\phi(B)$, $(1-B)^d$, $\theta(B)$ sono gli usuali polinomi caratteristici delle componenti non stagionali
- S è il periodo della stagionalità
- $\Phi(B^S)$ è il polinomio caratteristico della componente autoregressiva stagionale $AR(P)$, definito come

  $$\Phi(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS},$$

  con Φ_i $(i = 1, \ldots, P)$ parametri costanti.
- $(1-B^S)^D$ è il polinomio riferito alla componente integrata stagionale, dove D indica il numero di differenziazioni stagionali
- $\Theta(B^S)$ è il polinomio caratteristico della componente a media mobile stagionale $MA(Q)$, definito come

  $$\Theta(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \cdots - \Theta_Q B^{QS},$$

  con Θ_i $(i = 1, \ldots, Q)$ parametri costanti.

Esempio Se poniamo p=1, d=0, q=0, P=1, D=0, Q=0 e S=4, allora siamo in presenza di un $SARIMA(1,0,0)(1,0,0)_4$. Se consideriamo ϕ=0.5, Φ=0.4, il modello può essere riscritto come:

$$\begin{aligned}(1-0.5B)(1-0.4B^4)Y_t &= \varepsilon_t \\ Y_t &= 0.5Y_{t-1} + 0.4Y_{t-4} - 0.2Y_{t-5} + \varepsilon_t.\end{aligned} \tag{2.2}$$

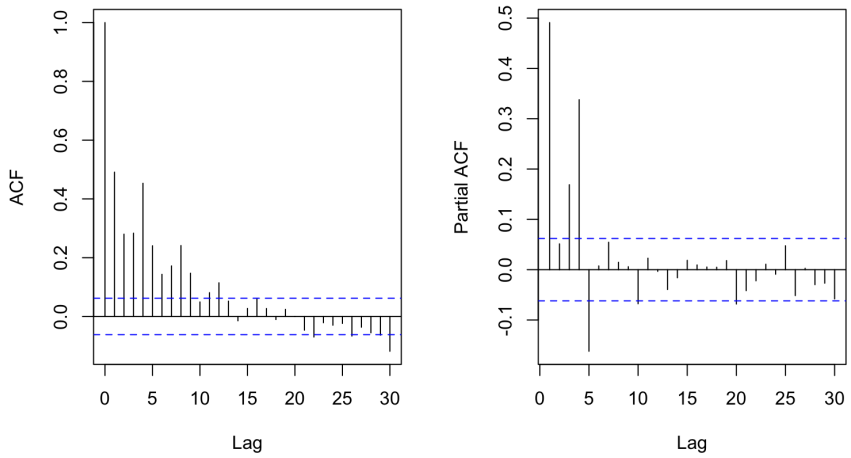


Figura 2.2: ACF e PACF di un processo (2.2) simulato

2.1.3 Il modello mSARIMA

Nella pratica, molte serie storiche reali contengono più di una sola componente periodica stagionale. In generale possono esserci m componenti periodiche che si ripetono ogni S_i osservazioni, con $i = 1, ..., m$.

Definizione Sia $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, con $\sigma_\varepsilon^2 < \infty$, il processo Y_t si dice $mSARIMA(p,d,q) \times (P_1, D_1, Q_1)_{S_1} \times \cdots \times (P_m, D_m, Q_m)_{S_m}$ se può essere espresso come

$$\left(\phi(B)(1-B)^d \prod_{i=1}^{m} \Phi(B^{S_i})(1-B^{S_i})^{D_i}\right) Y_t = \left(\theta(B) \prod_{i=1}^{m} \Theta(B^{S_i})\right) \varepsilon_t,$$

dove:

- $\phi(B)$, $(1-B)^d$, $\theta(B)$ sono gli usuali polinomi caratteristici delle componenti non stagionali
- m è il numero di stagionalità e S_i è il periodo dell'i-esima stagionalità, $i = 1, \ldots, m$
- $\Phi(B^{S_i})$ è il polinomio caratteristico dell'i-esima componente autoregressiva stagionale $AR(P_i)$, $i = 1, \ldots, m$, definito come

$$\Phi(B^{S_i}) = 1 - \Phi_{i,1}B^{S_i} - \Phi_{i,2}B^{2S_i} - \cdots - \Phi_{i,P_i}B^{P_i S_i},$$

con $\Phi_{i,j}$ $(i = 1, \ldots, m;\ j = 1, \ldots, P_i)$ parametri costanti.

- $(1 - B_i^S)^{D_i}$ è il polinomio riferito all'i-esima componente integrata stagionale, dove D_i indica il numero di differenziazioni stagionali, $i = 1, \ldots, m$
- $\Theta(B^{S_i})$ è il polinomio caratteristico dell'i-esima componente a media mobile stagionale $MA(Q_i)$, $i = 1, \ldots, m$, definito come

$$\Theta(B^{S_i}) = 1 - \Theta_{i,1}B^{S_i} - \Theta_{i,2}B^{2S_i} - \cdots - \Theta_{i,Q_i}B^{Q_i S_i},$$

con $\Theta_{i,j}$ $(i = 1, \ldots, m;\ j = 1, \ldots, Q_i)$ parametri costanti.

Esempio Se poniamo m=2, p=1, d=0, q=0, P_1=1, D_1=0, Q_1=0, P_2=1, D_2=0, Q_2=0, S_1=4 e S_1=7 , allora siamo in presenza di un $mSARIMA(1,0,0)(1,0,0)_4(1,0,0)_7$. Se consideriamo ϕ=0.5, Φ_1=0.4, Φ_2=0.3, il modello può essere riscritto come segue:

$$(1 - 0.5B)(1 - 0.4B^4)(1 - 0.3B^7)Y_t = \varepsilon_t \tag{2.3}$$
$$Y_t = 0.5Y_{t-1} + 0.4Y_{t-4} - 0.2Y_{t-5} + 0.3Y_{t-7} - 0.15Y_{t-8} + 0.12Y_{t-11} + 0.06Y_{t-12} + \varepsilon_t.$$

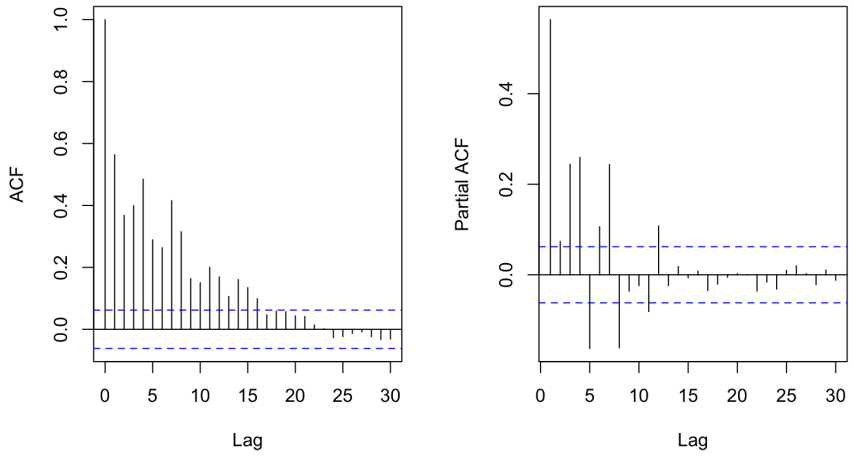


Figura 2.3: ACF e PACF di un processo (2.3) simulato

2.2 La procedura di Box-Jenkis

Perché le considerazioni teoriche affrontate nella sezione precedente siano utili nell'analisi di serie storiche è necessaria una procedura che, solo sulla base della serie storica osservata, consenta di costruire un modello mSARIMA che approssimi in maniera adeguata il processo generatore dei dati. La procedura di Box-Jenkis [5] è utile a questo scopo e si compone di diverse fasi.

2.2.1 Fase 1: Analisi preliminari

I dati devono essere opportunamente preparati, in particolare, dopo aver gestito correttamente l'eventuale presenza di outliers, bisogna rendere la serie stazionaria. Segnali di non stazionarietà possono essere individuati attraverso l'osservazione del grafico della serie storica e del comportamento della ACF. In particolare, se la serie non è stazionaria, è molto probabile che le autocorrelazioni tendano ad annullarsi molto lentamente. A tal fine può essere necessario applicare alla serie storica un certo numero di differenziazioni per ottenere stazionarietà in media, e/o opportune trasformazioni di Box-Cox [6] per ottenere stazionarietà in varianza.

É necessario porre particolare attenzione ai valori della serie storica ai ritardi stagionali, per assicurarsi che sia rispettata la stazionarietà anche in tutte le eventuali componenti periodiche. Al termine di questa fase, quindi, deve essere noto il valore degli ordini $d, D_1, \ldots, D_m$ per le differenziazioni, ed eventualmente il valore di λ per la trasformata di Box-Cox. Sulla base dei risultati ottenuti, si applicano le opportune trasformazioni per rendere la serie stazionaria per le fasi successive. In particolare indichiamo con $Y^{(\lambda)}$ la serie storica trasformata con le trasformazioni di Box-Cox e con Y_t^d la serie storica differenziata, e li definiamo come:

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \text{se } \lambda \neq 0 \\ \log(Y), & \text{se } \lambda = 0 \end{cases},$$

$$Y_t^d = Y_t - \left((1-B)^d \prod_{i=1}^{m} (1-B^{S_i})^{D_i} \right) Y_t.$$

D'ora in poi si considererà esclusivamente la serie storica stazionaria e trasformata, che per semplicità chiameremo Y_t.

2.2.2 Fase 2: Identificazione

In questa fase ci si occupa di specificare gli ordini corretti $p, q, P_1, \ldots, P_m, Q_1, \ldots, Q_m$ di tutte le componenti autoregressive e a media mobile del modello. Si utilizza l'identificazione strutturale, ovvero si cerca di riconoscere nelle funzioni di autocorrelazione empiriche (ACF e PACF) la struttura teorica di un modello mSARMA noto.

Si verificano dunque le significatività delle autocorrelazioni globali $p_k = \text{Corr}(Y_t, Y_{t-k})$, che misurano la correlazione tra Y_t e Y_{t-k}, e delle autocorrelazioni parziali $P_k = \text{Corr}(Y_t, Y_{t-k}|Y_{t-1}, \ldots, Y_{t-k+1})$, che misurano la correlazione tra Y_t e Y_{t-k} al netto delle variabili intermedie, attraverso i sistemi d'ipotesi:

- per la nullità delle autocorrelazioni:

$$\begin{cases} H_0 : p_k = 0 \\ H_1 : p_k \neq 0 \end{cases} ;$$

- per la nullità delle autocorrelazioni parziali:

$$\begin{cases} H_0 : P_k = 0 \\ H_1 : P_k \neq 0 \end{cases} .$$

Si assume che, sotto H_0, p_k e P_k abbiano una distribuzione approssimativamente normale con media 0 e varianza $\frac{1}{n}$, cioè $p_k, P_k \dot{\sim} \mathcal{N}(0, \frac{1}{n})$. Dunque le ipotesi nulle di incorrelazione non sono rifiutate, ad un livello di significatività del 95%, quando $\hat{p}_k \in \left[\frac{-1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}}\right]$ e $\hat{P}_k \in \left[\frac{-1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}}\right]$.

Dai risultati ottenuti, per identificare gli ordini delle componenti non stagionali:

- potremmo essere in presenza di un processo $AR(p)$ se p_k tende a zero all'aumentare di k (lentamente o velocemente sulla base dei valori grandi o piccoli dei parametri ϕ_i), mentre P_k è significativamente diversa da zero per le prime $k \leq p$ autocorrelazioni e si annulla per $k > p$;

- potremmo essere in presenza di un processo $MA(q)$ se p_k è significativamente diversa da zero per le prime $k \leq q$ autocorrelazioni e si annulla per $k > q$, mentre P_k tende a zero all'aumentare di k (lentamente o velocemente sulla base dei valori grandi o piccoli dei parametri θ_i);

- potremmo essere in presenza di un processo $ARMA(p, q)$ se c'è un comportamento misto.

Per identificare gli ordini delle componenti stagionali si utilizza lo stesso approccio, concentrandosi solo sui ritardi stagionali $k = S_i, 2S_i, 3S_i, \dots$ per $i = 1, \dots, m$.

2.2.3 Fase 3: Stima dei parametri

Dopo aver identificato il modello, bisogna stimarne i parametri. Indichiamo con δ il vettore dei parametri composto da:

$$\delta = (\phi_1, \dots, \phi_p, \Phi_{1,1}, \dots, \Phi_{1,P_1}, \cdots, \Phi_{m,1}, \dots, \Phi_{m,P_m}, \\ \theta_1, \dots, \theta_q, \Theta_{1,1}, \dots, \Theta_{1,Q_1}, \cdots, \Theta_{m,1}, \dots, \Theta_{m,Q_m}, \sigma_\varepsilon^2)'.$$

Esistono diversi metodi per stimare i parametri, come il metodo dei momenti (Yule-Walker) o il metodo dei minimi quadrati non lineari, ma ci soffermiamo sul metodo della massima verosimiglianza, che fornisce stimatori con migliori proprietà statistiche.

Questo metodo richiede la conoscenza della distribuzione del termine di errore, ma assumendo la normalità dei residui $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, un qualsiasi processo mSARIMA ha distribuzione normale, in quanto combinazione lineare di costanti e di eventuali combinazioni lineari del termine di errore.

Esempio Un processo $ARMA(1,1)$ con parametri $\phi = 0.5$, $\theta = 0.3$ è riscrivibile come

$$Y_t = 0.5Y_{t-1} + \varepsilon_t - 0.3\varepsilon_{t-1},$$

e per le proprietà della variabile casuale normale si ha che $Y_t|y_{t-1} \sim \mathcal{N}(0.5y_{t-1}, \sigma_\varepsilon^2)$.

Dunque l'obiettivo è quello di massimizzare la funzione di verosimiglianza di una variabile casuale normale.

$$L(\delta|y_1, y_2, \dots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \hat{y}_i)^2}{2\sigma^2}\right),$$

dove $\hat{y}_i$ è ricavabile attraverso la risoluzione dell'equazione che definisce il processo:

$$\left(\phi(B)\prod_{i=1}^{m}\Phi(B^{S_i})\right) Y_t = \left(\theta(B)\prod_{i=1}^{m}\Theta(B^{S_i})\right)\varepsilon_t. \tag{2.4}$$

Come è possibile notare, per procedere con i calcoli è necessario conoscere la serie degli ε_t, che però non è nota: l'unica informazione disponibile è la serie storica osservata $\{y_t\}_{t=1}^n$.

Di seguito sono illustrati due approcci diversi per la risoluzione di questo problema.

Rappresentazione degli ε_t tramite sviluppi di Taylor

Essendo nota solo la serie storica degli $\{y_t\}_{t=1}^n$, e non quella dei termini di errore $\{\varepsilon_t\}_{t=1}^n$, l'obiettivo, in questo approccio, è quello di ottenere un unico polinomio in termini di B che tenga conto sia delle componenti autoregressive sia delle componenti a media mobile del modello. Questo polinomio, moltiplicato per y_t, descriverà y_t in termini del suo passato e sarà quindi la sua stima, $\hat{y}_t$.

Per semplificare la notazione analizziamo una componente alla volta:

- Se consideriamo solo la componente autoregressiva il processo può essere riscritto come:

$$\left(\phi(B)\prod_{i=1}^{m}\Phi(B^{S_i})\right)Y_t = \varepsilon_t$$

$$Y_t = Y_t - \left(\phi(B)\prod_{i=1}^{m}\Phi(B^{S_i})\right)Y_t + \varepsilon_t. \tag{2.5}$$

  In questo modo abbiamo riscritto il processo Y_t in termini del suo passato. Infatti dopo lo sviluppo di tutti i prodotti, i due termini Y_t a destra dell'equazione si annulleranno a vicenda e rimarranno solo istanze passate del processo più un termine di errore.

- Se consideriamo solo la componente a media mobile il processo può essere riscritto come:

$$Y_t = \left(\theta(B)\prod_{i=1}^{m}\Theta(B^{S_i})\right)\varepsilon_t.$$

  E quindi è definito come una determinata combinazione lineare del passato degli errori del modello. Come anticipato, la serie degli errori ε_t non è nota in quanto il modello non è ancora stato stimato. È però possibile spostare il polinomio che moltiplica il termine di errore a sinistra dell'equazione, così che questo interagisca con il processo Y_t noto, ottenendo

$$\left(\theta(B)^{-1}\prod_{i=1}^{m}\Theta(B^{S_i})^{-1}\right)Y_t = \varepsilon_t.$$

Ora però è impossibile svolgere il prodotto tra l'inverso del polinomio delle componenti MA e Y_t, in quanto la funzione di B si trova al denominatore. Per risolvere tale problema è utile l'approssimazione di questa funzione attraverso lo sviluppo in serie di Taylor.

Una generica serie di Taylor è definita come:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots .$$

Se queste considerazioni vengono applicate alle funzioni $\theta(B)^{-1}$ e $\Theta(B^{S_i})^{-1}$, centrando in $a = 0$ otteniamo:

$$\theta(B)^{-1} = \theta(a)^{-1} + \frac{d}{da}\theta(a)^{-1}B + \frac{1}{2!}\frac{d^2}{da^2}\theta(a)^{-1}B^2 + \frac{1}{3!}\frac{d^3}{da^3}\theta(a)^{-1}B^3 + \cdots ;$$

$$\Theta(B^{S_i})^{-1} = \Theta(a)^{-1} + \frac{d}{da}\Theta(a)^{-1}B^{S_i} + \frac{1}{2!}\frac{d^2}{da^2}\Theta(a)^{-1}B^{2S_i} + \frac{1}{3!}\frac{d^3}{da^3}\Theta(a)^{-1}B^{3S_i} + \cdots .$$

E in maniera del tutto analoga con quanto visto per le componenti autoregressive, il processo è riscrivibile in termini del suo passato:

$$Y_t = Y_t - \left(\theta(B)^{-1}\prod_{i=1}^{m}\Theta(B^{S_i})^{-1}\right)Y_t + \varepsilon_t. \tag{2.6}$$

Unendo i risultati (2.5) e (2.6), è possibile moltiplicare i due polinomi per ottenerne uno unico. In particolare:

$$Y_t = Y_t - \left(\phi(B)\prod_{i=1}^{m}\Phi(B^{S_i})\right)\left(\theta(B)^{-1}\prod_{i=1}^{m}\Theta(B^{S_i})^{-1}\right)Y_t + \varepsilon_t.$$

È utile introdurre un parametro M, scelto per approssimare il polinomio di infiniti termini della componente a media mobile non stagionale $\theta(B)^{-1}$ ad un polinomio di grado M, e per approssimare i polinomi delle componenti a media mobile stagionali $\Theta(B^{S_i})^{-1}$, che sono di infiniti termini, a polinomi di grado MS_i, per $i = 1, \ldots, m$.

Esempio Se siamo in presenza di un $ARIMA(1,0,1)$ con parametri $\phi = 0.5, \theta = 0.4, M = 4$ il processo è riscrivibile come:

$$\begin{aligned} Y_t &= Y_t - (1 - 0.5B)\left(1 + 0.4B + 0.16B^2 + 0.064B^3 + 0.0256B^4\right) Y_t + \varepsilon_t \\ &= Y_t - \left(1 - 0.1B - 0.04B^2 - 0.016B^3 - 0.0064B^4 - 0.00128B^5\right) Y_t + \varepsilon_t \\ &= 0.1Y_{t-1} + 0.04Y_{t-2} + 0.016Y_{t-3} + 0.0064Y_{t-4} + 0.00128Y_{t-5} + \varepsilon_t. \end{aligned}$$

Lo scopo della creazione del polinomio era quello di ottenere una stima di y_t che potesse essere utilizzata per massimizzare la verosimiglianza. Essendo questa stima costruita con il passato della serie storica, il primo valore stimabile è il $(k+1)$-esimo, dove k rappresenta il grado del polinomio caratteristico complessivo, ed è dato da

$$k = p + d + qM + \sum_{i=1}^{m} (P_i + D_i + Q_i M) S_i.$$

Quindi ci si può limitare a massimizzare la funzione di verosimiglianza

$$L(\delta) = \prod_{t=k+1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right).$$

Per motivi computazionali è più efficiente lavorare sulla log-verosimiglianza $l(\delta) = \log(L(\delta))$, che essendo una trasformazione monotona non altera i risultati.

$$\begin{aligned} l(\delta) &= \log\left(\prod_{t=k+1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right)\right) \\ &= \sum_{t=k+1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{t=k+1}^{n} \log\left(\exp\left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right)\right) \\ &= (n-k)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{t=k+1}^{n} \left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right) \\ &= -\frac{n-k}{2}\log(2\pi) - \frac{n-k}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=k+1}^{n} (y_t - \hat{y}_t)^2 \\ &\propto -\frac{n-k}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=k+1}^{n} (y_t - \hat{y}_t)^2. \end{aligned} \tag{2.7}$$

Stima degli ε_t

A livello computazionale, utilizzare l'approssimazione di Taylor è svantaggioso perché implica il calcolo di molte derivate, anche di ordine elevato. Inoltre, se si vuole

ottenere una buona approssimazione, utilizzando quindi valori di M elevati, questo implica che un gran numero di residui non potranno essere calcolati, limitando la capacità del modello di poter massimizzare la log-verosimiglianza in modo adeguato specialmente per serie storiche di bassa numerosità, dove è difficile sfruttare tutta l'informazione disponibile.

Un modo più efficiente per calcolare la serie dei valori stimati $\hat{y}_t$ è la costruzione della serie degli errori ε_t tramite delle inizializzazioni, così da poterla utilizzare direttamente nell'equazione del modello (2.4). Per stimare un qualsiasi valore y_t del processo, serviranno dunque valori e/o errori passati fino all'istante $t-k$, dove k è il grado massimo tra quelli dei due distinti polinomi complessivi delle componenti AR e MA:

$$k = max(p + \sum_{i=1}^{m} P_i S_i, q + \sum_{i=1}^{m} Q_i S_i).$$

1. Inizializziamo una serie degli epsilon allungata, ε_t^*, e una serie storica allungata, y_t^*. Utilizzare queste serie ci permetterà di stimare anche i primi k valori della serie storica. È ragionevole assegnare ai primi k valori della serie degli y_t^* la media del processo, e ai primi k valori della serie degli ε_t^* la media degli errori, ovvero zero:

$$y_t^* = \begin{cases} \frac{1}{n}\sum_{t=1}^{n} y_t & t = 1, \ldots, k \\ y_{t-k} & t = k+1, \ldots, n+k \end{cases};$$

$$\varepsilon_t^* = 0 \quad t = 1, \ldots, n+k.$$

2. Calcoliamo la serie degli errori ε_t^* come differenza tra y_t^* e la sua stima, utilizzando anche gli errori appena calcolati:

$$\varepsilon_t^* = y_t^* - \left[\left(\phi(B) \prod_{i=1}^{m} \Phi(B^{S_i}) \right) y_t^* + \left(\theta(B) \prod_{i=1}^{m} \Theta(B^{S_i}) \right) \varepsilon_t^* \right],$$

   per $t = k+1, \ldots, n+k$.

3. Calcoliamo $\hat{y_t}^*$ anche attraverso l'utilizzo della serie degli ε_t^* appena stimati, come segue:

$$\hat{y_t}^* = \left(\phi(B) \prod_{i=1}^{m} \Phi(B^{S_i}) \right) y_t^* + \left(\theta(B) \prod_{i=1}^{m} \Theta(B^{S_i}) \right) \varepsilon_t^*,$$

   per $t = k+1, \ldots, n+k$.

4. Ricaviamo $\hat{y}_t$ e ε_t, eliminando gli elementi delle serie allungate, tramite:

$$\varepsilon_t = \varepsilon^*_{t+k} \quad t = 1, \ldots, n;$$

$$\hat{y}_t = \hat{y}^*_{t+k} \quad t = 1, \ldots, n.$$

5. Massimizziamo la verosimiglianza

$$L(\delta) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right).$$

Per motivi computazionali è più efficiente lavorare sulla log-verosimiglianza $l(\delta) = \log(L(\delta))$, che essendo una trasformazione monotona non altera i risultati.

$$\begin{aligned}
l(\delta) &= \log\left(\prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right)\right) \\
&= \sum_{t=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{t=1}^{n} \log\left(\exp\left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right)\right) \\
&= n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{t=1}^{n} \left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \\
&\propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \qquad (2.8)
\end{aligned}$$

2.2.4 Fase 4: Diagnostica

Dopo aver stimato il modello, vengono fatti dei controlli per verificare che le assunzioni fatte a priori siano rispettate, così da rendere validi i risultati. In particolare si era ipotizzato che i disturbi ε_t che componevano il modello fossero indipendenti e identicamente distribuiti come $WN(0, \sigma^2_\varepsilon)$. Sulla base di questa ipotesi abbiamo dato per nota la distribuzione del processo e siamo riusciti a stimarne i parametri attraverso la stima di massima verosimiglianza. Ora quindi dobbiamo verificare che i residui $e_t = y_t - \hat{y}_t$ siano stazionari, incorrelati e distribuiti normalmente.

- Attraverso l'osservazione del grafico dei residui non dovrebbe emergere alcuna regolarità, né valori particolarmente diversi gli uni dagli altri.

- Se trattiamo la serie dei residui e_t come una serie storica, possiamo calcolarne la funzione di autocorrelazione. Dovremmo poter affermare che tutte le autocorrelazioni non siano significative, ovvero che appartengano all'intervallo $\left[\frac{-1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}}\right]$.

- Per una verifica di assenza di correlazione dei residui ai ritardi non stagionali si può utilizzare la statistica di Ljung e Box [7]:

$$Q(h) = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \dot{\sim} \chi^2(h-p-q), \tag{2.9}$$

  dove h è il numero di autocorrelazioni che si vogliono includere nel test, p e q sono gli usuali ordini della componente autoregressiva non stagionale e della componente a media mobile non stagionale.

- È necessario invece modificare leggermente la statistica per verificare l'incorrelazione dei residui ai ritardi stagionali (consideriamo inizialmente una sola stagionalità).

$$Q_s(h) = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_{kS}^2}{n-kS} \dot{\sim} \chi^2(h-P-Q), \tag{2.10}$$

  dove h è il numero di autocorrelazioni che si vogliono includere nel test, P e Q sono gli usuali ordini della componente autoregressiva stagionale e della componente a media mobile stagionale.

- Proponiamo un'ulteriore modifica per verificare l'incorrelazione dei residui ai ritardi stagionali quando è presente più di una stagionalità.

$$Q_s(h_1,\ldots,h_m) = n(n+2)\sum_{i=1}^{m}\sum_{k=1}^{h_i}\frac{\hat{\rho}_{kS_i}^2}{n-kS_i} \dot{\sim} \chi^2\left(\sum_{i=1}^{m}h_i - \sum_{i=1}^{m}(P_i+Q_i)\right), \tag{2.11}$$

  dove, per ogni stagionalità $i = 1,\ldots,m$, h_i è il numero di autocorrelazioni che si vogliono includere nel test, P_i e Q_i sono gli usuali ordini della relativa componente autoregressiva e della relativa componente a media mobile.

  È opportuno pesare il fatto che, nel caso in cui i periodi stagionali fossero multipli tra loro, la stessa autocorrelazione potrebbe essere considerata più volte.

Introduciamo un termine c che conti il numero di volte in cui un'autocorrelazione allo stesso ritardo viene considerata all'interno del calcolo di $Q_s(h)$. Teniamo conto di c sottraendolo dai gradi di libertà con cui si distribuisce la statistica test.

Esempio Se i periodi considerati sono $S_1 = 4$ e $S_2 = 12$, e si considerano $h_1 = 6$ e $h_2 = 6$ ritardi per il calcolo del test, è evidente che le autocorrelazioni ai ritardi 12 e 24 saranno considerate due volte ciascuna. In questo caso $c = 2$.

Il test con i gradi di libertà modificati è definito come

$$Q_s^*(h_1, \ldots, h_m) = n(n+2) \sum_{i=1}^{m} \sum_{k=1}^{h_i} \frac{\hat{\rho}_{kS_i}^2}{n - kS_i} \dot{\sim} \chi^2 \left(\sum_{i=1}^{m} h_i - \sum_{i=1}^{m} (P_i + Q_i) - c \right). \tag{2.12}$$

- L'ipotesi di gaussianità dei residui è verificata attraverso la statistica di Jarque e Bera [8]

$$JB = \tilde{S}^2 + \tilde{K}^2 \dot{\sim} \chi^2(2),$$

dove

$$\begin{cases} \tilde{S} = \sqrt{\frac{n}{6}} \hat{S} \dot{\sim} \mathcal{N}(0,1) \\ \tilde{K} = \sqrt{\frac{n}{24}} (\hat{K} - 3) \dot{\sim} \mathcal{N}(0,1) \end{cases},$$

dove $\hat{S}$ è la stima campionaria del coefficiente di simmetria e $\hat{K}$ è la stima campionaria del coefficiente di curtosi.

Capitolo 3

Studio di simulazione

3.1 Introduzione

Per valutare l'affidabilità di metodi statistici spesso si ricorre a metodi di simulazione (medoti Monte Carlo). Questa operazione, infatti, permette di simulare dati con caratteristiche specifiche decise a priori, ed è possibile valutare se e in che misura il modello riesce a riconoscerle. Un altro vantaggio di lavorare attraverso uno studio di simulazione è la possibilità di esplorare scenari rari, difficili da ottenere nella realtà o di cui sarebbe troppo costoso raccogliere i dati.

Nel particolare caso degli mSARIMA, ad esempio, possiamo simulare serie storiche autocorrelate secondo determinati valori dei parametri che scegliamo a priori e di cui, di conseguenza, conosciamo il vero valore. Se l'obiettivo dei modelli è quello di stimare il valore di questi parametri, possiamo quindi valutarne l'accuratezza mediante misure di distorsione e consistenza.

Ciò non potrebbe essere effettuato applicando direttamente il modello a serie storiche reali, in quanto il vero valore dei parametri del processo generatore dei dati non sarebbe noto, e potrebbe solo essere stimato.

In determinati contesti, si può ipotizzare che se un modello funziona bene (secondo criteri prestabiliti) all'interno di uno studio di simulazione, allora potrebbe funzionare altrettanto bene nel mondo reale.

3.2 Piano di simulazioni

Sono state effettuate delle simulazioni per verificare il comportamento asintotico del modello mSARIMA. Lo studio è effettuato su modelli mSARIMA con solamente due componenti stagionali. Questa scelta deriva, oltre che dal minore sforzo computazionale nell'effettuare le simulazioni, dall'utilità nell'effettuarle: sebbene infatti avere più di due cicli stagionali sia possibile, la maggior parte delle serie storiche reali presenta al massimo due cicli stagionali. Inoltre i risultati possono essere facilmente estesi a casi con più stagionalità.

In particolare sono stati misurati gli effetti:

- di diverse combinazioni degli ordini del modello: facendo interagire nello stesso ciclo stagionale e tra cicli stagionali diversi determinate combinazioni di componenti autoregressive, integrate e a media mobile. Sono stati considerati i seguenti modelli:

  1. Modello 1: mSARIMA$(1,0,0)(1,0,0)_{S_1}(1,0,0)_{S_2}$
  2. Modello 2: mSARIMA$(1,0,0)(0,0,1)_{S_1}(0,0,1)_{S_2}$
  3. Modello 3: mSARIMA$(0,0,0)(1,0,1)_{S_1}(1,0,1)_{S_2}$
  4. Modello 4: mSARIMA$(1,0,1)(0,1,0)_{S_1}(0,1,0)_{S_2}$
  5. Modello 5: mSARIMA$(0,0,0)(1,1,0)_{S_1}(1,1,0)_{S_2}$
  6. Modello 6: mSARIMA$(0,0,0)(0,1,1)_{S_1}(0,1,1)_{S_2}$
  7. Modello 7: mSARIMA$(0,1,1)(0,1,1)_{S_1}(0,1,1)_{S_2}$
  8. Modello 8: mSARIMA$(1,1,0)(1,1,0)_{S_1}(1,1,0)_{S_2}$

  Era di particolare interesse studiare l'effetto di inserire o meno componenti integrate: i modelli 1,2,3 sono stazionari, mentre i modelli 4,5,6,7,8 sono non stazionari.

- di periodi stagionali diversi: per capire se periodi multipli tra loro, considerati contemporaneamente nello stesso modello, creano problemi nella stima dei parametri e/o nell'eliminazione di autocorrelazione nella serie. I periodi utilizzati sono: $S_1 = 4$ e $S_2 = 7$; $S_1 = 4$ e $S_2 = 11$; $S_1 = 4$ e $S_2 = 12$.

- di numerosità delle serie storiche diverse: per verificare che all'aumentare della numerosità vengano soddisfatte le condizioni di non distorsione e consistenza degli stimatori. Le numerosità utilizzate sono: $n = 200, n = 500, n = 1000$ e $n = 2000$.

Per ognuno dei suddetti casi sono state effettuate 2000 simulazioni. Possiamo descrivere le fasi dello studio di simulazione come segue:

1. Si sono generate 2000 serie storiche da un generatore mSARIMA con numerosità della serie storica, ordini del modello, valori dei parametri e periodi stagionali prestabiliti.

2. Si sono stimati i relativi modelli mSARIMA con gli stessi ordini e gli stessi periodi stagionali. Ci si aspetta che utilizzare un modello che imita esattamente gli ordini del modello con i quali la serie storica è stata generata porti a dei buoni risultati in termini di stima dei parametri.

3. Si è misurato l'errore medio nella stima dei parametri, che rappresenta una misura di distorsione, tramite il calcolo di $E[\hat{\delta} - \delta]$. Ci si aspetta una diminuzione della distorsione all'aumentare della numerosità della serie storica.

4. Si è misurata la varianza degli stimatori, tramite il calcolo di $Var(\hat{\delta})$. Ci si aspetta che la varianza degli stimatori, che è equivalente alla varianza degli errori di stima, diminuisca all'aumentare della numerosità della serie storica.

5. Si è utilizzata la statistica di Ljung-Box (2.11) per verificare l'ipotesi di incorrelazione dei residui ai ritardi stagionali e si è contato quante volte il test portasse a non rifiutare l'ipotesi di incorrelazione. Ci si aspetta che il test, essendo effettuato ad un livello di significatività del 95%, non rifiuti l'ipotesi di incorrelazione circa il 95% delle volte.

6. Si è utilizzato anche il test con i gradi di libertà modificati (2.12), in modo da tenere conto di autocorrelazioni considerate più volte. Ci si aspettano gli stessi risultati del test standard quando non ci sono lag in comune, e dei gradi di copertura inferiori quando l'autocorrelazione è considerata più volte.

Nelle pagine seguenti vengono riportati i risultati dello studio di simulazione relativi al Modello 1 e al Modello 8. Sono stati scelti questi due casi perché sono facilmente confrontabili, in quanto entrambi presentano una componente autoregressiva sia nella parte non stagionale, sia in tutte le parti stagionali, e l'unica differenza tra i due modelli è l'inserimento di componenti integrate nel Modello 8. Tutti gli altri risultati sono presenti in Appendice A.

3.2.1 Modello 1: mSARIMA(1,0,0)(1,0,0)$_{S_1}$(1,0,0)$_{S_2}$

Valori dei parametri: $\phi = 0.5$, $\Phi_1 = 0.5$, $\Phi_2 = 0.5$, $\sigma^2_\varepsilon = 1$

I. **Stagionalità**: $S_1 = 4$, $S_2 = 7$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	-0.0005	-0.0009	-0.0004	0.0007	0.0041	0.0015	0.0008	0.0004
Φ_1	0.0126	0.0062	0.0013	0.0010	0.0040	0.0015	0.0008	0.0004
Φ_2	0.0113	0.0038	0.0025	0.0016	0.0040	0.0015	0.0008	0.0004
σ^2_ε	-0.0820	-0.0320	-0.0155	-0.0089	0.0132	0.0044	0.0021	0.0010

Tabella 3.1: Distorsione e varianza degli stimatori, Modello 1, $S_1 = 4, S_2 = 7$

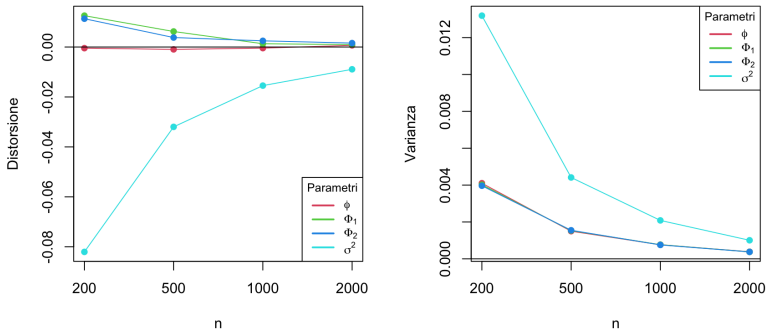


Figura 3.1: Distorsione e varianza degli stimatori, Modello 1, $S_1 = 4, S_2 = 7$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	95.35	94.65	95.20	95.20	95.35	94.65	95.20	95.20
$h_1 = 3, h_2 = 3$	95.15	95.10	94.80	94.70	95.15	95.10	94.80	94.70
$h_1 = 4, h_2 = 4$	95.20	95.60	95.35	95.15	95.20	95.60	95.35	95.15
$h_1 = 5, h_2 = 5$	94.90	96.00	94.95	95.25	94.90	96.00	94.95	95.25

Tabella 3.2: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 1, $S_1 = 4, S_2 = 7$

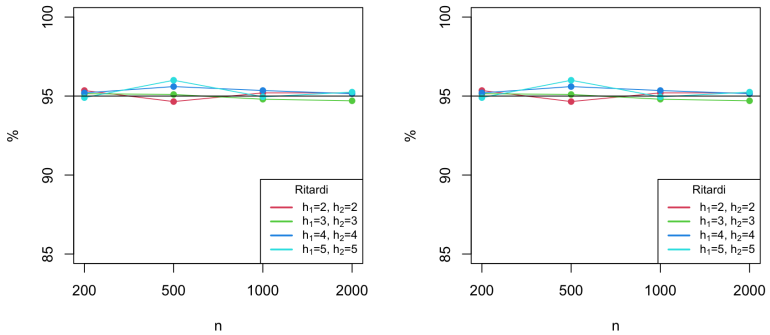


Figura 3.2: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 1, $S_1 = 4, S_2 = 7$

II. **Stagionalità**: $S_1 = 4$, $S_2 = 11$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.0023	0.0020	0.0003	0.0004	0.0039	0.0016	0.0008	0.0004
Φ_1	0.0136	0.0049	0.0036	0.0018	0.0039	0.0015	0.0007	0.0004
Φ_2	0.0117	0.0024	0.0014	0.0018	0.0043	0.0015	0.0007	0.0004
σ^2_ε	-0.1128	-0.0429	-0.0205	-0.0095	0.0136	0.0046	0.0021	0.0010

Tabella 3.3: Distorsione e varianza degli stimatori, Modello 1, $S_1 = 4, S_2 = 11$

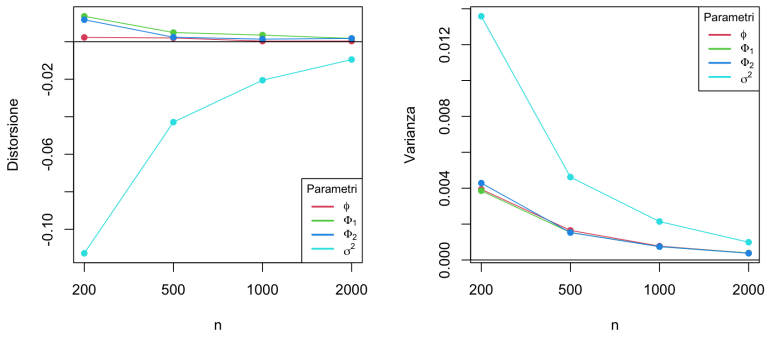


Figura 3.3: Distorsione e varianza degli stimatori, Modello 1, $S_1 = 4, S_2 = 11$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.55	95.05	93.80	95.00	94.55	95.05	93.80	95.00
$h_1 = 3, h_2 = 3$	94.80	95.25	94.80	95.00	94.80	95.25	94.80	95.00
$h_1 = 4, h_2 = 4$	95.50	95.55	95.50	95.45	95.50	95.55	95.50	95.45
$h_1 = 5, h_2 = 5$	94.65	95.80	95.65	95.00	94.65	95.80	95.65	95.00

Tabella 3.4: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 1, $S_1 = 4, S_2 = 11$

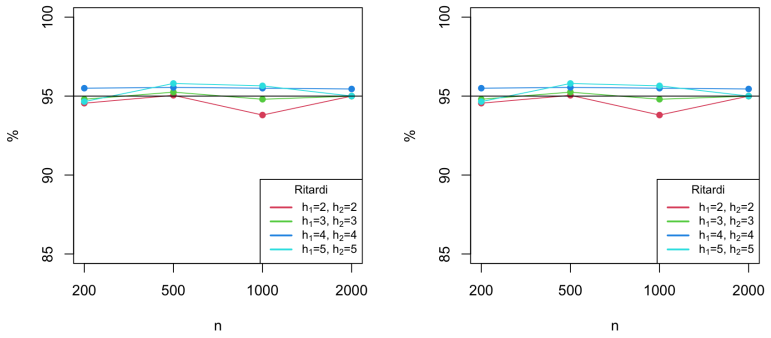


Figura 3.4: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 1, $S_1 = 4, S_2 = 11$

III. **Stagionalità**: $S_1 = 4$, $S_2 = 12$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.0067	0.0025	0.0019	0.0012	0.0043	0.0015	0.0008	0.0004
Φ_1	0.0068	0.0019	0.0001	0.0003	0.0042	0.0017	0.0008	0.0004
Φ_2	0.0094	0.0077	0.0028	0.0012	0.0043	0.0015	0.0008	0.0004
σ^2_ε	-0.1221	-0.0459	-0.0229	-0.0117	0.0135	0.0046	0.0021	0.0010

Tabella 3.5: Distorsione e varianza degli stimatori, Modello 1, $S_1 = 4, S_2 = 12$

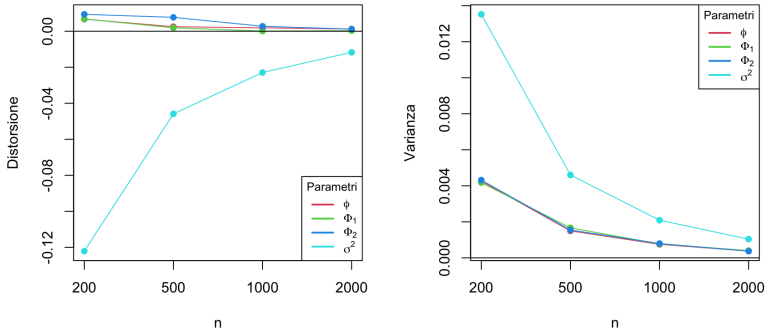


Figura 3.5: Distorsione e varianza degli stimatori, Modello 1, $S_1 = 4, S_2 = 12$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.45	94.80	94.50	94.95	94.45	94.80	94.50	94.95
$h_1 = 3, h_2 = 3$	96.90	96.65	96.85	96.25	93.75	93.30	92.85	93.20
$h_1 = 4, h_2 = 4$	96.50	96.30	96.05	96.30	93.75	93.45	93.75	93.80
$h_1 = 5, h_2 = 5$	96.50	96.45	96.05	96.55	94.45	93.85	94.10	94.25

Tabella 3.6: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 1, $S_1 = 4, S_2 = 12$

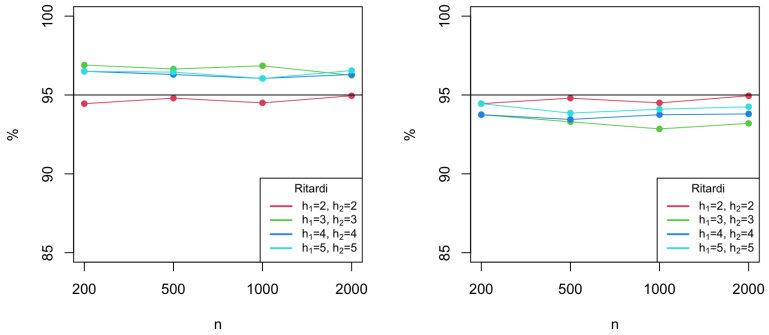


Figura 3.6: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 1, $S_1 = 4, S_2 = 12$

3.2.2 Modello 8: mSARIMA$(1,1,0)(1,1,0)_{S_1}(1,1,0)_{S_2}$

Valori dei parametri $\phi = 0.5, \Phi_1 = 0.5, \Phi_2 = 0.5, \sigma^2_\varepsilon = 1$

I. **Stagionalità**: $S_1 = 4$, $S_2 = 7$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.0076	0.0018	0.0019	0.0016	0.0068	0.0025	0.0013	0.0006
Φ_1	0.0109	0.0059	0.0018	0.0008	0.0070	0.0027	0.0013	0.0006
Φ_2	0.0120	0.0044	0.0021	0.0006	0.0070	0.0026	0.0013	0.0006
σ^2_ε	0.0230	0.0100	0.0041	0.0021	0.0177	0.0069	0.0034	0.0017

Tabella 3.7: Distorsione e varianza degli stimatori, Modello 8, $S_1 = 4, S_2 = 7$

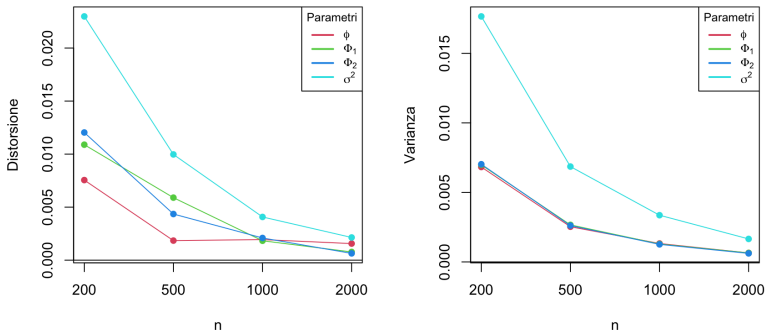


Figura 3.7: Distorsione e varianza degli stimatori, Modello 8, $S_1 = 4, S_2 = 7$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.15	94.50	94.60	94.20	94.15	94.50	94.60	94.20
$h_1 = 3, h_2 = 3$	94.50	94.95	93.70	94.55	94.50	94.95	93.70	94.55
$h_1 = 4, h_2 = 4$	95.20	94.60	94.40	95.60	95.20	94.60	94.40	95.60
$h_1 = 5, h_2 = 5$	95.15	94.05	95.00	95.35	95.15	94.05	95.00	95.35

Tabella 3.8: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 8, $S_1 = 4, S_2 = 7$

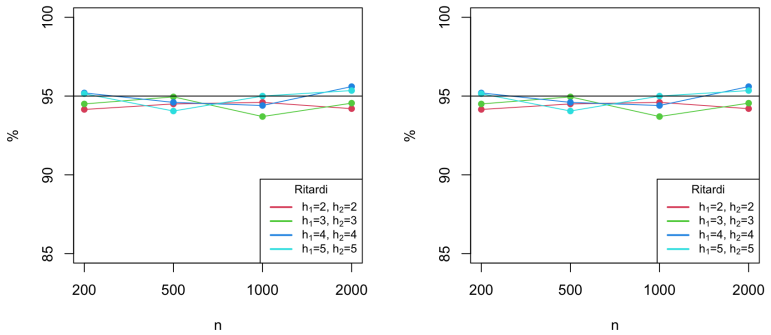


Figura 3.8: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 8, $S_1 = 4, S_2 = 7$

II. **Stagionalità**: $S_1 = 4$, $S_2 = 11$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.0064	0.0038	0.0009	0.0008	0.0073	0.0027	0.0013	0.0006
Φ_1	0.0123	0.0056	0.0027	0.0011	0.0073	0.0029	0.0013	0.0006
Φ_2	0.0115	0.0033	0.0010	0.0009	0.0071	0.0028	0.0012	0.0006
σ^2_ε	0.0279	0.0124	0.0036	0.0017	0.0176	0.0070	0.0033	0.0018

Tabella 3.9: Distorsione e varianza degli stimatori, Modello 8, $S_1 = 4, S_2 = 11$

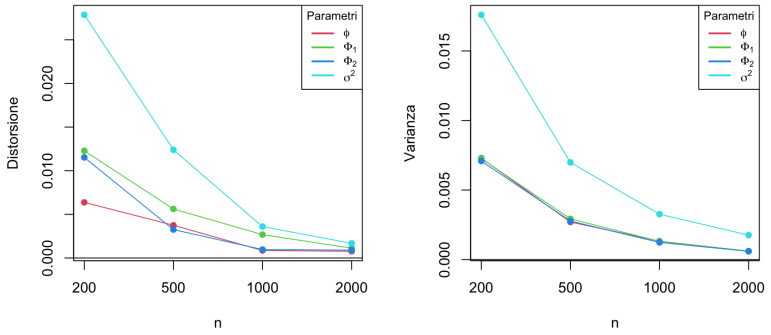


Figura 3.9: Distorsione e varianza degli stimatori, Modello 8, $S_1 = 4, S_2 = 11$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	95.40	93.95	94.85	94.15	95.40	93.95	94.85	94.15
$h_1 = 3, h_2 = 3$	95.50	94.50	94.70	95.00	95.50	94.50	94.70	95.00
$h_1 = 4, h_2 = 4$	95.30	94.45	95.05	94.25	95.30	94.45	95.05	94.25
$h_1 = 5, h_2 = 5$	95.10	95.15	94.40	94.20	95.10	95.15	94.40	94.20

Tabella 3.10: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 8, $S_1 = 4, S_2 = 11$

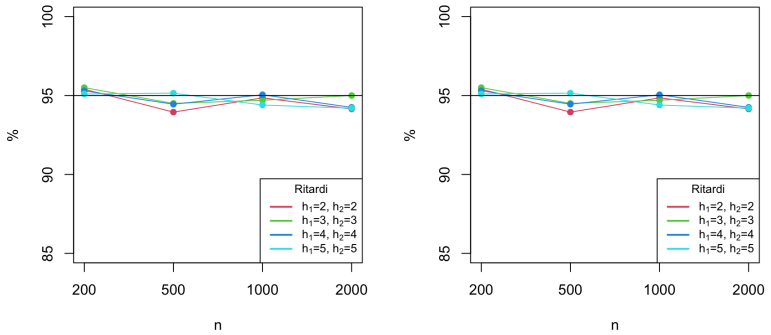


Figura 3.10: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 8, $S_1 = 4, S_2 = 11$

III. **Stagionalità**: $S_1 = 4$, $S_2 = 12$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.0097	0.0044	0.0015	0.0010	0.0074	0.0027	0.0013	0.0006
Φ_1	0.0077	0.0035	0.0019	0.0018	0.0081	0.0029	0.0013	0.0007
Φ_2	0.0146	0.0064	0.0017	0.0008	0.0078	0.0028	0.0014	0.0006
σ^2_ε	0.0315	0.0096	0.0039	0.0032	0.0184	0.0070	0.0033	0.0016

Tabella 3.11: Distorsione e varianza degli stimatori, Modello 8, $S_1 = 4, S_2 = 12$

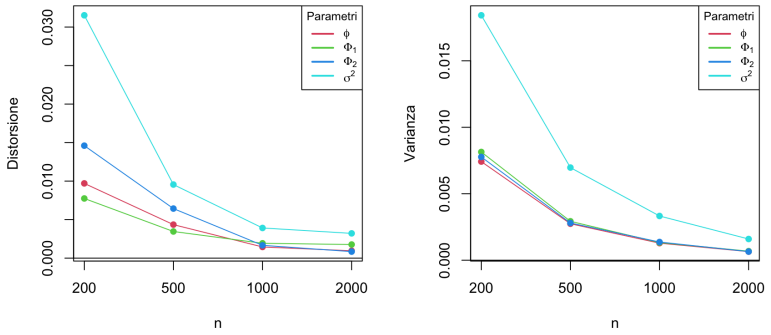


Figura 3.11: Distorsione e varianza degli stimatori, Modello 8, $S_1 = 4, S_2 = 12$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.70	94.80	94.65	95.20	94.70	94.80	94.65	95.20
$h_1 = 3, h_2 = 3$	96.05	96.75	96.50	96.75	92.70	93.55	93.05	93.85
$h_1 = 4, h_2 = 4$	95.85	96.45	96.35	96.45	92.85	94.00	94.25	94.20
$h_1 = 5, h_2 = 5$	95.25	96.45	95.70	96.50	92.85	94.20	93.70	94.65

Tabella 3.12: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 8, $S_1 = 4, S_2 = 12$

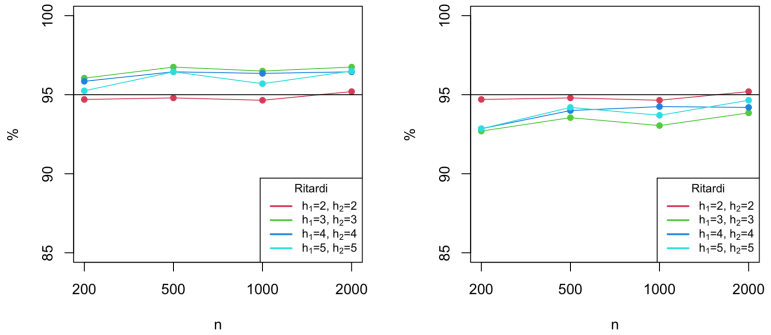


Figura 3.12: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 8, $S_1 = 4, S_2 = 12$

3.2.3 Valutazioni

In tutti i casi considerati è evidente la differenza che si ottiene utilizzando serie storiche di numerosità diverse: più è alta la numerosità, più la distorsione nella stima dei parametri diminuisce e con essa la varianza degli stimatori. Ciò conferma il buon funzionamento asintotico del modello mSARIMA, gli stimatori sono consistenti. Piccole variazioni, come si può notare ad esempio in Figura (A.25), contrarie a quanto affermato, sono da imputare al basso numero di simulazioni.

Per quanto riguarda le analisi diagnostiche effettuate, in generale si evince che il test di Ljung-Box ai ritardi stagionali non rifiuta l'ipotesi di incorrelazione dei residui circa il 95% delle volte. Questo risultato è piuttosto positivo in quanto significa che il modello riesce ad individuare ed eliminare correttamente la correlazione presente nella serie storica ai ritardi stagionali. Inoltre conferma le aspettative in quanto il test è effettuato con un livello di significatività del 95%, dunque è normale che abbia un grado di copertura osservato simile. Si notano delle difficoltà con le numerosità molto basse della serie storica, nei modelli 6 e 7, in particolare con la stagionalità $S_1 = 4$, $S_2 = 12$.

È stata valutata l'efficacia dell'utilizzo del test di Ljung-Box con i gradi di libertà modificati, che nelle simulazioni effettuate ha ottenuto valori diversi rispetto al test standard esclusivamente per la combinazione di stagionalità $S_1 = 4$ e $S_2 = 12$, e solo quando si consideravano almeno 3 lag per il periodo 4 relativo alla stagionalità S_1. In questi casi, l'autocorrelazione al ritardo 12 veniva considerata due volte, e quindi la statistica test si distribuiva come un chi-quadro, con un grado di libertà in meno rispetto a quelli del test standard. In queste situazioni, con un grado di libertà in meno, il test tende a rifiutare l'ipotesi nulla più spesso. Nella maggior parte delle situazioni il suo utilizzo risulta quindi adeguato, perché il test standard sovrastimava il numero di volte in cui i residui fossero incorrelati, mentre il test con le opportune modifiche ai gradi di libertà, rifiutando più spesso, portava ad un grado di copertura leggermente inferiore, più vicino al 95%. In qualche caso, e in particolare nel modello 4, il suo utilizzo è risultato peggiore rispetto a quello standard, perché quest'ultimo aveva già un grado di copertura inferiore al 95%.

Capitolo 4

Previsioni

4.1 Introduzione

Uno degli scopi dell'analisi delle serie storiche è quello di fornire previsioni sull'andamento delle variabili oggetto di interesse. Si presuppone in particolare che se si trattano adeguatamente le informazioni disponibili per il passato delle variabili di interesse, queste possano essere utili a stimare, con un'inevitabile approssimazione, gli avvenimenti futuri.

L'accuratezza che un modello possiede nell'effettuare una previsione, inoltre, può essere determinante nella scelta tra diversi modelli alternativi. Infatti se si analizza solamente la capacità di adattamento ai dati, ad esempio attraverso indici come R^2 o AIC, si può rischiare di selezionare un modello eccellente nello spiegare i dati *in-sample*, ma completamente inutile per effettuare previsioni. È utile quindi confrontare la capacità previsiva attraverso diversi indicatori come MAE o MAPE e utilizzare anche questi criteri per la scelta del modello.

Data una serie storica $y_1, \dots, y_n$, fare previsione significa cercare di ottenere i valori di y_{n+k}. Questi valori non sono altro che la realizzazione di una variabile casuale, Y_{n+k}, che non è nota, ma può essere stimata attraverso il passato del processo. Il passato è chiamato *informazione al tempo t*, ed è composto dalle infinite variabili casuali che compongono il processo fino all'istante t. È indicato come

$$I_t = \{Y_t, Y_{t-1}, \dots, Y_0, Y_{-1}, \dots\}.$$

Y_{n+k} sarà quindi stimata con una variabile casuale chiamata *previsore*, $\hat{Y}_{n+k}$, che è funzione del passato.

$$\hat{Y}_{n+k} = g(I_t).$$

Questa funzione $g(\cdot)$ dev'essere tale da minimizzare una funzione di perdita, comunemente l'errore di previsione, definito come

$$e_{t+k} = Y_{t+k} - \hat{Y}_{t+k}.$$

In particolare deve fare in modo che l'errore di previsione sia nullo in media e che la sua variabilità sia minima a priori.

Il previsore che soddisfa queste proprietà è il valore atteso condizionato [9], definito come

$$\hat{Y}_{n+k} = E\left[Y_{n+k} \middle| I_t\right],$$

e quando questo valore atteso esiste, definisce il *previsore ottimo*.

Nel contesto dei modelli mSARIMA, data una serie storica $\{y_t\}_{t=1}^n$, e dato un modello $mSARIMA(p,d,q) \times (P_1,D_1,Q_1)_{S_1} \times \cdots \times (P_m,D_m,Q_m)_{S_m}$ che cerca di descrivere il processo generatore dei dati Y_t, vogliamo ottenere i valori di y_{n+k}. Come abbiamo visto quindi, la sua previsione è

$$\begin{aligned}\hat{y}_{n+k} &= E\left[Y_{n+k} \middle| I_t\right] \\ &\approx E\left[Y_{n+k} \middle| Y_n = y_n, \ldots, Y_1 = y_1\right].\end{aligned}$$

in cui abbiamo considerato questa approssimazione perché non disponiamo dell'intera informazione al tempo t, ma solo quella fornita dalla serie storica $\{y_t\}_{t=1}^n$.

Per il calcolo del valore atteso condizionato valgolo le seguenti regole:

$$E\left[Y_{t+j} \middle| Y_t = y_t, \ldots, Y_1 = y_1\right] = \begin{cases} y_{t+j} & \text{se } j \leq 0 \\ \hat{y}_{t+j} & \text{se } j > 0 \end{cases};$$

$$E\left[\varepsilon_{t+j} \middle| Y_t = y_t, \ldots, Y_1 = y_1\right] = \begin{cases} e_{t+j} & \text{se } j \leq 0 \\ 0 & \text{se } j > 0 \end{cases}.$$

Spesso può essere utile una previsione intervallare, e sapendo che

$$\hat{Y}_{t+k} \sim \mathcal{N}(y_{t+k}, \mathrm{Var}(e_{t+k}))$$

si può definire un intervallo di previsione di livello $(1-\alpha)$ come

$$\hat{y}_{t+k} \pm z_{1-\frac{\alpha}{2}} \sqrt{\mathrm{Var}(e_{t+k})},$$

dove $z_{1-\frac{\alpha}{2}}$ è il percentile di ordine $1-\frac{\alpha}{2}$ di una distribuzione $\mathcal{N}(0,1)$.

Per misurare l'accuratezza di una previsione, sono disponibili diversi indicatori:

- **Mean Error**: $ME = \frac{1}{n}\sum_{t=1}^{n} e_t$

  Questo indicatore tende ad essere piccolo perché gli errori positivi e negativi si compensano. A questo scopo ci informa se è presente distorsione nella previsione.

- **Mean Absolute Error**: $MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|$

  È una misura facilmente interpretabile perché rappresenta quanto sono grandi gli errori assoluti in media, nella stessa unità di misura della variabile analizzata.

- **Mean Square Error**: $MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$

  Viene spesso utilizzato per aspetti di ottimizzazione.

  Per questioni di interpretabilità viene spesso usato il *Root Mean Square Error*: $RMSE = \sqrt{MSE}$, che riporta i valori alla stessa scala della variabile.

- **Mean Absolute Percentage Error**: $MAPE = \frac{1}{n}\sum_{t=1}^{n} |\frac{e_t}{y_t}|$

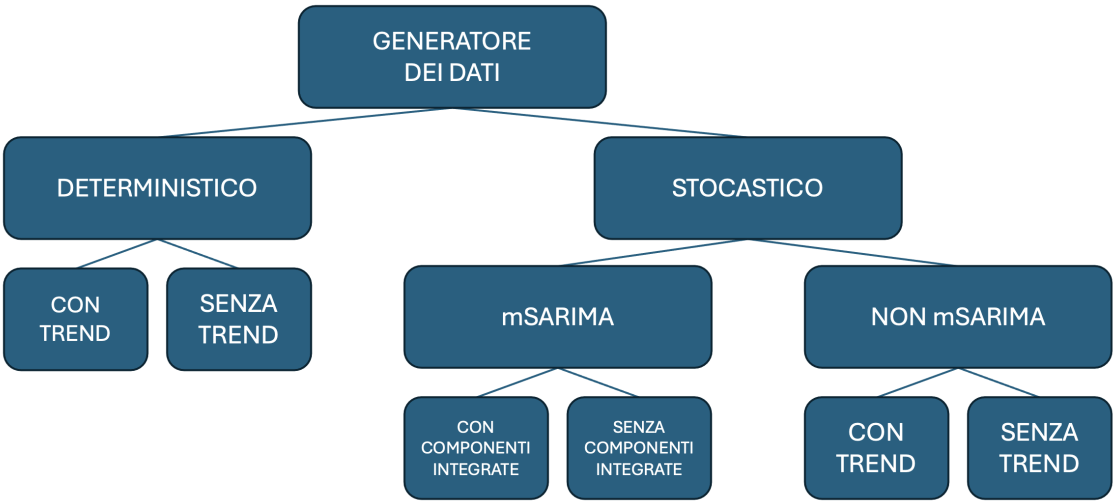
  È l'errore assoluto medio definito in termini percentuali, così non dipende dalla scala della variabile ed è più facilmente confrontabile.

Per misurare questi indicatori in una serie storica è utile dividere quest'ultima in una parte *in-sample* da utilizzare per stimare i parametri del modello, e una parte *out-of-sample* di cui il modello cercherà di prevedere i valori. Disponendo dei valori veri si potrà valutare l'accuratezza delle previsioni con uno degli indicatori precedentemente elencati.

4.2 Analisi effettuate

È interessante studiare l'accuratezza previsiva del modello mSARIMA in relazione al tipo di processo generatore dei dati e attraverso un confronto con un altro modello generalmente utilizzato per effettuare previsioni da serie storiche.

Nelle seguenti simulazioni sono stati utilizzati tre tipi di generatori, e di ognuno si è valutata la differenza tra la generazione di serie storiche stazionarie e serie storiche non stazionarie:



- Un generatore deterministico secondo la funzione (descritta in [10]):

$$Y_t = \delta T_t + \alpha S_t^1 + \beta S_t^2 + \gamma \varepsilon_t, \qquad t = 1, \ldots, n, \tag{4.1}$$

dove Y_t è la serie storica generata, T_t è il trend, S_t^1 e S_t^2 sono rispettivamente la prima e la seconda componente stagionale e ε_t è un termine di errore. δ, α, β e γ sono dei parametri che controllano il contributo delle varie componenti. Infine n denota la lunghezza della serie storica.

Quando il generatore era deterministico le componenti sono state così calcolate: il trend è stato generato attraverso una funzione lineare del tempo $T_t = N_1(t + \frac{n}{2}(N_2 - 1))$ dove N_1 e N_2 sono due variabili casuali $\mathcal{N}(0,1)$ indipendenti; entrambe le componenti stagionali sono composte da cinque coppie di termini di Fourier con coefficienti casuali $\mathcal{N}(0,1)$; sia il trend sia le componenti stagionali sono state normalizzate per ottenere media nulla e varianza unitaria.

Si è voluta studiare la differenza tra la presenza e l'assenza di trend, facendo dunque variare il parametro δ nei valori 1 e 0.

- Un generatore stocastico secondo la stessa funzione (4.1).

  Quando il generatore era stocastico le componenti sono state così calcolate: il trend è stato generato attraverso un modello ARIMA(0,1,0) con errori normali standard; entrambe le componenti stagionali sono composte da cinque coppie di termini di Fourier con coefficienti casuali $\mathcal{N}(0,1)$ ma con l'aggiunta di un termine d'errore $\mathcal{N}(0, \sigma^2)$, in cui il parametro σ^2 controlla quanto dev'essere stocastica la componente stagionale; sia il trend sia le componenti stagionali sono state normalizzate per ottenere media nulla e varianza unitaria.

  Anche in questo caso si è voluta studiare la differenza tra la presenza e l'assenza di trend, facendo variare il parametro δ nei valori 1 e 0.

- Un generatore stocastico da modelli mSARIMA. Sono stati utilizzati gli stessi otto modelli generatori dello studio di simulazione e si è fatta particolare attenzione alla differenza tra modelli che generavano serie storiche stazionarie e quelli che invece avevano componenti integrate.

Le performance sono confrontate con quelle di uno dei modelli più affermati in letteratura, il TBATS. I modelli TBATS (Trigonometric, Box-Cox, ARMA errors, Trend and Seasonal components) sono una classe di modelli di decomposizione, che suddividono dunque la serie nelle componenti di trend, livello e stagionalità. Il modello TBATS si basa su una trasformazione Box-Cox per stabilizzare la varianza della serie, su componenti trigonometriche per catturare la stagionalità, su componenti autoregressive e a media mobile (ARMA) per modellare gli errori e su componenti per il trend e la stagionalità che possono variare nel tempo.

Il modello TBATS, descritto nel dettaglio in [4], può essere scritto come segue:

$$
\begin{aligned}
y_t^{(\omega)} &= \begin{cases} \frac{y_t^{\omega}-1}{\omega}, & \omega \neq 0, \\ \log\ y_t, & \omega = 0, \end{cases} \\
y_t^{(\omega)} &= l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t, \\
l_t &= l_{t-1} + \phi b_{t-1} + \alpha d_t, \\
b_t &= (1-\phi)b + \phi b_{t-1} + \beta d_t, \\
s_t^{(i)} &= \sum_{j=1}^{k_i} s_{j,t}, \\
s_{j,t}^{(i)} &= s_{j,t-1}^{(i)} \cos\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin\lambda_j^{(i)} + \gamma_1^{(i)} d_t, \\
s_{j,t}^{*(i)} &= -s_{j,t-1} \sin\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos\lambda_j^{(i)} + \gamma_2^{(i)} d_t, \\
d_t &= \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t,
\end{aligned}
$$

dove y_t è la serie storica originale al tempo t, ω è il parametro di trasformazione di Box-Cox, $y_t^{(\omega)}$ è la serie storica trasformata al tempo t, l_t è il livello locale al tempo t, b_t è il trend di breve termine al tempo t, b è il trend di lungo termine, $m_1, \ldots, m_T$ sono i periodi stagionali, $s_t^{(i)}$ è la i-esima componente stagionale al tempo t, k_i è il numero di armoniche per la i-esima componente stagionale, ϕ, α, β, $\gamma_1^{(i)}$ e $\gamma_2^{(i)}$ sono parametri di lisciamento e $\gamma_j^{(i)} = \frac{2\pi j}{m_i}$, d_t è un processo $ARMA(p,q)$ con parametri φ_i per la componente autoregressiva e θ_i per la componente a media mobile, ϵ_t è un white noise con media zero e varianza costante σ^2.

Quando il processo generatore dei dati era un modello mSARIMA, il modello che si è utilizzato per stimarne i parametri e per fare previsioni è il medesimo che si è utilizzato per generare i dati, ovvero ne rispecchia tutti gli ordini. Quando, invece il processo generatore dei dati derivava dalla funzione (4.1), le componenti stagionali erano presenti nella funzione in maniera additiva, e quindi era possibile rimuoverle tramite differenziazioni. Applicare le differenziazioni in un modello di questo tipo porta a generarsi della correlazione con gli errori ai ritardi stagionali e ai ritardi delle combinazioni dei vari periodi stagionali.

Esempio Se il processo generatore è del tipo

$$y_t = S_t^1 + S_t^2 + \varepsilon_t,$$

dove S_t^1 e S_t^2 sono la prima e la seconda stagionalità, con relativi periodi 4 e 7, queste sono tali che

$$\begin{cases} S_t^1 = S_{t+4}^1 = S_{t+8}^1 = \cdots \\ S_t^2 = S_{t+7}^2 = S_{t+14}^2 = \cdots \end{cases}, \tag{4.2}$$

allora, applicando la prima differenziazione di periodo 4 e sfruttando le proprietà (4.2), si ha

$$\begin{aligned} y_t^* &= y_t - y_{t-4} \\ &= S_t^1 + S_t^2 + \varepsilon_t - (S_{t-4}^1 + S_{t-4}^2 + \varepsilon_{t-4}) \\ &= S_t^2 - S_{t-4}^2 + \varepsilon_t - \varepsilon_{t-4}, \end{aligned}$$

e, applicando un'altra differenziazione di periodo 7,

$$\begin{aligned} y_t^{**} &= y_t^* - y_{t-7}^* \\ &= S_t^2 - S_{t-4}^2 + \varepsilon_t - \varepsilon_{t-4} - (S_{t-7}^2 - S_{t-11}^2 + \varepsilon_{t-7} - \varepsilon_{t-11}) \\ &= \varepsilon_t - \varepsilon_{t-4} - \varepsilon_{t-7} + \varepsilon_{t-11}. \end{aligned}$$

È quindi sensato inserire nel modello, oltre alle due differenziazioni ai ritardi stagionali, delle componenti a media mobile stagionali, che colgono esattamente le autocorrelazioni rimaste dopo le differenziazioni.

Alla luce di queste considerazioni, i modelli mSARIMA scelti per gestire queste serie storiche sono del tipo $mSARIMA(0,0,0)(0,1,1)_4(0,1,1)_7$ in assenza di trend, e $mSARIMA(0,1,1)(0,1,1)_4(0,1,1)_7$ in presenza di trend. Quando è presente il trend è stata inserita una componente integrata non stagionale, e la relativa componente a media mobile che l'applicazione della differenziazione comporta. A causa della natura additiva della stagionalità in questo tipo di generatori, indipendentemente dal fatto che questi fossero deterministici o stocastici, il modello stimato aveva gli stessi ordini sia per generatori deterministici che per generatori stocastici.

In tutti i casi, invece, il modello TBATS è stato stimato con delle procedure automatiche: il modello è in grado di determinare automaticamente se inserire o meno il trend o il trend smorzato. Per garantire la confrontabilità dei risultati, è stato imposto di non effettuare trasformazioni di Box-Cox e di utilizzare errori ARMA, e si sono definite a priori le stagionalità corrette.

Per ogni modello generatore di dati e per ogni numerosità diversa della serie storica, sono state effettuate 500 simulazioni. Possiamo descrivere le fasi di questo

studio come segue:

1. Sono state generate 500 serie storiche da ogni tipo di generatore con valori dei parametri prestabiliti e per le seguenti numerosità: $n = 228, n = 528, n = 1028, n = 2028$.

2. Ogni serie generata è stata suddivisa: le parti *in-sample* sono costituite dalle rispettive prime $n = 200, n = 500, n = 1000, n = 2000$ osservazioni; le parti *out-of-sample* sono costituite dalle relative ultime 28 osservazioni.

3. Attraverso il solo utilizzo della parte *in-sample* della serie storica, sono stati stimati i relativi modelli TBATS e mSARIMA, con le componenti adeguate mostrate in precedenza.

4. Sono state effettuate previsioni a 28 passi in avanti attraverso entrambi i modelli.

5. Le previsioni si sono confrontate con i veri valori *out-of-sample*, e se ne è misurata l'accuratezza tramite l'indicatore MAE.

Nelle seguenti pagine sono mostrati i confronti delle prestazioni previsive in termini di MAE per i vari processi generatori. Per quanto riguarda i generatori mSARIMA, si sono selezionati esclusivamente i Modelli 1 e 8, per questioni di confrontabilità: entrambi hanno componenti autoregressive sia non stagionali sia stagionali, e l'unica differenza tra i due modelli è la presenza di componenti integrate nel Modello 8. Tutti i risultati relativi agli altri generatori mSARIMA sono presenti nell'Appendice B.

4.2.1 Generatore deterministico con trend

Valori dei parametri $\delta = 1, \alpha = 1, \beta = 1, \gamma = 0.4, \sigma_\varepsilon^2 = 1$

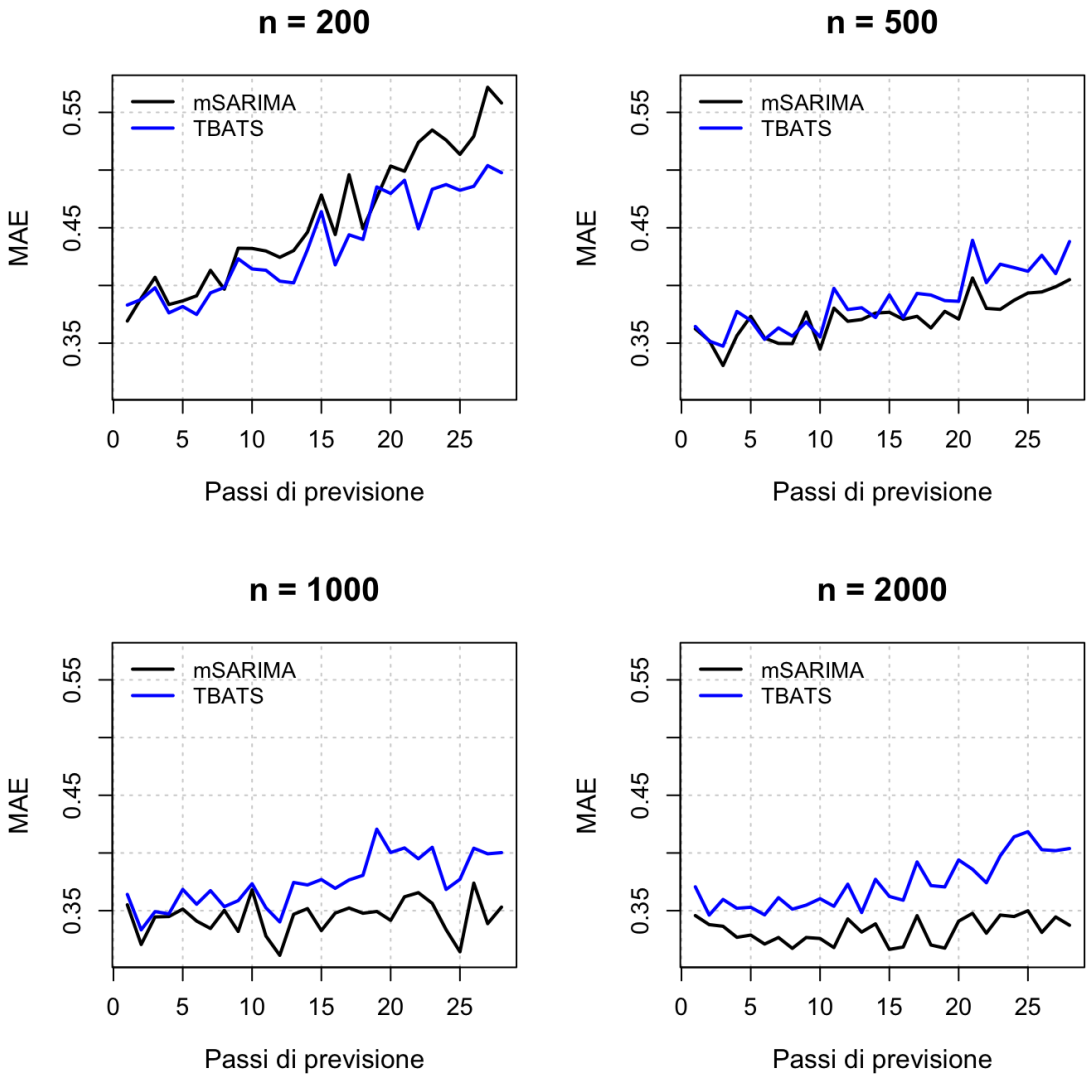


Figura 4.1: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Generatore deterministico con trend.

Il modello TBATS performa meglio rispetto al modello mSARIMA, in termini di MAE, quando la numerosità della serie storica è molto bassa (n=200). In casi di serie storica di numerosità più elevata, invece, il modello migliore è mSARIMA, che aumenta le proprie prestazioni ottenendo stime sempre più precise all'aumentare della lunghezza della serie storica, più di quanto faccia TBATS.

4.2.2 Generatore deterministico senza trend

Valori dei parametri $\delta = 0, \alpha = 1, \beta = 1, \gamma = 0.4, \sigma_\varepsilon^2 = 1$

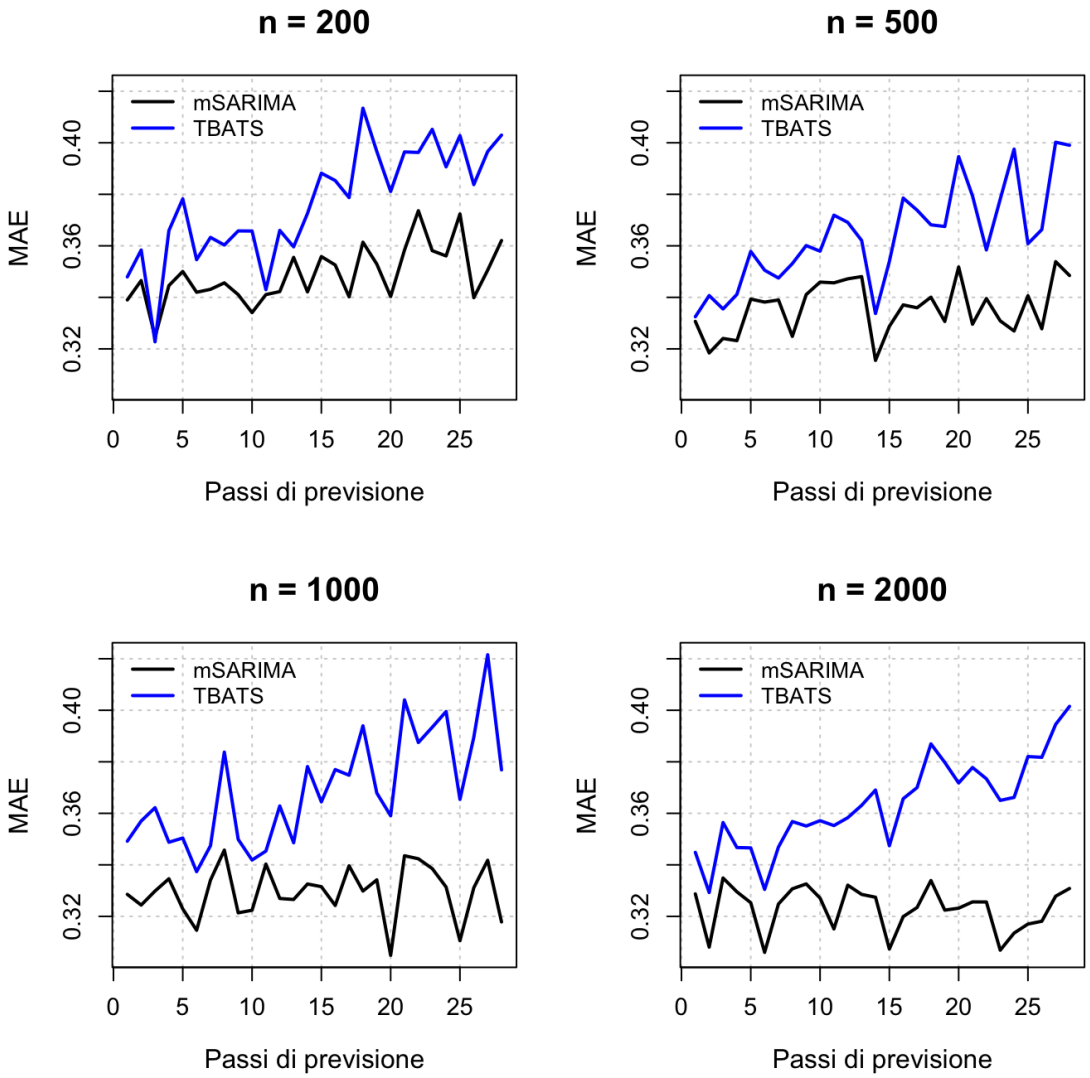


Figura 4.2: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Generatore deterministico senza trend.

Il modello mSARIMA ottiene MAE sempre più bassi rispetto a quelli ottenuti dal modello TBATS. Il divario tra i due modelli aumenta all'aumentare della lunghezza della serie storica. L'assenza di trend mantiene il MAE basso anche con serie di bassa numerosità.

4.2.3 Generatore stocastico con trend

Valori dei parametri $\delta = 1, \alpha = 1, \beta = 1, \gamma = 0.4, \sigma_\varepsilon^2 = 1, \sigma^2 = 0.025$

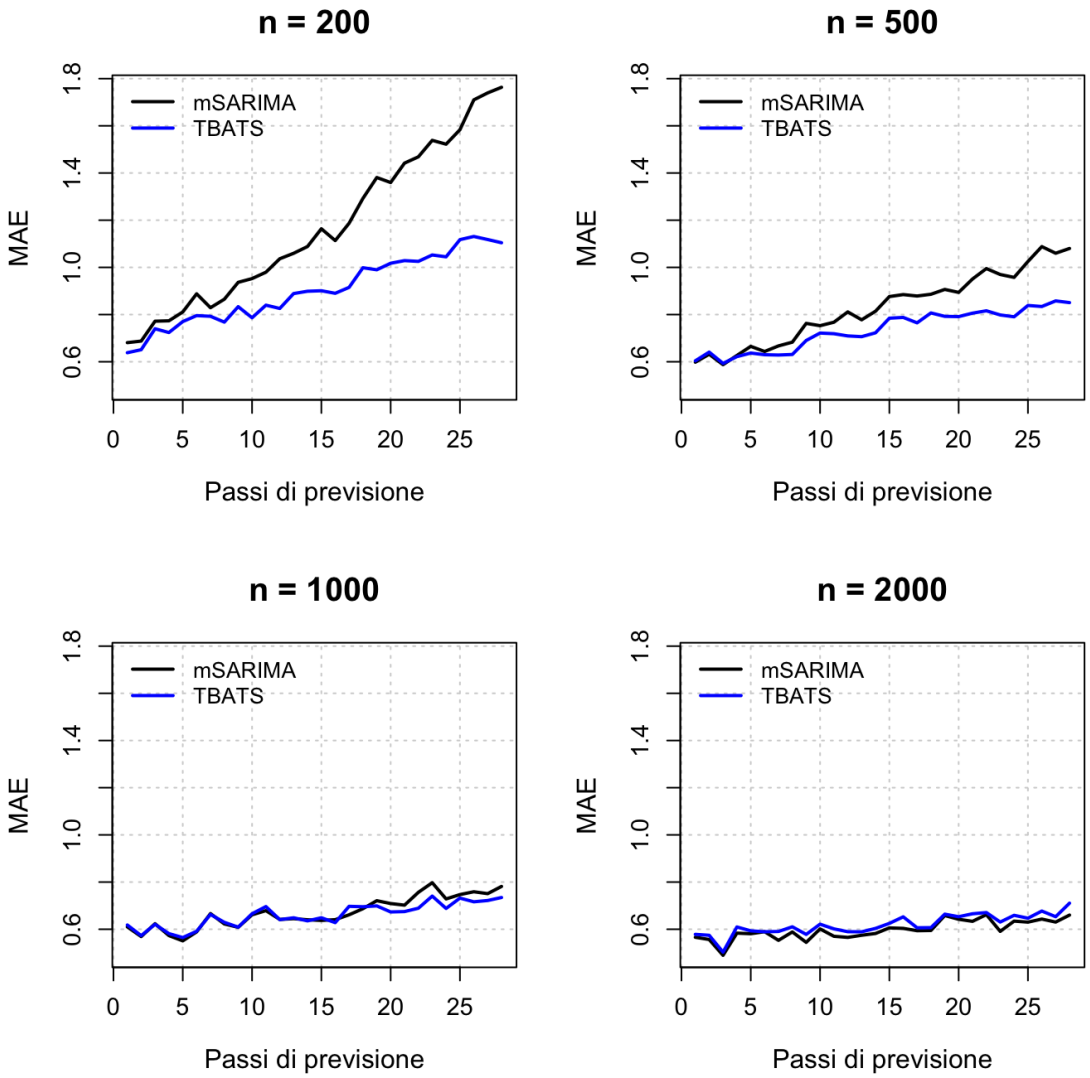


Figura 4.3: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Generatore stocastico con trend.

La presenza del trend associata ad un generatore di tipo stocastico rende peggiori le prestazioni di mSARIMA rispetto a quelle di TBATS per numerosità basse, comparabili con n=1000 e leggermente migliori quando la serie ha lunghezza n=2000.

4.2.4 Generatore stocastico senza trend

Valori dei parametri $\delta = 0, \alpha = 1, \beta = 1, \gamma = 0.4, \sigma_\varepsilon^2 = 1, \sigma^2 = 0.025$

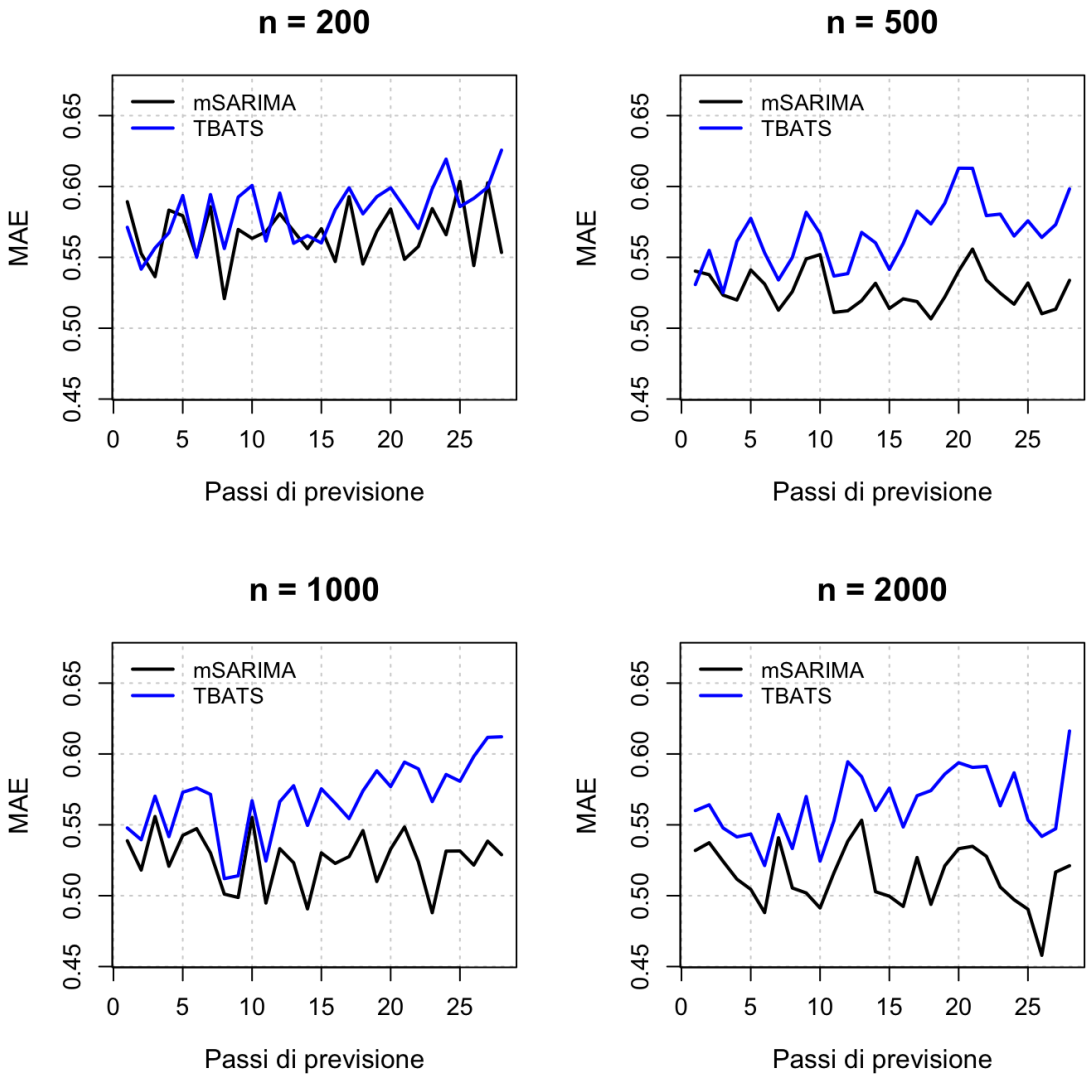


Figura 4.4: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Generatore stocastico senza trend.

Il modello mSARIMA ottiene MAE sempre più bassi rispetto a quelli ottenuti dal modello TBATS. Il divario tra i due modelli aumenta all'aumentare della lunghezza della serie storica.

4.2.5 Modello 1: mSARIMA$(1,0,0)(1,0,0)_4(1,0,0)_7$

Valori dei parametri $\phi = 0.5, \Phi_1 = 0.5, \Phi_2 = 0.5, \sigma_\varepsilon^2 = 1$

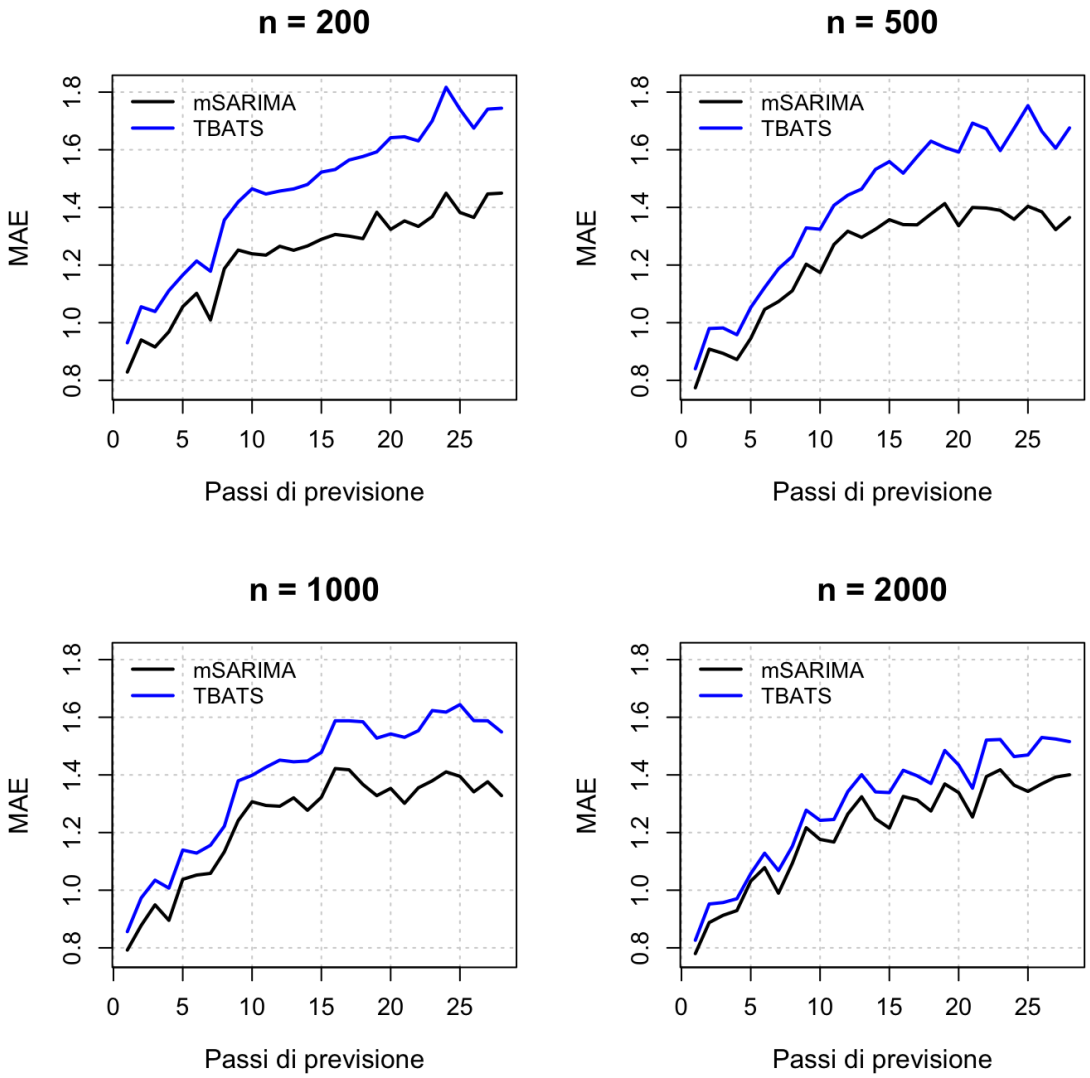


Figura 4.5: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Modello 1.

Il modello mSARIMA performa sempre meglio di TBATS, in termini di MAE, specialmente per serie con bassa numerosità. Il comportamento del MAE di mSARIMA, in realtà, è piuttosto simile indipendentemente dalla numerosità della serie storica, a differenza di quello del TBATS, che assume valori più bassi quando le serie sono più lunghe.

4.2.6 Modello 8: mSARIMA$(1,1,0)(1,1,0)_4(1,1,0)_7$

Valori dei parametri $\phi = 0.5, \Phi_1 = 0.5, \Phi_2 = 0.5, \sigma_\varepsilon^2 = 1$

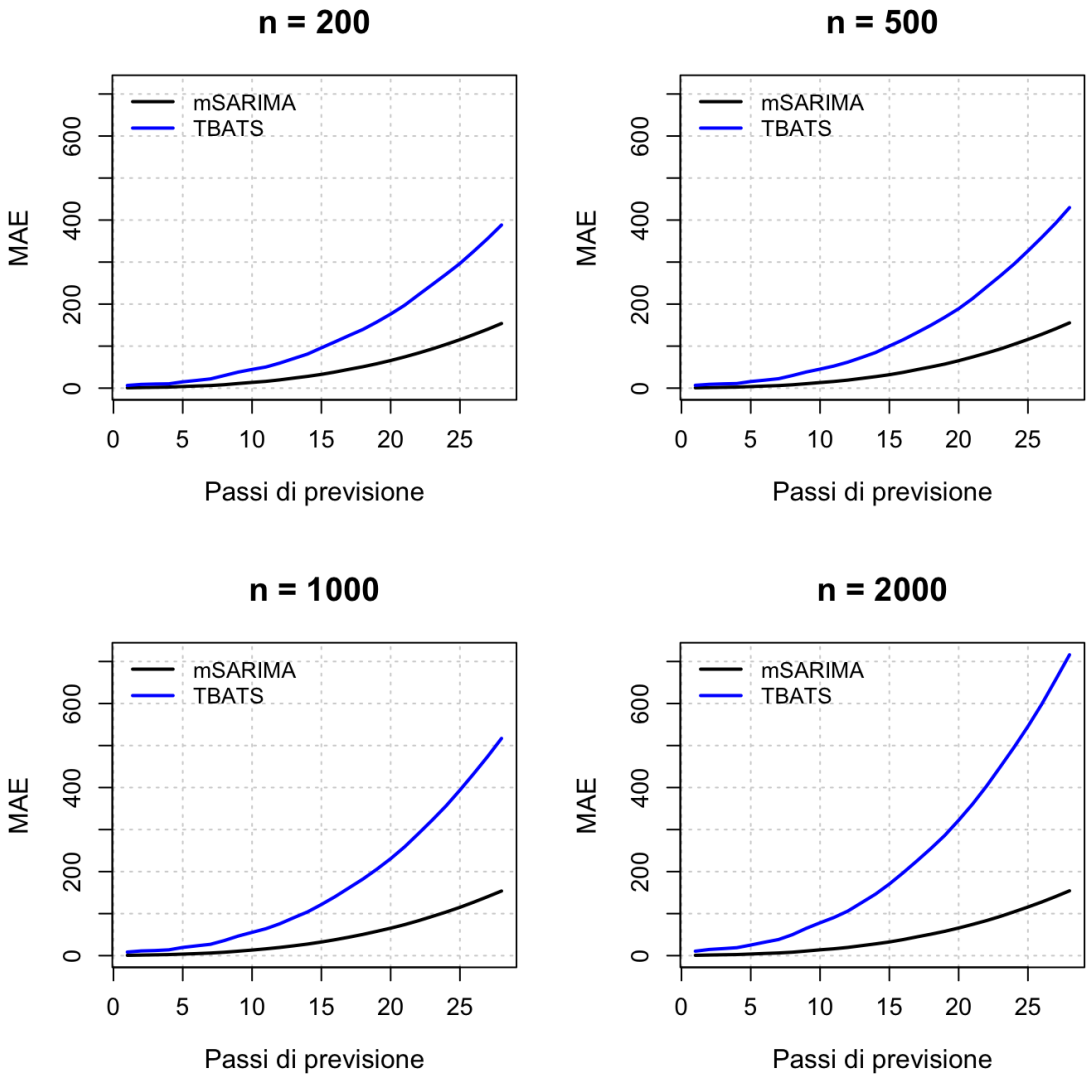


Figura 4.6: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Modello 8.

Ai primi passi di previsione i due modelli ottengono dei valori di MAE simili, ma con l'aumentare dei passi di previsione le performance di TBATS peggiorano più che proporzionalmente rispetto a quelle di mSARIMA, in particolare per serie storiche con numerosità più elevata. Le performance di mSARIMA, invece, rimangono pressoché invariate all'aumentare della numerosità della serie.

4.2.7 Valutazioni

Complessivamente il modello mSARIMA risulta adeguato a effettuare previsioni, riuscendo a performare meglio del modello TBATS, in termini di MAE, la maggior parte delle volte. Di seguito vengono analizzati i singoli aspetti di interesse, di cui si cerca di valutare l'impatto indipendentemente da tutti gli altri fattori.

La presenza di trend, indipendentemente dalla funzione che genera i dati, porta a dei peggioramenti nell'accuratezza delle previsioni, facendo registrare valori di MAE più alti rispetto a quando il trend non è presente. Ciò accade sia per modelli mSARIMA che per modelli TBATS. Questo accade perché la presenza di trend porta ad una propagazione degli errori molto pronunciata. Le previsioni successive alla prima si basano sulle previsioni precedenti, che includono già una componente di errore, e ciò avviene in maniera molto più marcata quando sono presenti differenziazioni.

La numerosità diversa delle serie storiche porta a delle conclusioni differenti a seconda del processo generatore dei dati utilizzato e dalla presenza o assenza di trend. Quando i dati erano generati dalla funzione 4.1, sia nel caso di generatore deterministico sia in quello di generatore stocastico, si nota un miglioramento delle prestazioni in termini di MAE all'aumentare della numerosità campionaria. Quanto detto vale per entrambi i modelli. Quando il generatore dei dati era esso stesso un mSARIMA, il comportamento del MAE del modello mSARIMA è generalmente simile indipendentemente dalla lunghezza della serie storica. Nello stesso contesto il comportamento del MAE del modello TBATS si differenzia: quando la serie è stazionaria il MAE rimane invariato all'aumentare di n; quando siamo in presenza di componenti integrate, il MAE assume generalmente valori più elevati all'aumentare della numerosità della serie. Questo risultato è particolarmente interessante perché ci si aspettava che all'aumentare della lunghezza della serie storica le stime dei parametri migliorassero, e con esse le prestazioni previsive.

La differenza tra processo generatore deterministico e stocastico si evince da dei risultati migliori quando il generatore era deterministico, indipendentemente dalla presenza o assenza di trend e dalla numerosità campionaria. Per queste conclusioni ci limitiamo a confrontare i risultati delle serie generate dalla funzione 4.1, in quanto quelle generate da mSARIMA presentano parametri e dinamiche completamente diverse che portano il MAE ad assumere valori difficilmente confrontabili. È ragionevole ottenere risultati migliori per serie storiche con componenti deterministiche, perché sono soggette a meno variabilità.

Capitolo 5

Le maree a Venezia

5.1 Contesto

La previsione del livello del mare a Venezia è cruciale per la vita socio-economica della città. Quello dell'acqua alta è un problema storico, aggravato da fenomeni come l'innalzamento del mare, la subsidenza e mareggiate sempre più intense. Il sistema di dighe mobili MOSE, costituito da 78 paratoie che si sollevano bloccando le bocche di porto in caso di alta marea, richiede previsioni accurate per essere attivato in tempo e prevenire inondazioni. Errori nelle previsioni comportano ingenti perdite economiche: se le paratoie vengono alzate troppo tardi ci saranno danni economici legati all'acqua alta, al contrario se le previsioni portano ad anticipare l'innalzamento delle dighe, bisogna comunque tenere conto dei problemi legati alla chiusura temporanea dei porti turistico e industriale. Inoltre sulla base delle previsioni, in caso di emergenze, è attivo un sistema di sirene di allertamento, che emette suoni di diverso tipo in base all'altezza prevista dell'acqua alta, per differenziare le diverse gravità.

In generale, il fenomeno della marea è dovuto alla somma di due componenti: la marea astronomica e il contributo meteorologico.

La marea astronomica è legata al moto dei corpi celesti, principalmente Luna e Sole, e all'attrazione gravitazionale che essi esercitano sulle masse d'acqua della Terra. Il moto di questi corpi nelle relative orbite è molto regolare, dunque la marea astronomica è caratterizzata da periodicità ricorrenti:

- La maggiore differenza tra l'alta e la bassa marea viene spiegata con il passaggio della Luna al meridiano. La Terra compie una rotazione in circa 23h56'04" (giorno siderale), ma nel frattempo anche la Luna sta lentamente effettuando una rivoluzione attorno alla Terra, quindi questi due moti si combinano

e la Luna è osservata in corrispondenza dello stesso meridiano in media ogni 24h54'33" (questa periodicità non è costante a causa dell'orbita ellittica della Luna).

Inoltre in ogni momento si verifica un'alta marea sia nella zona della Terra rivolta verso la Luna, sia nel lato opposto. Ciò porta al verificarsi di due picchi di alta marea al giorno, con periodicità pari alla metà di 24h54'33", ovvero circa ogni 12h27'17".

- La seconda componente è legata alle fasi lunari, che sono diverse a seconda della posizione della Luna nella sua orbita intorno alla Terra e contemporaneamente intorno al Sole. Durante le fasi di luna nuova e di luna piena gli effetti del Sole e della Luna si sommano, determinando le massime oscillazioni di marea (sizigie). Nei periodi di primo e ultimo quarto, invece, la marea è meno ampia e meno regolare (quadratura) e possono esserci giorni con un solo minimo e un solo massimo.

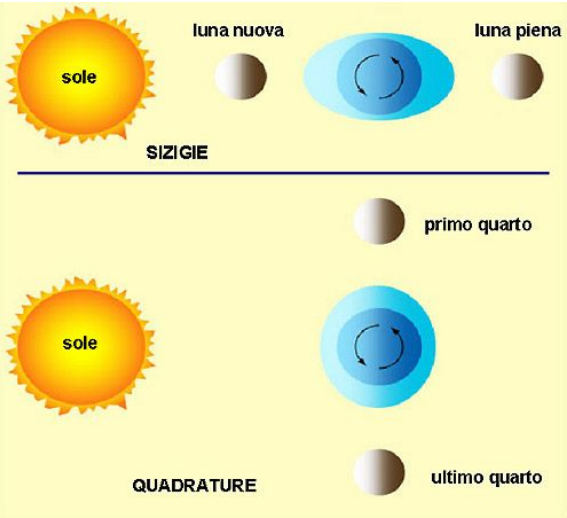


Figura 5.1: Rappresentazione fasi lunari

La Luna compie una rivoluzione intorno alla terra ogni circa 27 giorni e 7h43'12" (mese siderale), ma nel frattempo la Terra sta lentamente orbitando attorno al Sole, quindi, affinché la Luna torni ad essere osservata nella stessa posizione anche rispetto al Sole, il tempo impiegato è di 29 giorni e 12h44'3" (mese sinodico). Dunque la periodicità delle fasi lunari è descritta dalla durata del mese sinodico. Si verificheranno maree più alte sia nella fase di luna piena che nella fase di luna nuova, quindi circa ogni 14 giorni e 18h, ovvero la metà della periodicità complessiva.

- Una terza componente è legata ad una ciclicità annuale. Periodicamente varia l'inclinazione del Sole sul piano equatoriale: minore è l'inclinazione e maggiore sarà l'ampiezza della marea. Quindi, ad esempio, in prossimità degli equinozi le maree sono più alte, mentre in prossimità dei solstizi invernale ed estivo sono più basse. Nello stesso periodo varia anche la vicinanza tra Terra e Sole. Il perielio, ovvero il periodo di massima vicinanza, si verifica intorno al 3-5 gennaio e a causa dell'attrazione gravitazionale del Sole più forte si verificano maree più elevate. L'afelio, ovvero il periodo di massima lontananza, si verifica intorno al 4-7 luglio, e l'attrazione gravitazionale più debole porta a maree più basse.

- Ulteriori componenti sono l'inclinazione della Luna sul piano equatoriale, la distanza della Luna e la distanza del Sole, ecc. Ciò porta in particolare ad un quarto ciclo di circa quattro anni e mezzo.

È importante sottolineare che è riconosciuta una certa discordanza tra teoria e osservazioni in quanto, a causa della complessità del fenomeno, viene a mancare l'assunzione che in ogni istante la superficie delle acque si disponga in condizioni di equilibrio rispetto al campo delle forze perturbatrici. In particolare il fenomeno è di natura dinamica, il corpo idrico ha forze d'inerzia non trascurabili, e sono importanti le condizioni meteorologiche, le caratteristiche dei fondali e altri fattori. Ad esempio le maree hanno comportamenti molto diversi a seconda che queste vengano osservate negli oceani, nei laghi o in mari chiusi come il Mar Mediterraneo o il Mare Adriatico.

Le condizioni meteorologiche, dovute allo stato dell'atmosfera, possono essere importanti: a Venezia se si verificano condizioni di bassa pressione o forti venti di scirocco contemporaneamente ad un massimo di marea astronomica può verificarsi il fenomeno dell'acqua alta; viceversa, se c'è alta pressione contemporaneamente ad un minimo di marea astronomica possono verificarsi notevoli basse maree. Tuttavia in condizioni normali il contributo meteorologico è piccolo, e il livello osservato coincide approssimativamente con la marea astronomica.

Il livello del mare, in particolare a Venezia, è condizionato anche da altri fenomeni come l'eustatismo, ovvero l'innalzamento del livello del mare, e la subsidenza, ovvero l'abbassamento del suolo.

Attualmente, presso il Centro Previsioni e Segnalazioni Maree (CPSM), sono operativi diversi modelli numerici per le previsioni delle maree a Venezia. Questi si suddividono in modelli statistici e modelli deterministici, o idrodinamici.

Nei modelli statistici, il fenomeno viene scomposto nelle componenti di marea astronomica, calcolata con precisione anche in largo anticipo attraverso funzioni ar-

moniche, e nella componente meteorologica. Quest'ultima viene modellata tramite un numero di variabili che è diventato sempre più elevato col passare del tempo, aumentando la complessità dei modelli ma anche la loro accuratezza. Il modello 'ScontraTower 2.0' del 2021, ad esempio, include 158 regressori. Tra i più importanti troviamo la pressione atmosferica, i valori del gradiente barico, i valori del vento, le previsioni di questi fenomeni, tutti misurati in diverse stazioni collocate in determinati punti strategici.

Nei modelli deterministici, invece, vengono integrate equazioni della fluidodinamica, calcolando lo stato del mare, il livello e le correnti, sotto l'azione delle forzanti meteorologiche, tipicamente il vento e la pressione atmosferica.

5.2 Analisi

5.2.1 Il dataset

Il sito web del Centro Previsioni e Segnalazioni Maree del Comune di Venezia mette a disposizione un archivio storico [3] dei valori (orari, massimi e minimi) di livello della marea registrati a Venezia dal 1983 allo scorso anno.

Per comprendere a pieno le caratteristiche del fenomeno in esame, vengono mostrati i grafici della serie storica dei valori orari del livello dell'acqua e le funzioni di autocorrelazione, considerando diversi intervalli temporali e diversi lag. È difficile, infatti, riconoscere tutte le stagionalità contemporaneamente in maniera chiara.

A causa della complessità del fenomeno, che porta a delle difficoltà nello stabilire quali siano i cicli reali, e a causa dell'impossibilità di gestire periodicità non intere, per le analisi effettuate non si farà riferimento alla teoria ma si seguirà un approccio data-driven.

Viene innanzitutto presentato il grafico del livello dell'acqua dell'ultima settimana del 2023. In questo intervallo temporale è facile notare la periodicità giornaliera delle maree, che, come abbiamo visto solitamente succedere, presenta due picchi di alta marea e due picchi di bassa marea al giorno.

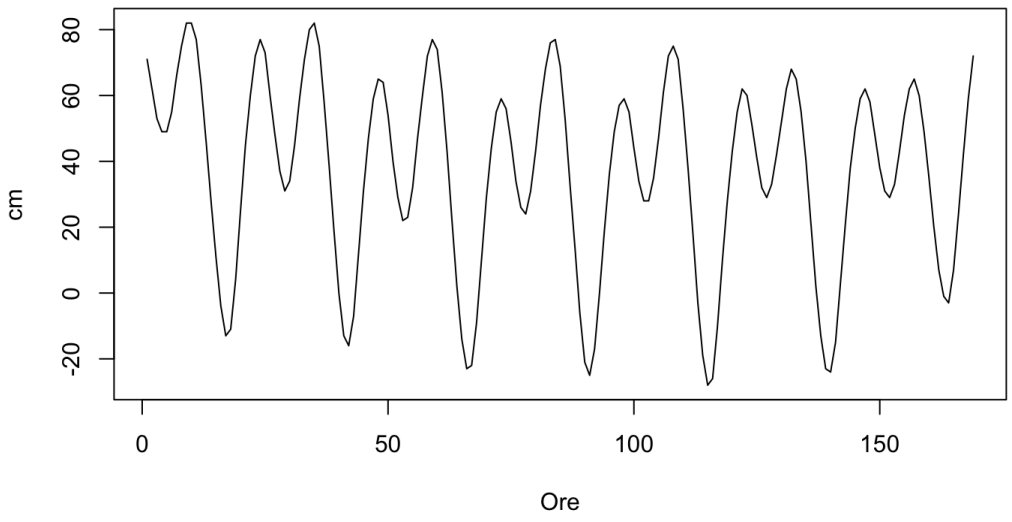


Figura 5.2: Livello dell'acqua in cm a Punta della Salute, Venezia, ultima settimana 2023

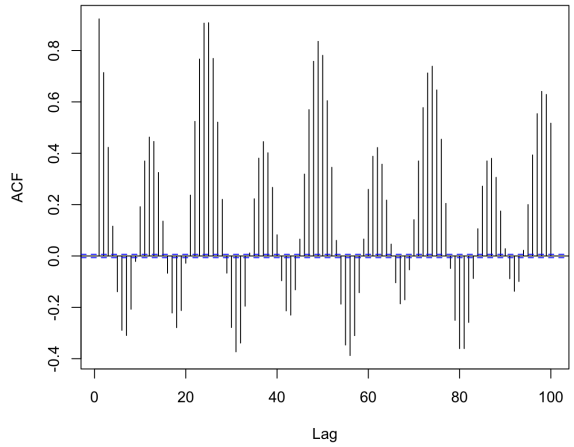


Figura 5.3: ACF considerando 100 lag

Osservando il grafico della ACF in Figura (5.3), è evidente la presenza di una stagionalità periodo 24-25, che descrive la relazione tra i livelli di marea a circa un giorno di distanza. Le correlazioni ai ritardi 12-13 sono più deboli a causa del fatto che il livello dell'acqua non sempre dipende da quello di 12-13 ore precedenti, ad esempio molto spesso i due picchi giornalieri hanno altezze diverse, e ci sono periodi in cui c'è un solo picco di marea al giorno. Per questi motivi, ai fini della rimozione della correlazione nei dati e per migliorare le prestazioni previsive è risultato molto più efficiente utilizzare il periodo intero 24-25 rispetto al semi-ciclo. Questa stagionalità è associabile al tempo percorso dalla Luna per passare in corrispondenza del medesimo meridiano, e sappiamo che in realtà non sarebbe intera, ma compresa tra 24 e 25. Dovendo scegliere un compromesso, a causa dell'impossibilità di gestire periodi non interi, è risultato più efficiente utilizzare il periodo 24.

Viene poi illustrato l'andamento del fenomeno durante l'intero anno 2023. In questa finestra temporale si riesce a notare un'altra stagionalità ricorrente: quella legata alle fasi lunari.

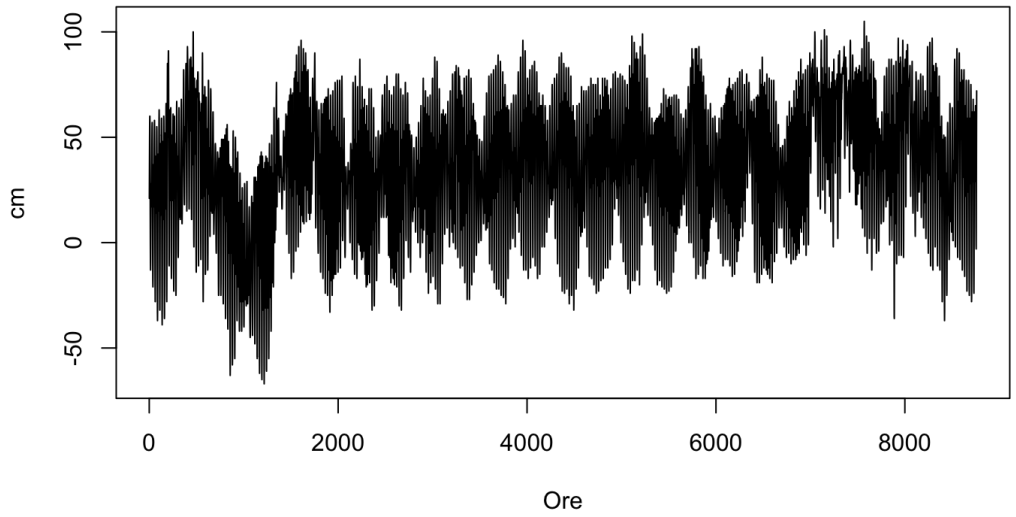


Figura 5.4: Livello dell'acqua in cm a Punta della Salute, Venezia, anno 2023

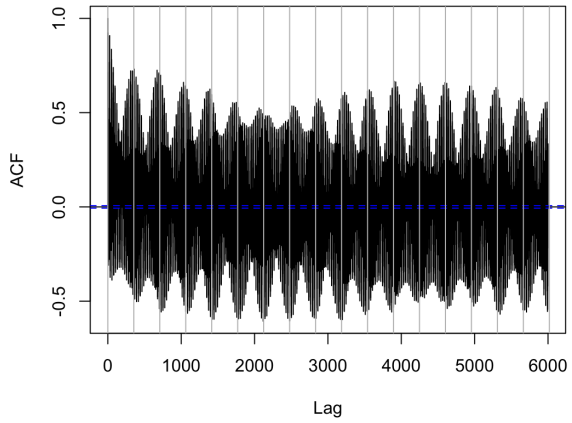


Figura 5.5: ACF considerando 6000 lag

Dal grafico della serie storica in Figura (5.4), e in particolare dal grafico della ACF quando consideriamo un numero sufficientemente elevato di lag, in questo caso 6000 (Figura (5.5)), si nota la presenza di un periodo di lunghezza 354, equivalenti a circa 14 giorni e 18 ore. Questa stagionalità è associabile alla metà della durata del mese siderale, che regola la presenza di maree più alte nelle fasi di luna nuova e luna piena, e di maree più basse nelle fasi di primo e ultimo quarto. A differenza di quanto accade quando osserviamo il grafico della stagionalità giornaliera in Figura (5.3), in questo caso i valori del livello dell'acqua sono più fortemente correlati a quelli di un semi-ciclo precedente, piuttosto che a quelli di un ciclo intero di 29.5 giorni. Tutto ciò è dovuto al fatto che i due semi-cicli sono molto simili tra loro, al contrario di quanto succede nella componente giornaliera.

In riferimento al secondo periodo, quello corretto a livello teorico sarebbe 354.37, ma, seguendo un approccio data-driven, quello che risulta spiegare in maniera più adeguata le autocorrelazioni e minimizzare il MAE nelle previsioni è 360. Si ritiene che questo avvenga a causa delle approssimazioni che si devono applicare al periodo giornaliero, oltre che all'irregolarità dei cicli e alla complessità del fenomeno.

Infine, consideriamo il grafico da un punto di vista più ampio, ovvero osservando il livello dell'acqua da gennaio 2020 a dicembre 2023. Da qui è possibile intravedere un'altra stagionalità di periodicità annuale, che, come abbiamo visto in chiave teorica, è associabile all'influenza della forza gravitazionale tra Sole e Terra.

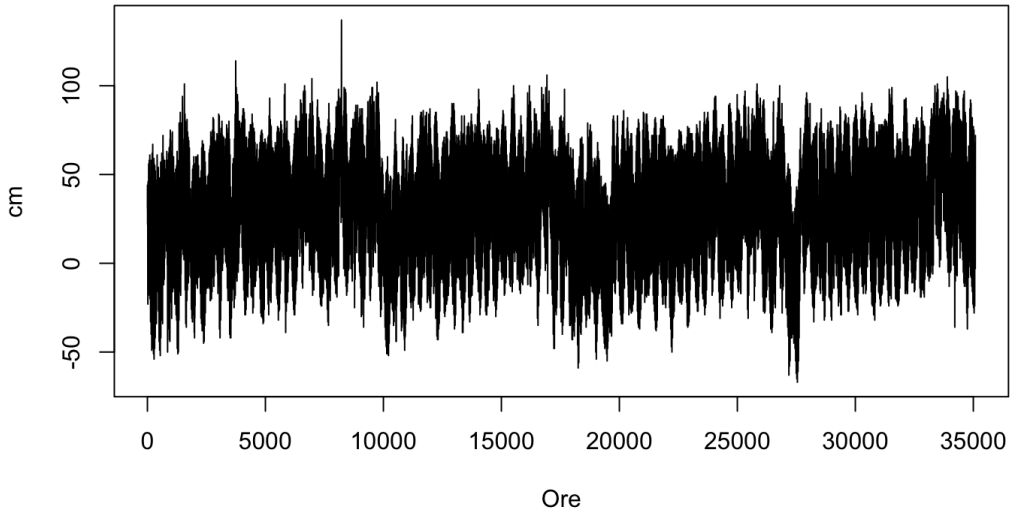


Figura 5.6: Livello dell'acqua in cm a Punta della Salute, Venezia, anni 2020-2023

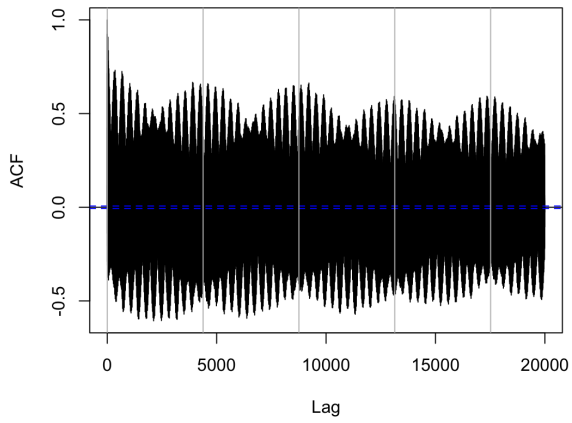


Figura 5.7: ACF considerando 20000 lag

È possibile notare questa stagionalità anche attraverso il grafico delle autocorrelazioni, se consideriamo un numero di lag ancora più elevato, in questo caso 20000

(Figura 5.7). Il periodo che emerge dalla ACF è di 4380 ore, che corrispondono alla metà di un anno solare. Significa che viene colta maggiormente l'influenza che ha il semi-ciclo rispetto che il ciclo intero, similmente a quanto accade per le fasi lunari. L'influenza della stagionalità annuale non è risultata rilevante né per la rimozione della correlazione dai dati, né per le performance previsive (dato che è utile fare previsioni con un orizzonte massimo di 5 giorni) e dunque non è stata inserita nel modello.

Vista la natura delle autocorrelazioni, che rimangono significative anche a molti lag di distanza e decadono molto lentamente, si ritiene che l'analisi dei dati con un modello a memoria lunga sarebbe più adeguato. In questo caso studio ci limitiamo ad utilizzare il modello mSARIMA, ma questa analisi rappresenta comunque un'opportunità per future ricerche ed eventuali estensioni.

5.2.2 Il modello mSARIMA in-sample

In questa fase viene utilizzata la serie storica delle rilevazioni degli ultimi due anni, 2022 e 2023. Dalle analisi preliminari non si evidenziano valori anomali. La serie risulta stazionaria in varianza, per cui non è necessario applicare alcuna trasformazione di Box-Cox. Risulta invece non stazionaria in media.

È di nostro interesse in questa fase trovare un modello della classe mSARIMA che riesca a individuare ed eliminare la correlazione presente nei dati. Tramite l'identificazione strutturale, il modello individuato è del tipo $mSARIMA(3,2,3)(1,1,1)_{24}(1,1,1)_{360}$. Dopo aver applicato il modello alla serie storica emergono i seguenti risultati.

Parametro	**Coefficiente**	**Standard Error**	**p-value**
ϕ_1	0.64639055	0.010701292	0.000000e+00
ϕ_2	0.80728191	0.014704061	0.000000e+00
ϕ_3	-0.73326850	0.009387740	0.000000e+00
θ_1	0.91961272	0.008554021	0.000000e+00
θ_2	0.95576866	0.013411923	0.000000e+00
θ_3	-0.89348714	0.007447359	0.000000e+00
Φ_1	0.23154815	0.009717019	1.666766e-125
Θ_1	0.90695933	0.003543339	0.000000e+00
Φ_2	-0.07810904	0.012327987	2.359413e-10
Θ_2	0.62568495	0.010629072	0.000000e+00

Tabella 5.1: Coefficienti, standard-error e p-value dei parametri

Dalla Tabella (5.1) si evince che tutti i parametri risultano ampiamente significativi. È stato inoltre calcolato l'indice di adattamento del modello ai dati, definito come

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Secondo questo indicatore il modello si adatta molto bene ai dati, spiegando quasi la totalità della devianza, infatti è risultato che $R^2 = 0.9917$.

Viene presentata la struttura di autocorrelazione dei residui del mese di dicembre 2023.

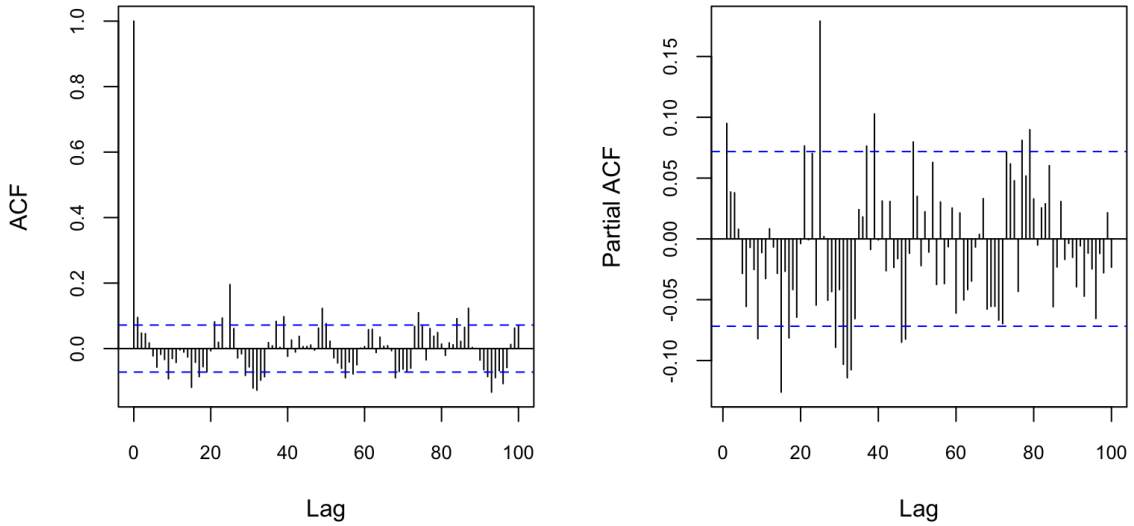


Figura 5.8: ACF e PACF considerando 100 lag

Il modello è riuscito a spiegare gran parte della correlazione presente nei dati. Quasi tutte le autocorrelazioni non risultano significative e infatti sono contenute all'interno degli intervalli di confidenza definiti nel grafico con la linea blu. Molte delle autocorrelazioni significative non rappresentano un problema perché non corrispondono ai ritardi stagionali. Nei dati reali ci sono molti fattori che possono causare deviazioni rispetto al modello teorico perfetto, per cui è normale che rimanga qualche autocorrelazione significativa. Si può notare però la presenza di diverse autocorrelazioni significative attorno al lag 24 (escluso), in particolare al 25° ritardo, e ciò è dovuto al fatto che il vero periodo che influenza il fenomeno delle maree non è esattamente 24. Utilizzando la periodicità 25 nel modello, l'autocorrelazione al lag 24 sarebbe più elevata di quanto non sia quella al lag 25 utilizzando il periodo 24, per tale motivo si è optato per l'utilizzo di questo modello.

5.2.3 Le previsioni di marea

Per utilizzare un modello mSARIMA a scopi previsivi in questo ambito, si è deciso di fornirsi di un modello diverso rispetto a quello individuato nella fase di identificazione. Sebbene il modello scelto nella sezione precedente riuscisse a rimuovere l'autocorrelazione dei dati in maniera soddisfacente, non era altrettanto adeguato a effettuare previsioni, in particolare quelle a medio/lungo termine. Infatti, essendo i dati collezionati ogni ora, il valore che assume il livello dell'acqua in un istante y_t è piuttosto simile a quello assunto all'istante precedente y_{t-1}, e per questo risultavano molto significativi i parametri delle componenti non stagionali. Inserire molti parametri non stagionali porta sicuramente a dei buoni risultati se si vuole spiegare la serie storica, ma a risultati del tutto insoddisfacenti se si vogliono effettuare delle previsioni: il modello sarà abbastanza adeguato per prevedere il livello dell'acqua nei primissimi passi di previsione, ma non sarà mai realmente utile in una previsione di periodo più lungo, perché non riesce a cogliere le componenti stagionali, dando troppa importanza a quelle non stagionali.

Si è notato che l'introduzione di componenti integrate facesse divergere le previsioni a valori non realistici, e che anche per quanto riguarda le componenti a media mobile queste portavano a risultati peggiori in termini di MAE.

Il modello scelto è del tipo $mSARIMA(1,0,0)(2,0,0)_{24}(2,0,0)_{360}$.

Le prestazioni sono state confrontate con quelle del modello TBATS, i cui parametri sono stati stimati tramite procedure automatiche. Data la sua capacità di gestire periodi non interi, si è confrontato sia il modello con gli stessi periodi utilizzati dal modello mSARIMA (indicato con 'TBATS 1'), sia quello con i periodi reali, pari a $S_1 = 24.91$ e $S_2 = 354.37$ (indicato con 'TBATS 2').

I parametri dei due modelli sono stati stimati attraverso l'utilizzo dei dati relativi alle misurazioni del livello dell'acqua a Punta Salute da gennaio 2017 a dicembre 2019. Successivamente, per ogni ora dal 1° gennaio 2020 al 26 dicembre 2023, sono state effettuate previsioni a 120 passi in avanti (120 ore di anticipo di previsione, equivalenti a 5 giorni). Si è misurato l'errore medio assoluto, MAE, e si sono confrontate le prestazioni dei due modelli.

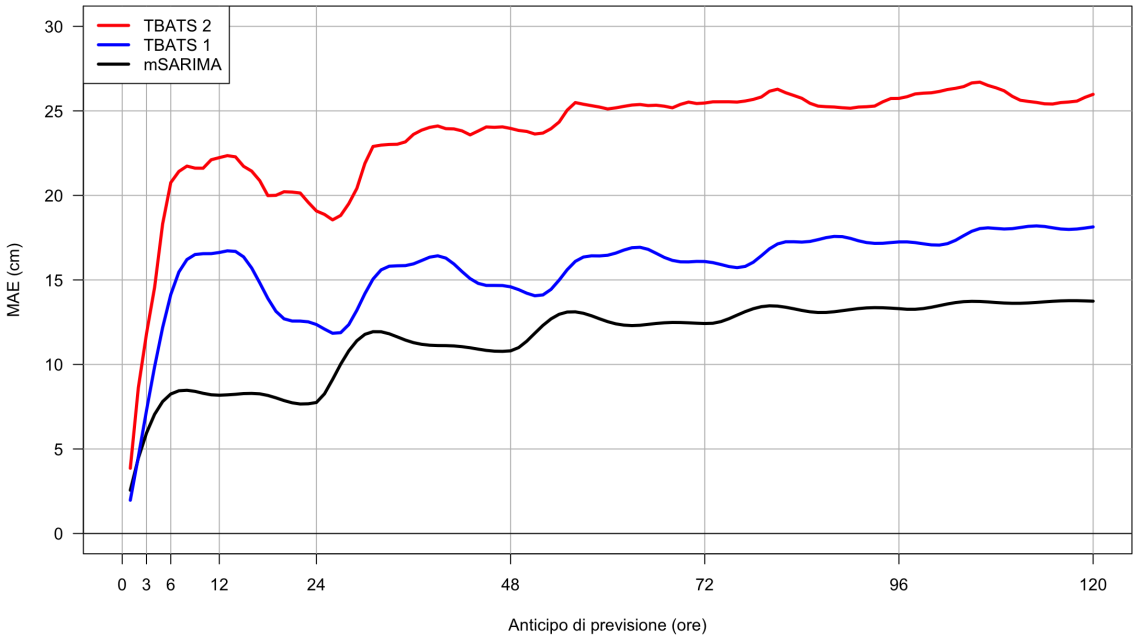


Figura 5.9: Confronto tra MAE (in cm) di mSARIMA e TBATS per diversi orizzonti di previsione, in operatività da gennaio 2020 a dicembre 2023

Orizzonte (ore)	3	6	12	24	48	72	96	120
MAE mSARIMA	5.94	8.25	8.18	7.75	10.80	12.42	13.30	13.74
MAE TBATS 1	7.27	14.13	16.62	12.36	14.59	16.09	17.25	18.14
MAE TBATS 2	11.81	20.76	22.23	19.08	23.96	25.46	25.74	25.98
RMSE mSARIMA	7.87	10.87	10.86	10.51	14.16	16.10	17.15	17.70
RMSE TBATS 1	9.34	17.42	20.12	15.51	18.17	19.94	21.32	22.42
RMSE TBATS 2	15.64	26.32	27.39	24.54	30.15	31.86	32.14	32.49

Tabella 5.2: Confronto tra MAE (in cm) e RMSE di mSARIMA e TBATS per diversi orizzonti di previsione, in operatività da gennaio 2020 a dicembre 2023

Dal grafico in Figura (5.9) è evidente come mSARIMA performi meglio rispetto a TBATS 1. Quest'ultimo, già a 12 ore di anticipo di previsione raggiunge un errore medio assoluto maggiore di 16cm, che lo rende piuttosto inadeguato per un utilizzo reale di previsione; all'aumentare dei passi di previsione la situazione peggiora arrivando a errori in media più grandi di 18cm a 5 giorni di anticipo. Il modello mSARIMA, invece, mantiene gli errori in media sotto gli 8.5cm per le previsioni fino ad un giorno di anticipo, e tra i 10cm e i 14cm per le previsioni fino a 5 giorni. Si nota come ogni 24 ore ci sia un graduale peggioramento delle performance, dovuto al fatto che le previsioni successive utilizzano, invece dei valori osservati di marea, le

previsioni passate, che inevitabilmente saranno composte anche da una componente di errore. Il modello TBATS 2, con i periodi reali, ha ottenuto i risultati peggiori.

Per scopi conoscitivi è stato effettuato un confronto del MAE anche con il Modello Esteso, un modello statistico di riferimento per il CPSM che è stato operativo dal 1996 al 2008. Esso utilizza oltre ai valori osservati in diverse stazioni, anche i valori previsti di pressione atmosferica, calcolati dal centro europeo ECMWF e diffusi dal Centro Nazionale di Meteorologia e Climatologia Aeronautica (CNMCA) dell'Aeronautica Militare Italiana. Nell'articolo [11] in cui viene introdotto il Modello Esteso BIG SUMDP (ancora più avanzato), viene mostrato un confronto tra le sue prestazioni e quelle del Modello Esteso, misurate facendo previsioni fino a 5 giorni di anticipo dal 1° ottobre 2008 al 31 marzo 2009. Il grafico in Figura (5.10) riporta, oltre al MAE, l'*Indice di accuratezza*, definito come errore medio ± due volte la deviazione standard, che equivale quindi ad un intervallo di confidenza del 95.44%.

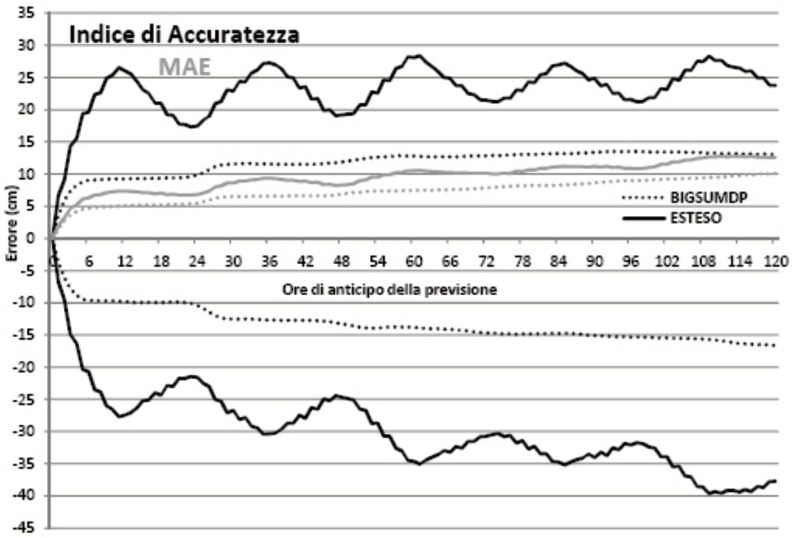


Figura 5.10: MAE Modello Esteso (linea grigia continua) in operatività da ottobre 2008 a marzo 2009, a diversi passi di previsione

Per il medesimo periodo si sono create le previsioni anche con il modello mSARIMA, addestrato con le misurazioni del livello dell'acqua dal 1° gennaio 2005 al 31 dicembre 2007. Le performance ottenute sono le seguenti:

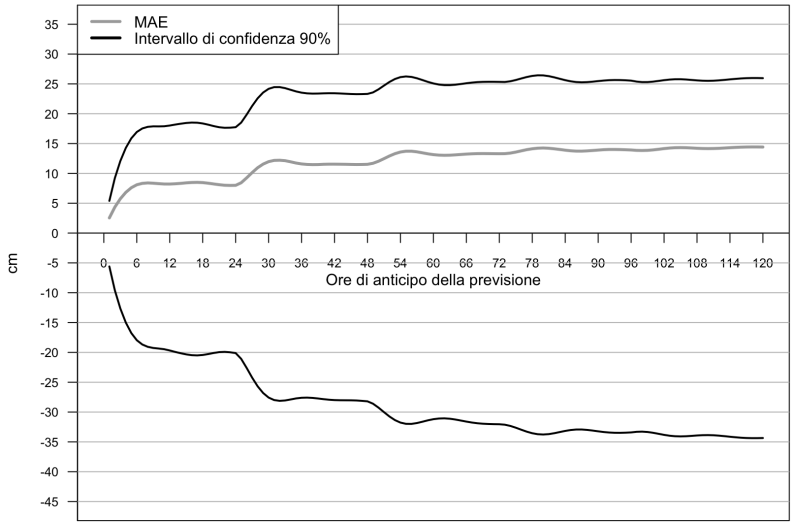


Figura 5.11: MAE Modello mSARIMA (linea grigia continua) in operatività da ottobre 2008 a marzo 2009, a diversi passi di previsione

Emerge che con mSARIMA riusciamo ad ottenere un MAE generalmente più alto di 2cm o 3cm rispetto al Modello Esteso, a fronte però di un modello molto più semplice, che considera esclusivamente i valori della serie storica osservata senza nessun regressore esterno. Chiaramente aggiungendo al modello variabili capaci di spiegare le componenti meteorologiche le prestazioni aumenterebbero. Inoltre si è notato che un intervallo di confidenza del 90% calcolato con mSARIMA è simile, in termini di scala, ad un intervallo con una confidenza del 95% calcolato con il Modello Esteso.

Capitolo 6

Conclusioni

È stata proposta un'estensione della classe dei modelli SARIMA per consentire la gestione di più cicli stagionali. Si sono discussi gli opportuni accorgimenti da implementare nella procedura di Box-Jenkis per individuare il modello mSARIMA corretto.

Si è effettuato uno studio di simulazione per verificare le proprietà asintotiche del modello. Gli stimatori dei parametri del modello risultano non distorti e consistenti, infatti l'errore di stima diminuisce all'aumentare della numerosità della serie storica e con esso anche la sua varianza. Sulla base di un controllo diagnostico sui residui effettuato attraverso il test di Ljung-Box ai ritardi stagionali, è evidente come il modello riesca a cogliere e rimuovere in maniera adeguata l'autocorrelazione presente nei dati ai ritardi stagionali. Talvolta risulta adeguato modificare i gradi di libertà del test di Ljung-Box per tenere conto del numero di volte che nel calcolo del test vengono considerate le stesse autocorrelazioni, dato che con più stagionalità, se i periodi sono multipli tra loro, questo è uno scenario possibile. Il modello risulta quindi adatto per la gestione di fenomeni con stagionalità multiciclo.

Si è effettuato un confronto delle prestazioni previsive, misurate tramite MAE, con uno dei modelli più affermati in letteratura, il TBATS. Si riscontra che generalmente mSARIMA performa meglio di TBATS, in particolare per numerosità campionarie elevate. Quando la serie storica è stazionaria il modello riesce ad ottenere delle previsioni buone e con un'accuratezza piuttosto costante anche all'aumentare dei passi di previsione; quando la serie storica presenta una componente di trend, il MAE cresce di più all'aumentare dei passi di previsione, e le prestazioni migliorano molto quando la serie ha numerosità più elevata.

Si è applicato il modello al fenomeno delle maree a Venezia, caratterizzato dalla presenza di molte stagionalità che influenzano il livello dell'acqua. È stato indivi-

duato un modello del tipo $mSARIMA(3,2,3)(1,1,1)_{24}(1,1,1)_{360}$, che riesce a cogliere la presenza di stagionalità e i cui residui risultano incorrelati. Alcune limitazioni sono dovute all'impossibilità del modello mSARIMA di gestire stagionalità complesse, non intere, che invece caratterizzano le maree. Si è inoltre studiato un modello adeguato per le previsioni, che è del tipo $mSARIMA(1,0,0)(2,0,0)_{24}(2,0,0)_{360}$. Il modello utilizza solamente i valori osservati del livello dell'acqua alla stazione di Punta della Salute, a Venezia, e ottiene performance previsive migliori del modello TBATS, con un errore medio assoluto che è inferiore agli 8.5cm per le previsioni entro le 24 ore, e fino a meno di 14cm per previsioni a 5 giorni di anticipo. Le prestazioni di mSARIMA portano ad errori generalmente di soli 2-3cm più elevati rispetto al modello statistico Esteso, utilizzato dal CPSM dal 1996 al 2008, con la differenza che mSARIMA utilizza solamente i valori osservati del livello dell'acqua, mentre il modello Esteso gestiva in maniera precisa la marea astronomica ed era composto da molteplici regressori per modellare le componenti atmosferiche.

Bibliografia e sitografia

[1] Aruã, Souza, *23 years of hourly eletric energy demand (Brazil)*, Disponibile online, URL: https://www.kaggle.com/datasets/arusouza/23-years-of-hourly-eletric-energy-demand-brazil, 2023.

[2] Maryam Shoaei, *US traffic data with weather and calendar dataset*, Disponibile online, URL: https://www.kaggle.com/datasets/maryamshoaei/us-traffic-data-with-weather-and-calendar-dataset, 2024.

[3] Comune di Venezia, *Archivio storico: livello di marea a Venezia*, Disponibile online, URL: https://www.comune.venezia.it/it/content/archivio-storico-livello-marea-venezia-1, 2024.

[4] A. M. De Livera, R. J. Hyndman e R. D. Snyder, «Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing,» *Journal of the American Statistical Association*, 2011.

[5] G. E. P. Box e G. Jenkis, *Time Series Analysis, Forecasting and Control*. Holden-Day, 1976.

[6] G. E. P. Box e D. Cox, *An analysis of transformations*. Journal of the Royal Statistical Society, 1964, pp. 211–243.

[7] G. M. Ljung e G. E. P. Box, *On a measure of lack of fit in time series models*. Biometrika, 1978, pp. 297–303.

[8] C. M. Jarque e A. K. Bera, *Efficient tests for normality homoskedasticity and serial indipendence of regression residuals*. Economic Letters, 1980, pp. 255–259.

[9] J. D. Hamilton, *Time series analysis*. Princeton University Press, Princeton, NJ, 1994, p. 73.

[10] K. Bandara, R. J. Hyndman e C. Bergmeir, «MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns,» 2021.

[11] A. Tosoni e P. Canestrelli, *Il modello stocastico per la previsione di marea a Venezia*. 2010.

Appendice A

Studio di simulazione

A.0.1 Modello 2: mSARIMA(1,0,0)(0,0,1)$_{S_1}$(0,0,1)$_{S_2}$

Valori dei parametri $\phi = 0.5, \Theta_1 = 0.5, \Theta_2 = 0.5, \sigma^2_\varepsilon = 1$

I. **Stagionalità**: $S_1 = 4$, $S_2 = 7$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.0026	0.0001	0.0009	0.0003	0.0040	0.0017	0.0008	0.0004
Θ_1	0.0039	0.0001	0.0009	-0.0002	0.0042	0.0016	0.0008	0.0004
Θ_2	0.0089	0.0033	0.0015	0.0010	0.0040	0.0016	0.0008	0.0004
σ^2_ε	-0.0709	-0.0257	-0.0131	-0.0071	0.0117	0.0043	0.0020	0.0010

Tabella A.1: Distorsione e varianza degli stimatori, Modello 2, $S_1 = 4, S_2 = 7$

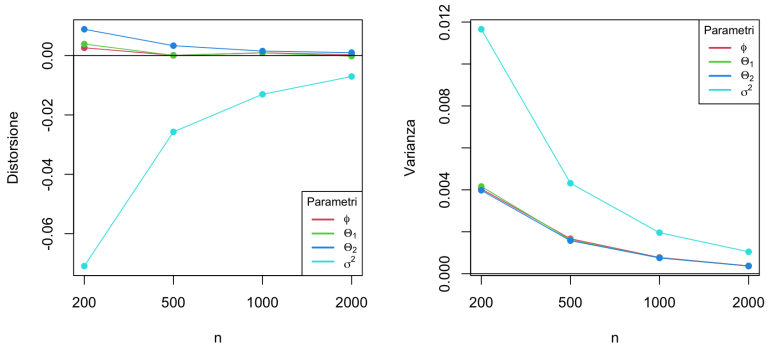


Figura A.1: Distorsione e varianza degli stimatori, Modello 2, $S_1 = 4, S_2 = 7$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.50	94.85	94.25	94.70	94.50	94.85	94.25	94.70
$h_1 = 3, h_2 = 3$	94.95	95.45	94.60	95.55	94.95	95.45	94.60	95.55
$h_1 = 4, h_2 = 4$	95.15	94.80	94.50	95.20	95.15	94.80	94.50	95.20
$h_1 = 5, h_2 = 5$	95.75	95.00	94.20	94.95	95.75	95.00	94.20	94.95

Tabella A.2: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 2, $S_1 = 4, S_2 = 7$

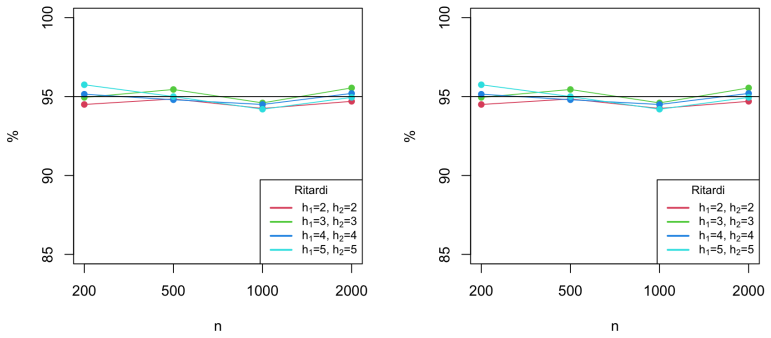


Figura A.2: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 2, $S_1 = 4, S_2 = 7$

II. **Stagionalità**: $S_1 = 4$, $S_2 = 11$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.0063	0.0017	0.0018	0.0002	0.0040	0.0016	0.0007	0.0004
Θ_1	0.0026	0.0017	0.0012	0.0009	0.0044	0.0016	0.0008	0.0004
Θ_2	0.0137	0.0034	0.0024	0.0018	0.0044	0.0016	0.0008	0.0004
σ^2_ε	-0.0986	-0.0378	-0.0198	-0.0089	0.0130	0.0046	0.0021	0.0010

Tabella A.3: Distorsione e varianza degli stimatori, Modello 2, $S_1 = 4, S_2 = 11$

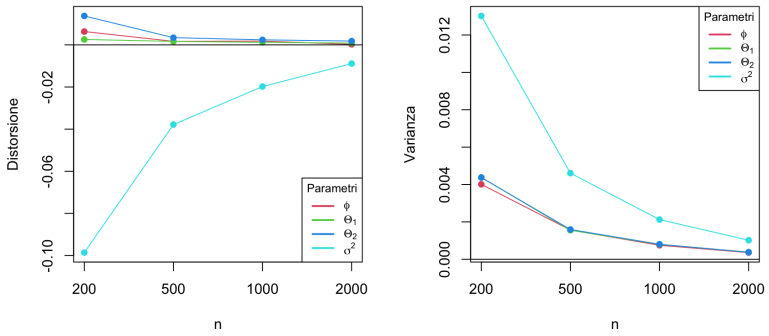


Figura A.3: Distorsione e varianza degli stimatori, Modello 2, $S_1 = 4, S_2 = 11$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	95.00	94.45	94.65	94.25	95.00	94.45	94.65	94.25
$h_1 = 3, h_2 = 3$	95.75	95.10	95.20	95.25	95.75	95.10	95.20	95.25
$h_1 = 4, h_2 = 4$	95.95	95.55	95.25	95.45	95.95	95.55	95.25	95.45
$h_1 = 5, h_2 = 5$	96.30	95.65	95.40	95.75	96.30	95.65	95.40	95.75

Tabella A.4: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 2, $S_1 = 4, S_2 = 11$

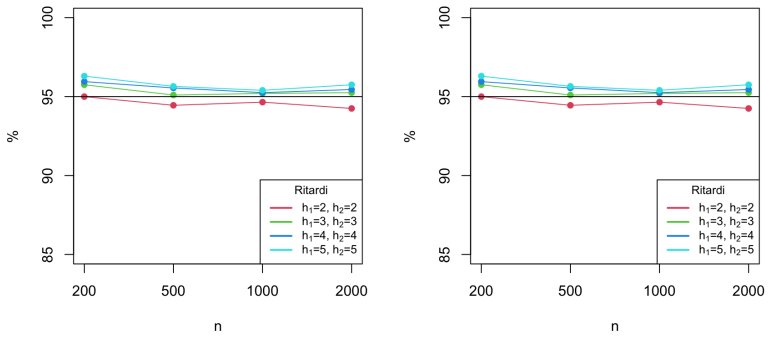


Figura A.4: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 2, $S_1 = 4, S_2 = 11$

III. **Stagionalità**: $S_1 = 4$, $S_2 = 12$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.0090	0.0047	0.0021	0.0014	0.0039	0.0016	0.0008	0.0004
Θ_1	0.0164	0.0055	0.0022	0.0018	0.0044	0.0016	0.0008	0.0004
Θ_2	0.0197	0.0063	0.0035	0.0018	0.0043	0.0017	0.0008	0.0004
σ^2_ε	-0.1160	-0.0426	-0.0208	-0.0118	0.0134	0.0047	0.0022	0.0011

Tabella A.5: Distorsione e varianza degli stimatori, Modello 2, $S_1 = 4, S_2 = 12$

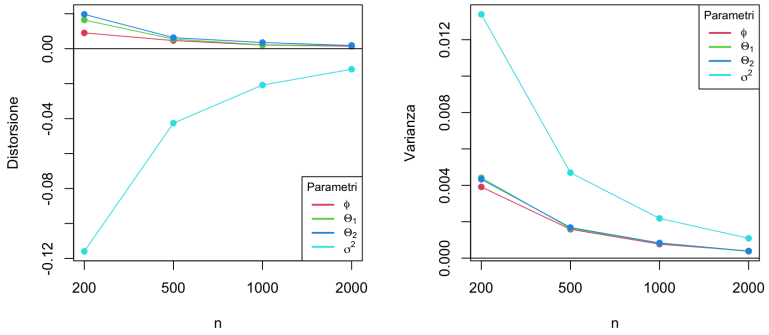


Figura A.5: Distorsione e varianza degli stimatori, Modello 2, $S_1 = 4, S_2 = 12$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	95.05	94.50	94.50	94.65	95.05	94.50	94.50	94.65
$h_1 = 3, h_2 = 3$	97.20	96.65	96.35	96.50	94.45	93.20	93.30	92.85
$h_1 = 4, h_2 = 4$	96.70	96.50	95.75	96.20	94.35	93.75	93.15	93.35
$h_1 = 5, h_2 = 5$	97.25	95.75	96.10	96.35	95.25	93.15	93.80	93.85

Tabella A.6: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 2, $S_1 = 4, S_2 = 12$

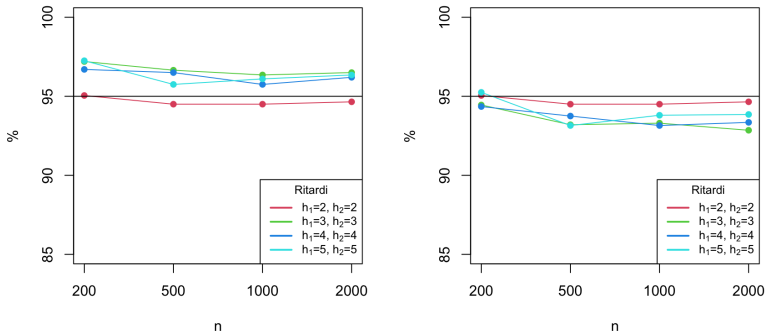


Figura A.6: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 2, $S_1 = 4, S_2 = 12$

A.0.2 Modello 3: mSARIMA(0,0,0)(1,0,1)$_{S_1}$(1,0,1)$_{S_2}$

Valori dei parametri $\Phi_1 = 0.4, \Phi_2 = 0.5, \Theta_1 = 0.5, \Theta_2 = 0.4, \sigma^2_\varepsilon = 1$

I. **Stagionalità**: $S_1 = 4$, $S_2 = 7$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
Φ_1	0.1142	0.0948	0.0349	0.0123	0.2299	0.1546	0.0780	0.0322
Φ_2	0.1169	0.0832	0.0413	0.0149	0.2365	0.1524	0.0667	0.0339
Θ_1	0.1030	0.0915	0.0329	0.0110	0.2304	0.1536	0.0726	0.0297
Θ_2	0.1190	0.0831	0.0426	0.0145	0.2551	0.1597	0.0719	0.0373
σ^2_ε	-0.0359	-0.0156	-0.0078	-0.0042	0.0109	0.0042	0.0019	0.0010

Tabella A.7: Distorsione e varianza degli stimatori, Modello 3, $S_1 = 4, S_2 = 7$

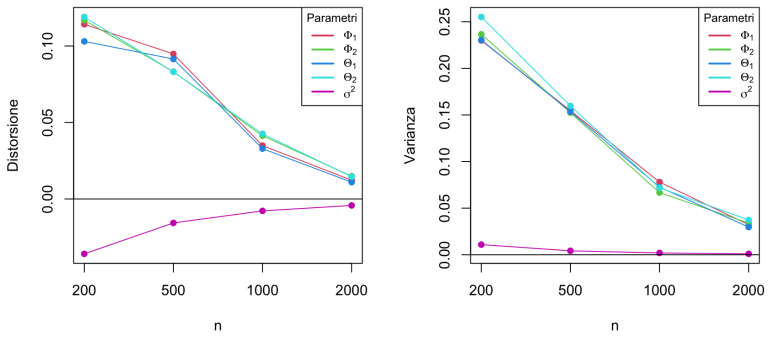


Figura A.7: Distorsione e varianza degli stimatori, Modello 3, $S_1 = 4, S_2 = 7$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$h_1 = 3, h_2 = 3$	90.90	93.75	92.55	92.05	90.90	93.75	92.55	92.05
$h_1 = 4, h_2 = 4$	92.50	94.40	94.40	94.55	92.50	94.40	94.40	94.55
$h_1 = 5, h_2 = 5$	93.15	95.15	94.80	95.25	93.15	95.15	94.80	95.25

Tabella A.8: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 3, $S_1 = 4, S_2 = 7$

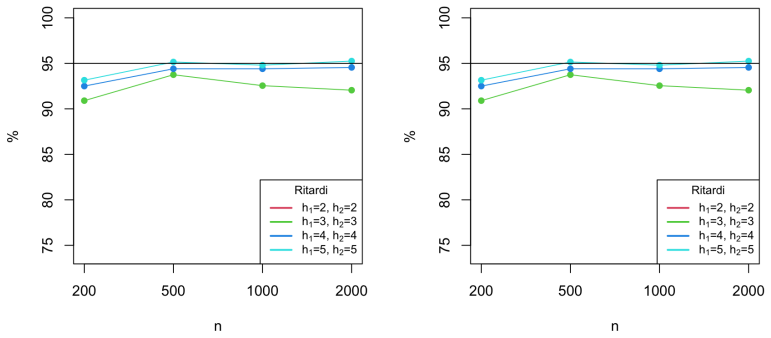


Figura A.8: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 3, $S_1 = 4, S_2 = 7$

II. **Stagionalità**: $S_1 = 4$, $S_2 = 11$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
Φ_1	0.1213	0.0959	0.0342	0.0186	0.2284	0.1548	0.0759	0.0350
Φ_2	0.0993	0.0775	0.0430	0.0198	0.2809	0.1366	0.0702	0.0315
Θ_1	0.1096	0.0932	0.0324	0.0173	0.2318	0.1542	0.0706	0.0326
Θ_2	0.0933	0.0789	0.0444	0.0200	0.3027	0.1440	0.0751	0.0345
σ^2_ε	-0.0563	-0.0215	-0.0115	-0.0057	0.0110	0.0041	0.0020	0.0010

Tabella A.9: Distorsione e varianza degli stimatori, Modello 3, $S_1 = 4, S_2 = 11$

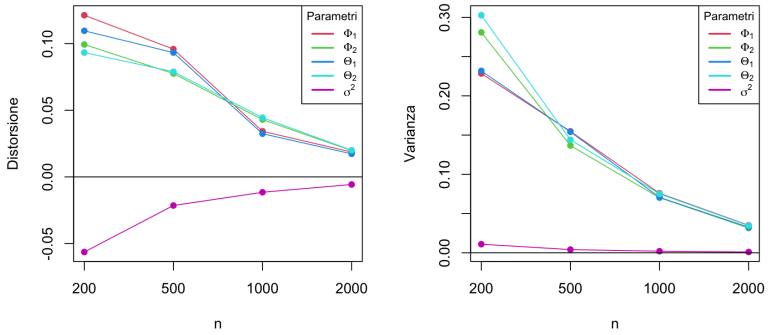


Figura A.9: Distorsione e varianza degli stimatori, Modello 3, $S_1 = 4, S_2 = 11$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$h_1 = 3, h_2 = 3$	89.85	91.70	92.55	93.05	89.85	91.70	92.55	93.05
$h_1 = 4, h_2 = 4$	91.60	93.90	94.50	94.20	91.60	93.90	94.50	94.20
$h_1 = 5, h_2 = 5$	92.50	94.05	94.70	94.25	92.50	94.05	94.70	94.25

Tabella A.10: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 3, $S_1 = 4, S_2 = 11$

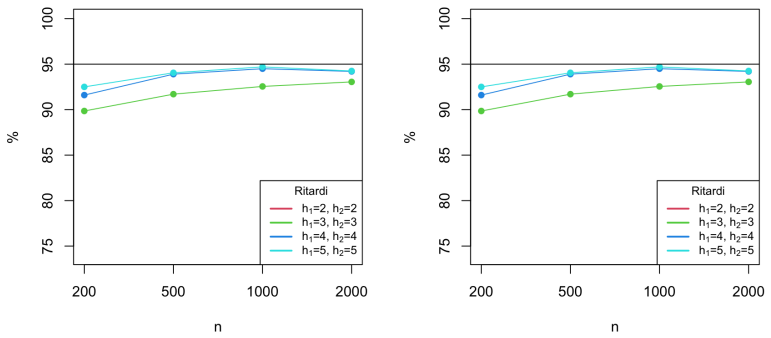


Figura A.10: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 3, $S_1 = 4, S_2 = 11$

III. **Stagionalità**: $S_1 = 4$, $S_2 = 12$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
Φ_1	0.1193	0.0937	0.0415	0.0128	0.2649	0.1553	0.0749	0.0377
Φ_2	0.0713	0.0640	0.0174	0.0114	0.2720	0.1512	0.0776	0.0317
Θ_1	0.1090	0.0905	0.0395	0.0110	0.2665	0.1553	0.0706	0.0350
Θ_2	0.0754	0.0675	0.0185	0.0134	0.2901	0.1559	0.0831	0.0345
σ^2_ε	-0.0566	-0.0244	-0.0108	-0.0074	0.0120	0.0041	0.0021	0.0011

Tabella A.11: Distorsione e varianza degli stimatori, Modello 3, $S_1 = 4, S_2 = 12$

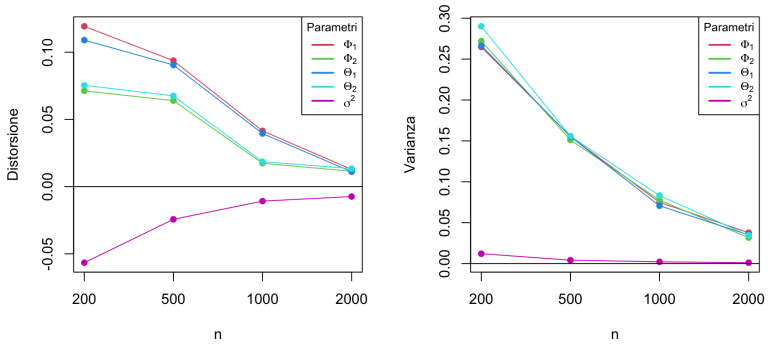


Figura A.11: Distorsione e varianza degli stimatori, Modello 3, $S_1 = 4, S_2 = 12$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$h_1 = 3, h_2 = 3$	93.40	95.85	96.55	97.30	85.00	85.95	88.55	89.25
$h_1 = 4, h_2 = 4$	94.15	96.05	97.10	97.30	89.65	91.35	93.85	94.15
$h_1 = 5, h_2 = 5$	93.45	96.25	96.70	97.25	91.05	93.45	94.55	94.85

Tabella A.12: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 3, $S_1 = 4, S_2 = 12$

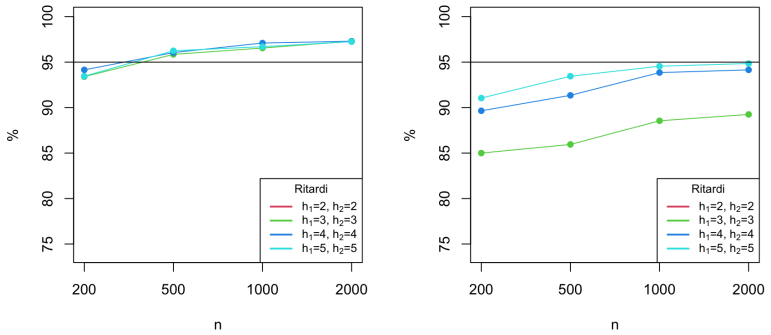


Figura A.12: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 3, $S_1 = 4, S_2 = 12$

A.0.3 Modello 4: mSARIMA(1,0,1)(0,1,0)$_{S_1}$(0,1,0)$_{S_2}$

Valori dei parametri $\phi = 0.5, \theta = 0.6, \sigma^2_\varepsilon = 1$

I. **Stagionalità**: $S_1 = 4$, $S_2 = 7$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.2978	0.1335	0.0973	0.0273	0.2537	0.1636	0.1276	0.0432
θ	0.2827	0.1251	0.0927	0.0242	0.2837	0.1620	0.1288	0.0388
σ^2_ε	0.0212	0.0053	0.0020	0.0013	0.0164	0.0069	0.0035	0.0017

Tabella A.13: Distorsione e varianza degli stimatori, Modello 4, $S_1 = 4, S_2 = 7$

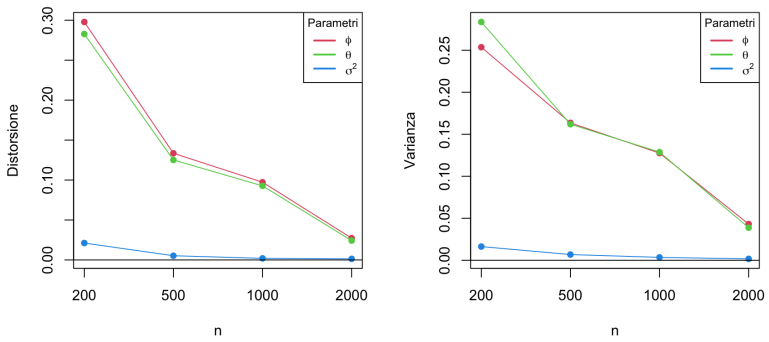


Figura A.13: Distorsione e varianza degli stimatori, Modello 4, $S_1 = 4, S_2 = 7$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.90	96.15	94.65	96.00	94.90	96.15	94.65	96.00
$h_1 = 3, h_2 = 3$	95.15	95.50	95.25	96.40	95.15	95.50	95.25	96.40
$h_1 = 4, h_2 = 4$	94.75	95.30	94.90	96.45	94.75	95.30	94.90	96.45
$h_1 = 5, h_2 = 5$	95.00	95.70	95.45	96.40	95.00	95.70	95.45	96.40

Tabella A.14: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 4, $S_1 = 4, S_2 = 7$

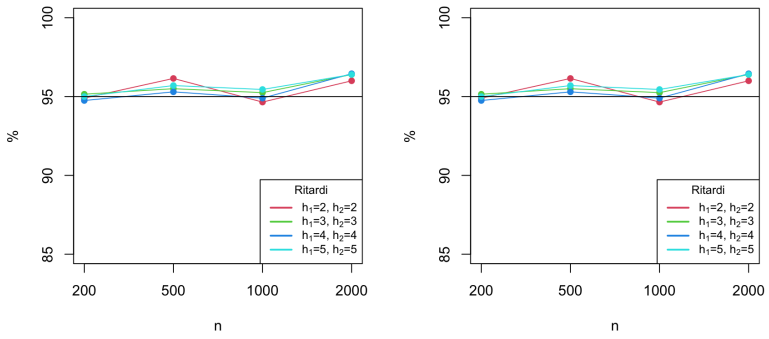


Figura A.14: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 4, $S_1 = 4, S_2 = 7$

II. **Stagionalità**: $S_1 = 4$, $S_2 = 11$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.2880	0.1474	0.1017	0.0303	0.2556	0.1732	0.1328	0.0438
θ	0.2715	0.1391	0.0989	0.0275	0.2829	0.1735	0.1346	0.0400
σ_ε^2	0.0221	0.0091	0.0020	0.0026	0.0179	0.0072	0.0034	0.0017

Tabella A.15: Distorsione e varianza degli stimatori, Modello 4, $S_1 = 4, S_2 = 11$

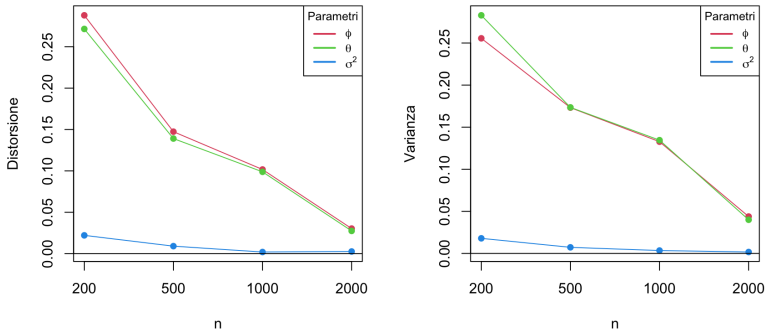


Figura A.15: Distorsione e varianza degli stimatori, Modello 4, $S_1 = 4, S_2 = 11$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.70	95.95	95.15	97.15	94.70	95.95	95.15	97.15
$h_1 = 3, h_2 = 3$	94.75	95.70	95.60	95.90	94.75	95.70	95.60	95.90
$h_1 = 4, h_2 = 4$	94.70	95.15	95.05	95.85	94.70	95.15	95.05	95.85
$h_1 = 5, h_2 = 5$	93.90	95.15	95.45	95.80	93.90	95.15	95.45	95.80

Tabella A.16: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 4, $S_1 = 4, S_2 = 11$

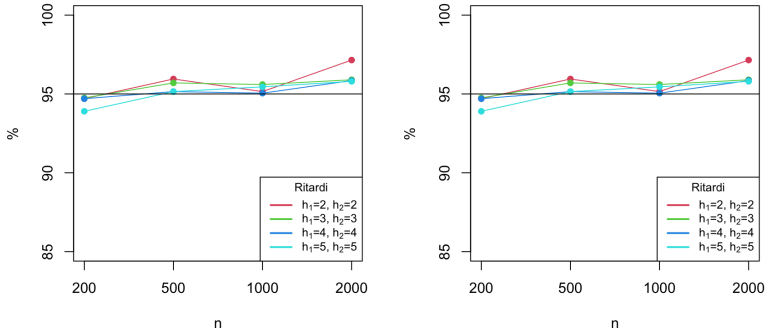


Figura A.16: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 4, $S_1 = 4, S_2 = 11$

III. **Stagionalità**: $S_1 = 4$, $S_2 = 12$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
ϕ	0.2903	0.1375	0.1106	0.0319	0.2459	0.1676	0.1238	0.0454
θ	0.2756	0.1287	0.1059	0.0284	0.2710	0.1683	0.1251	0.0411
σ_ε^2	0.0227	0.0083	0.0021	0.0017	0.0180	0.0069	0.0034	0.0017

Tabella A.17: Distorsione e varianza degli stimatori, Modello 4, $S_1 = 4, S_2 = 12$

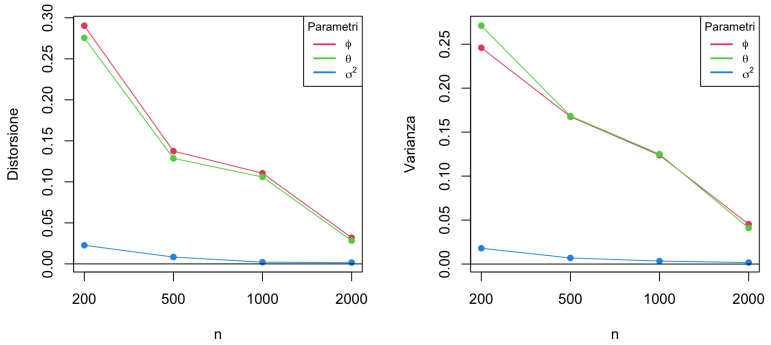


Figura A.17: Distorsione e varianza degli stimatori, Modello 4, $S_1 = 4, S_2 = 12$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.95	95.85	95.85	96.10	94.95	95.85	95.85	96.10
$h_1 = 3, h_2 = 3$	92.10	94.20	94.35	93.70	88.75	91.50	90.55	90.35
$h_1 = 4, h_2 = 4$	92.65	94.10	94.50	94.05	89.70	92.00	91.30	90.95
$h_1 = 5, h_2 = 5$	92.35	94.30	93.75	94.65	89.45	92.60	91.65	92.00

Tabella A.18: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 4, $S_1 = 4, S_2 = 12$

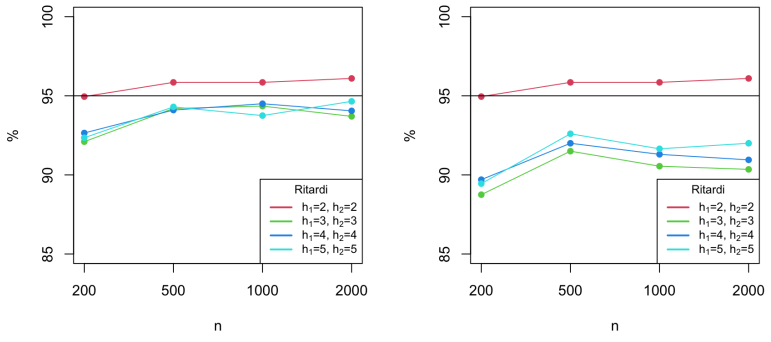


Figura A.18: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 4, $S_1 = 4, S_2 = 12$

A.0.4 Modello 5: mSARIMA(0,0,0)(1,1,0)$_{S_1}$(1,1,0)$_{S_2}$

Valori dei parametri $\Phi_1 = 0.5, \Phi_2 = 0.5, \sigma^2_\varepsilon = 1$

I. **Stagionalità**: $S_1 = 4$, $S_2 = 7$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
Φ_1	0.0070	0.0038	0.0023	0.0002	0.0068	0.0026	0.0013	0.0007
Φ_2	0.0042	0.0040	0.0003	0.0003	0.0069	0.0026	0.0013	0.0006
σ^2_ε	0.0170	0.0090	0.0059	0.0012	0.0174	0.0068	0.0033	0.0017

Tabella A.19: Distorsione e varianza degli stimatori, Modello 5, $S_1 = 4, S_2 = 7$

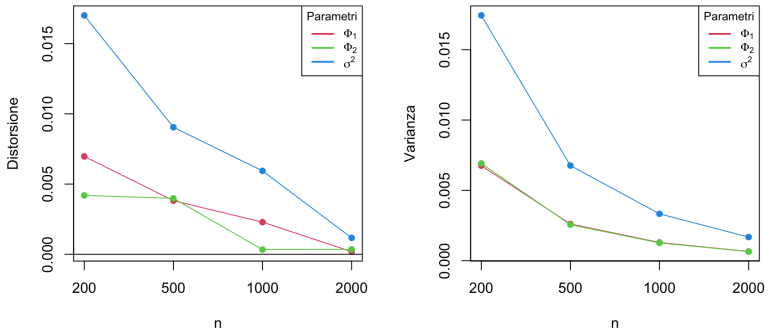


Figura A.19: Distorsione e varianza degli stimatori, Modello 5, $S_1 = 4, S_2 = 7$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	93.95	94.50	94.55	94.40	93.95	94.50	94.55	94.40
$h_1 = 3, h_2 = 3$	94.05	94.60	95.30	94.90	94.05	94.60	95.30	94.90
$h_1 = 4, h_2 = 4$	94.30	95.05	94.65	95.20	94.30	95.05	94.65	95.20
$h_1 = 5, h_2 = 5$	94.35	94.55	95.35	95.15	94.35	94.55	95.35	95.15

Tabella A.20: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 5, $S_1 = 4, S_2 = 7$

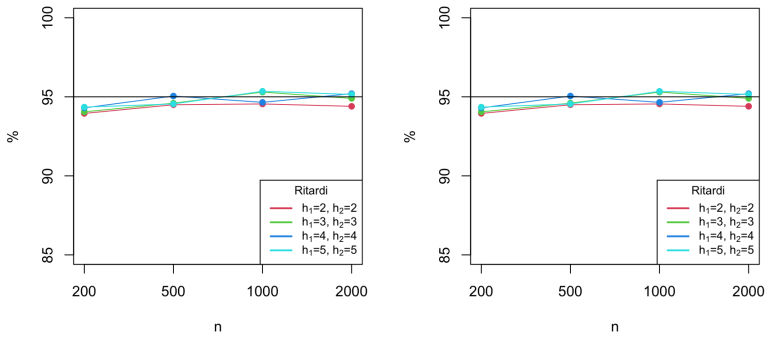


Figura A.20: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 5, $S_1 = 4, S_2 = 7$

II. **Stagionalità**: $S_1 = 4$, $S_2 = 11$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
Φ_1	0.0046	0.0011	0.0014	0.0010	0.0068	0.0027	0.0012	0.0007
Φ_2	0.0079	0.0045	0.0019	0.0008	0.0070	0.0026	0.0012	0.0006
σ^2_ε	0.0139	0.0085	0.0039	0.0013	0.0185	0.0067	0.0033	0.0018

Tabella A.21: Distorsione e varianza degli stimatori, Modello 5, $S_1 = 4, S_2 = 11$

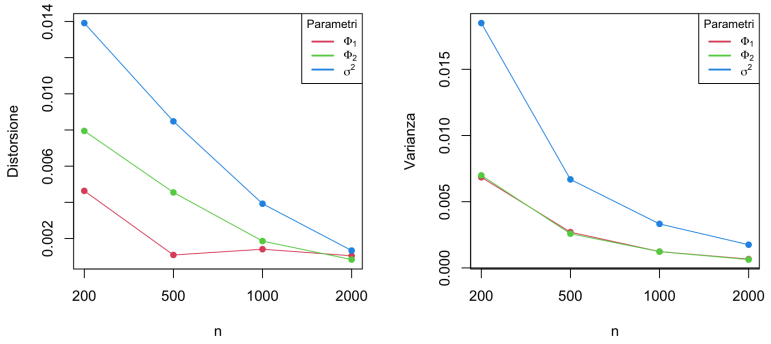


Figura A.21: Distorsione e varianza degli stimatori, Modello 5, $S_1 = 4, S_2 = 11$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.70	95.05	94.75	94.00	94.70	95.05	94.75	94.00
$h_1 = 3, h_2 = 3$	94.70	94.25	95.05	93.80	94.70	94.25	95.05	93.80
$h_1 = 4, h_2 = 4$	95.30	94.50	95.40	94.15	95.30	94.50	95.40	94.15
$h_1 = 5, h_2 = 5$	95.05	94.30	94.75	94.45	95.05	94.30	94.75	94.45

Tabella A.22: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 5, $S_1 = 4, S_2 = 11$

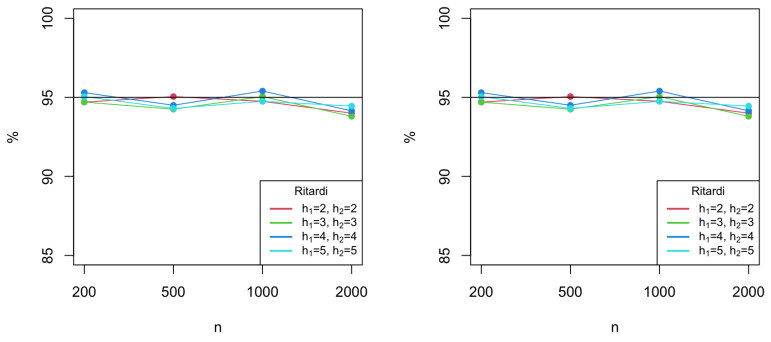


Figura A.22: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 5, $S_1 = 4, S_2 = 11$

III. **Stagionalità**: $S_1 = 4$, $S_2 = 12$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
Φ_1	0.0037	0.0027	0.0002	-0.0007	0.0071	0.0027	0.0014	0.0006
Φ_2	0.0155	0.0093	0.0037	0.0021	0.0076	0.0028	0.0013	0.0006
σ_ε^2	0.0174	0.0053	0.0037	0.0018	0.0177	0.0068	0.0035	0.0017

Tabella A.23: Distorsione e varianza degli stimatori, Modello 5, $S_1 = 4, S_2 = 12$

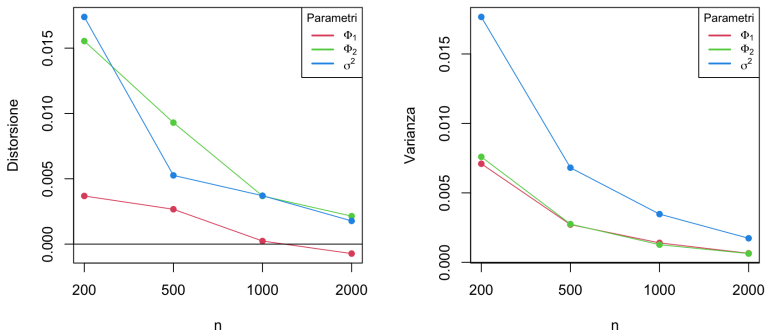


Figura A.23: Distorsione e varianza degli stimatori, Modello 5, $S_1 = 4, S_2 = 12$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	95.00	94.90	94.00	94.45	95.00	94.90	94.00	94.45
$h_1 = 3, h_2 = 3$	96.50	95.65	96.50	96.50	93.20	92.85	93.05	92.85
$h_1 = 4, h_2 = 4$	96.30	95.35	96.40	96.30	93.70	93.15	93.40	93.10
$h_1 = 5, h_2 = 5$	95.55	94.85	96.75	96.20	94.00	92.70	94.20	93.95

Tabella A.24: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 5, $S_1 = 4, S_2 = 12$

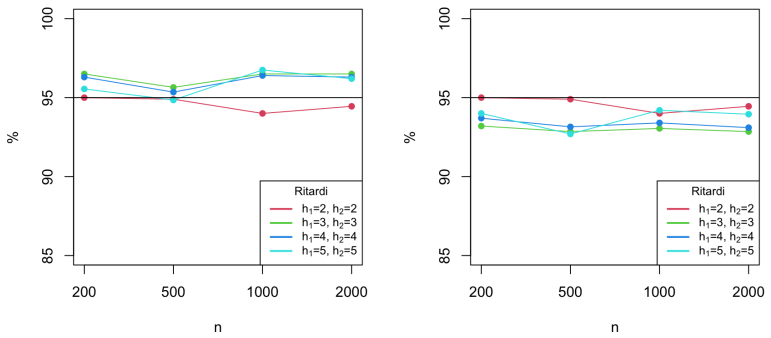


Figura A.24: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 5, $S_1 = 4, S_2 = 12$

A.0.5 Modello 6: mSARIMA(0,0,0)(0,1,1)$_{S_1}$(0,1,1)$_{S_2}$

Valori dei parametri $\Theta_1 = 0.5, \Theta_2 = 0.5, \sigma^2_\varepsilon = 1$

I. **Stagionalità**: $S_1 = 4$, $S_2 = 7$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
Θ_1	-0.0056	-0.0001	-0.0020	-0.0002	0.0077	0.0026	0.0013	0.0006
Θ_2	-0.0026	-0.0018	-0.0006	-0.0003	0.0072	0.0027	0.0013	0.0006
σ^2_ε	0.0169	0.0096	0.0047	0.0017	0.0176	0.0066	0.0032	0.0017

Tabella A.25: Distorsione e varianza degli stimatori, Modello 6, $S_1 = 4, S_2 = 7$

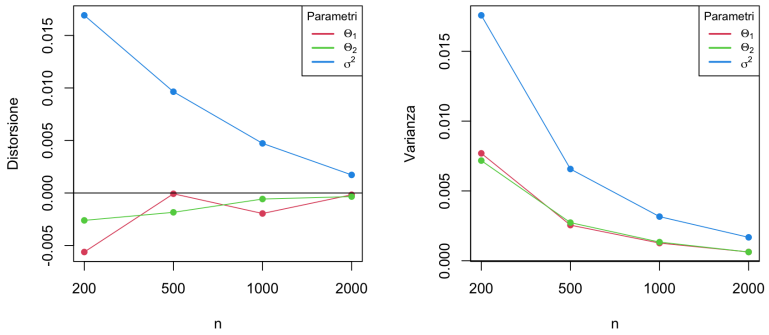


Figura A.25: Distorsione e varianza degli stimatori, Modello 6, $S_1 = 4, S_2 = 7$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	93.30	94.20	94.55	94.90	93.30	94.20	94.55	94.90
$h_1 = 3, h_2 = 3$	93.50	94.90	94.50	94.85	93.50	94.90	94.50	94.85
$h_1 = 4, h_2 = 4$	94.40	94.95	94.70	94.90	94.40	94.95	94.70	94.90
$h_1 = 5, h_2 = 5$	94.40	94.95	95.40	95.00	94.40	94.95	95.40	95.00

Tabella A.26: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 6, $S_1 = 4, S_2 = 7$

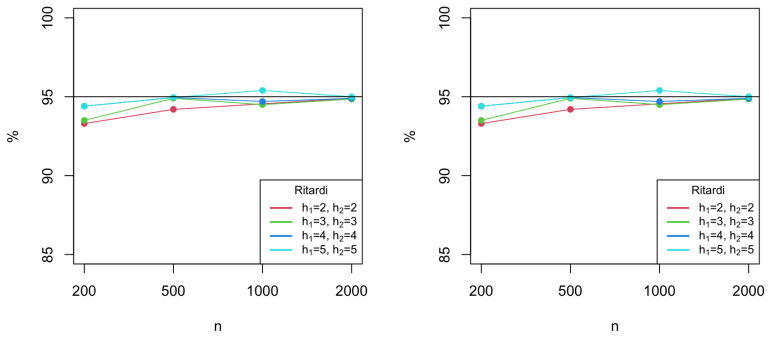


Figura A.26: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 6, $S_1 = 4, S_2 = 7$

II. **Stagionalità**: $S_1 = 4$, $S_2 = 11$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
Θ_1	-0.0056	-0.0017	-0.0005	-0.0002	0.0074	0.0028	0.0012	0.0007
Θ_2	-0.0038	-0.0035	-0.0010	-0.0013	0.0074	0.0027	0.0013	0.0006
σ^2_ε	0.0164	0.0090	0.0017	0.0008	0.0181	0.0063	0.0034	0.0017

Tabella A.27: Distorsione e varianza degli stimatori, Modello 6, $S_1 = 4, S_2 = 11$

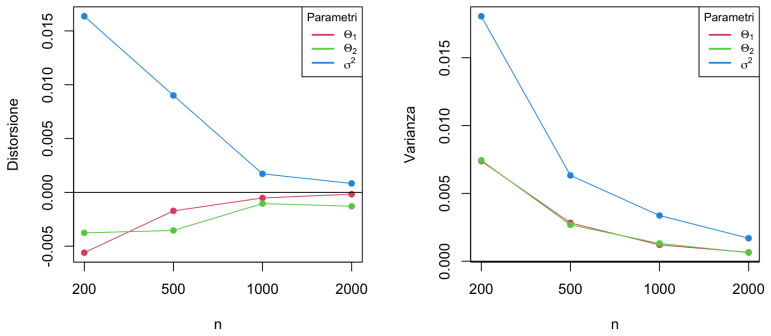


Figura A.27: Distorsione e varianza degli stimatori, Modello 6, $S_1 = 4, S_2 = 11$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	94.25	95.20	94.80	94.15	94.25	95.20	94.80	94.15
$h_1 = 3, h_2 = 3$	94.55	95.15	95.35	94.80	94.55	95.15	95.35	94.80
$h_1 = 4, h_2 = 4$	95.25	94.70	94.40	94.35	95.25	94.70	94.40	94.35
$h_1 = 5, h_2 = 5$	94.80	94.95	95.25	94.95	94.80	94.95	95.25	94.95

Tabella A.28: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 6, $S_1 = 4, S_2 = 11$

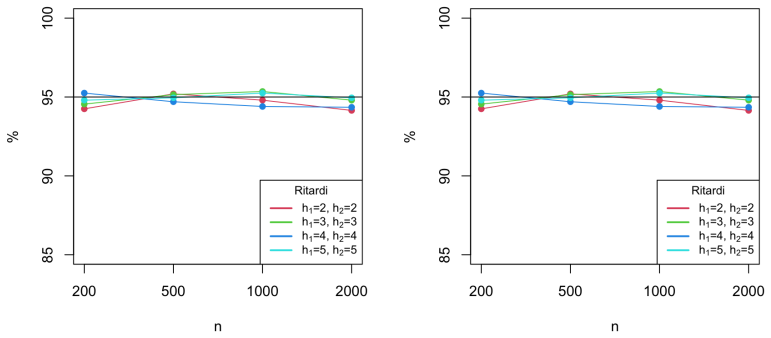


Figura A.28: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 6, $S_1 = 4, S_2 = 11$

III. **Stagionalità**: $S_1 = 4$, $S_2 = 12$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
Θ_1	-0.0036	-0.0015	0.0001	-0.0004	0.0078	0.0025	0.0014	0.0007
Θ_2	-0.0034	-0.0002	-0.0020	0.0007	0.0078	0.0028	0.0014	0.0007
σ^2_ε	0.0226	0.0061	0.0035	0.0015	0.0177	0.0069	0.0034	0.0017

Tabella A.29: Distorsione e varianza degli stimatori, Modello 6, $S_1 = 4, S_2 = 12$

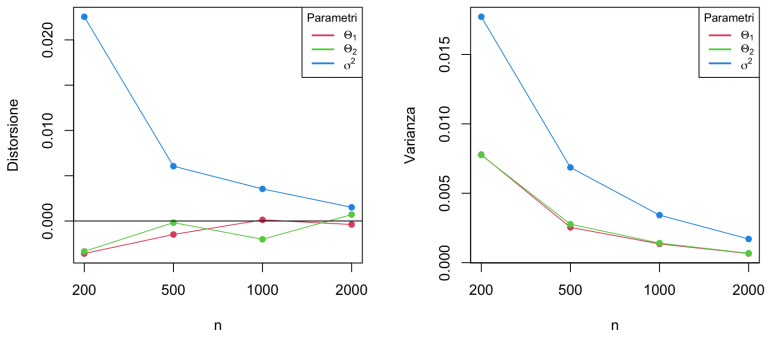


Figura A.29: Distorsione e varianza degli stimatori, Modello 6, $S_1 = 4, S_2 = 12$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	74.50	94.75	94.15	93.80	74.50	94.75	94.15	93.80
$h_1 = 3, h_2 = 3$	78.60	97.00	96.20	96.00	72.65	93.45	93.00	92.10
$h_1 = 4, h_2 = 4$	79.35	96.50	95.75	95.70	75.90	93.75	93.50	92.65
$h_1 = 5, h_2 = 5$	79.70	95.65	96.05	96.35	76.65	93.50	93.50	93.20

Tabella A.30: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 6, $S_1 = 4, S_2 = 12$

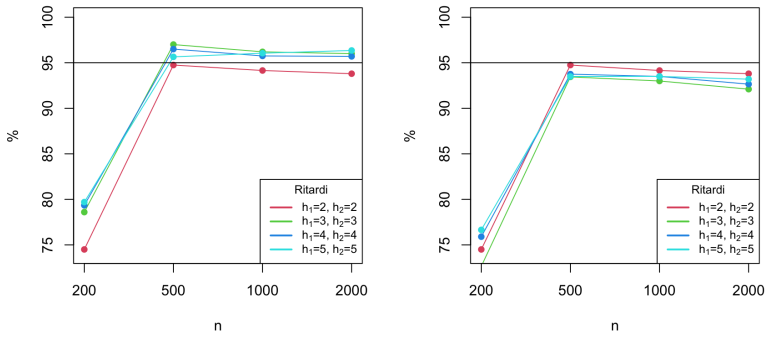


Figura A.30: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 6, $S_1 = 4, S_2 = 12$

A.0.6 Modello 7: mSARIMA$(0,1,1)(0,1,1)_{S_1}(0,1,1)_{S_2}$

Valori dei parametri $\theta = 0.5, \Theta_1 = 0.5, \Theta_2 = 0.5, \sigma^2_\varepsilon = 1$

I. **Stagionalità**: $S_1 = 4$, $S_2 = 7$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
θ	-0.0049	0.0001	0.0006	0.0009	0.0077	0.0026	0.0013	0.0007
Θ_1	-0.0032	-0.0014	-0.0004	-0.0007	0.0080	0.0026	0.0013	0.0006
Θ_2	-0.0038	-0.0019	-0.0007	-0.0017	0.0078	0.0027	0.0013	0.0006
σ^2_ε	0.0289	0.0103	0.0051	0.0027	0.0166	0.0067	0.0034	0.0017

Tabella A.31: Distorsione e varianza degli stimatori, Modello 7, $S_1 = 4, S_2 = 7$

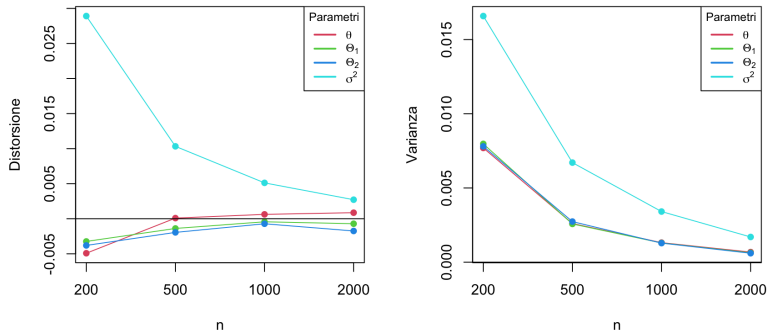


Figura A.31: Distorsione e varianza degli stimatori, Modello 7, $S_1 = 4, S_2 = 7$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	90.90	94.15	94.65	94.75	90.90	94.15	94.65	94.75
$h_1 = 3, h_2 = 3$	91.45	94.95	95.25	95.75	91.45	94.95	95.25	95.75
$h_1 = 4, h_2 = 4$	92.60	94.45	95.05	95.80	92.60	94.45	95.05	95.80
$h_1 = 5, h_2 = 5$	92.75	94.05	94.50	95.45	92.75	94.05	94.50	95.45

Tabella A.32: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 7, $S_1 = 4, S_2 = 7$

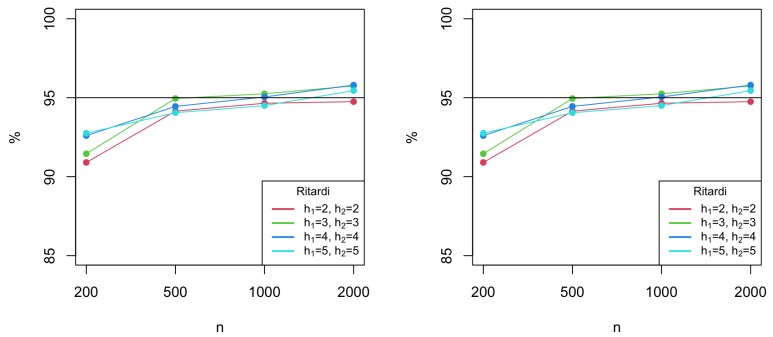


Figura A.32: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 7, $S_1 = 4, S_2 = 7$

II. **Stagionalità**: $S_1 = 4$, $S_2 = 11$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
θ	-0.0081	-0.0012	-0.0008	-0.0003	0.0078	0.0027	0.0013	0.0006
Θ_1	-0.0042	-0.0048	-0.0010	0.0001	0.0070	0.0028	0.0013	0.0006
Θ_2	0.0002	-0.0016	-0.0004	-0.0001	0.0074	0.0027	0.0013	0.0006
σ^2_ε	0.0276	0.0091	0.0054	0.0025	0.0179	0.0067	0.0033	0.0016

Tabella A.33: Distorsione e varianza degli stimatori, Modello 7, $S_1 = 4, S_2 = 11$

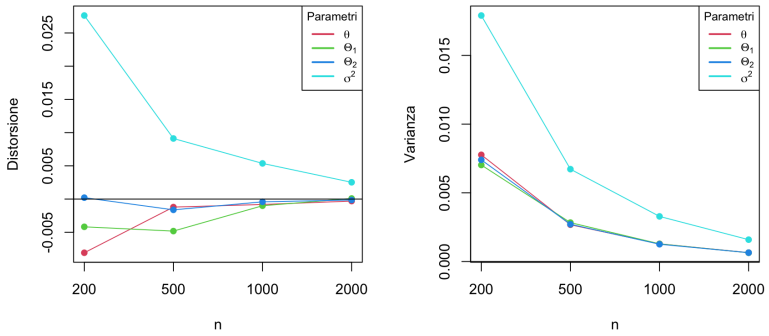


Figura A.33: Distorsione e varianza degli stimatori, Modello 7, $S_1 = 4, S_2 = 11$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	79.60	94.70	94.05	94.95	79.60	94.70	94.05	94.95
$h_1 = 3, h_2 = 3$	81.05	95.75	95.05	95.15	81.05	95.75	95.05	95.15
$h_1 = 4, h_2 = 4$	82.50	95.60	94.90	95.10	82.50	95.60	94.90	95.10
$h_1 = 5, h_2 = 5$	82.85	95.55	94.65	94.60	82.85	95.55	94.65	94.60

Tabella A.34: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 7, $S_1 = 4, S_2 = 11$

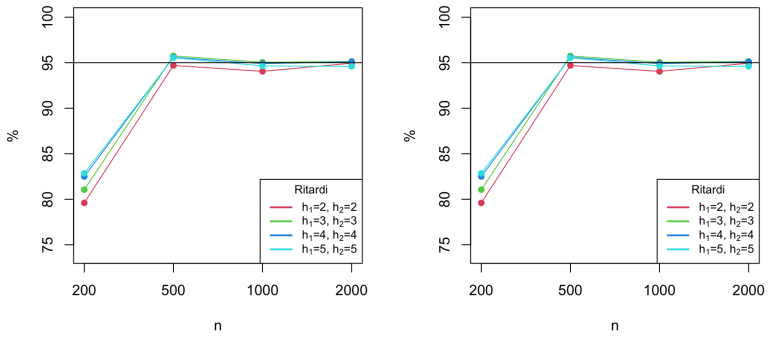


Figura A.34: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 7, $S_1 = 4, S_2 = 11$

III. **Stagionalità**: $S_1 = 4$, $S_2 = 12$

	Distorsione				Varianza			
Parametri	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
θ	-0.0037	-0.0005	-0.0010	0.0002	0.0079	0.0027	0.0014	0.0007
Θ_1	-0.0053	-0.0011	-0.0005	-0.0003	0.0075	0.0029	0.0014	0.0007
Θ_2	-0.0017	-0.0000	-0.0013	-0.0009	0.0079	0.0029	0.0013	0.0007
σ^2_ε	0.0221	0.0085	0.0038	0.0016	0.0171	0.0067	0.0031	0.0016

Tabella A.35: Distorsione e varianza degli stimatori, Modello 7, $S_1 = 4, S_2 = 12$

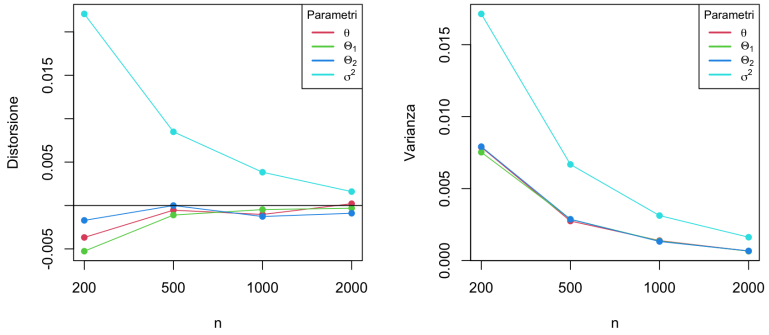


Figura A.35: Distorsione e varianza degli stimatori, Modello 7, $S_1 = 4, S_2 = 12$

	Ljung-Box classico				Ljung-Box modificato			
Ritardi	n=200	n=500	n=1000	n=2000	n=200	n=500	n=1000	n=2000
$h_1 = 2, h_2 = 2$	60.45	94.95	95.10	94.80	60.45	94.95	95.10	94.80
$h_1 = 3, h_2 = 3$	63.05	97.10	96.15	96.75	58.10	93.75	93.55	93.60
$h_1 = 4, h_2 = 4$	65.20	96.65	96.55	96.75	61.05	94.35	94.05	93.70
$h_1 = 5, h_2 = 5$	66.10	96.55	96.50	96.35	63.10	94.30	94.30	94.00

Tabella A.36: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 7, $S_1 = 4, S_2 = 12$

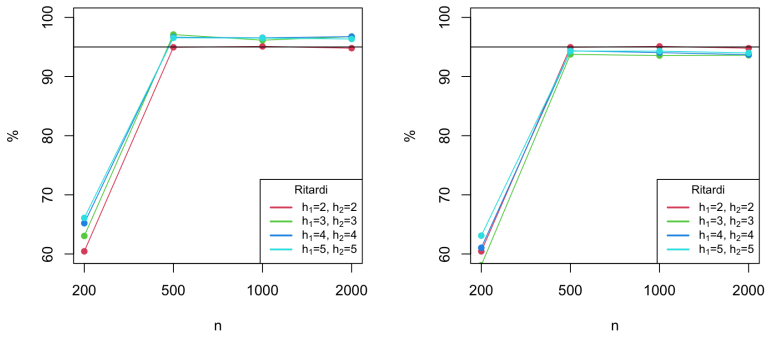


Figura A.36: Percentuale di non rifiuto test di Ljung-Box standard (a sinistra) e con i gdl modificati (a destra), Modello 7, $S_1 = 4, S_2 = 12$

Appendice B

Previsioni

B.0.1 Modello 2: mSARIMA$(1,0,0)(0,0,1)_4(0,0,1)_7$

Valori dei parametri $\phi = 0.5, \Theta_1 = 0.5, \Theta_2 = 0.5, \sigma_\varepsilon^2 = 1$

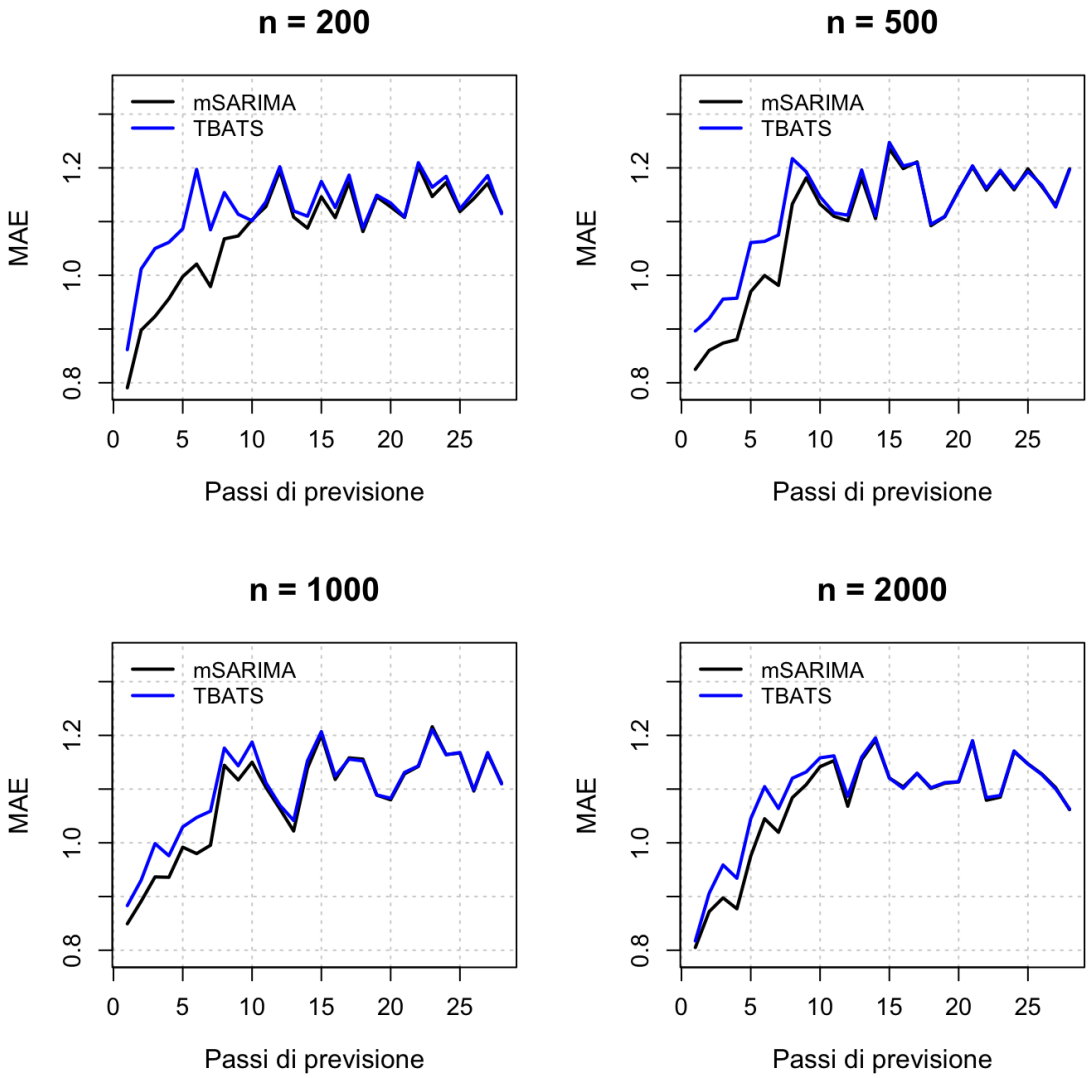


Figura B.1: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Modello 2.

I due modelli hanno prestazioni simili sebbene mSARIMA riesca ad avere un MAE inferiore nei primi passi di previsione.

B.0.2 Modello 3: mSARIMA$(0,0,0)(1,0,1)_4(1,0,1)_7$

Valori dei parametri $\Phi_1 = 0.4, \Phi_2 = 0.5, \Theta_1 = 0.5, \Theta_2 = 0.4, \sigma_\varepsilon^2 = 1$

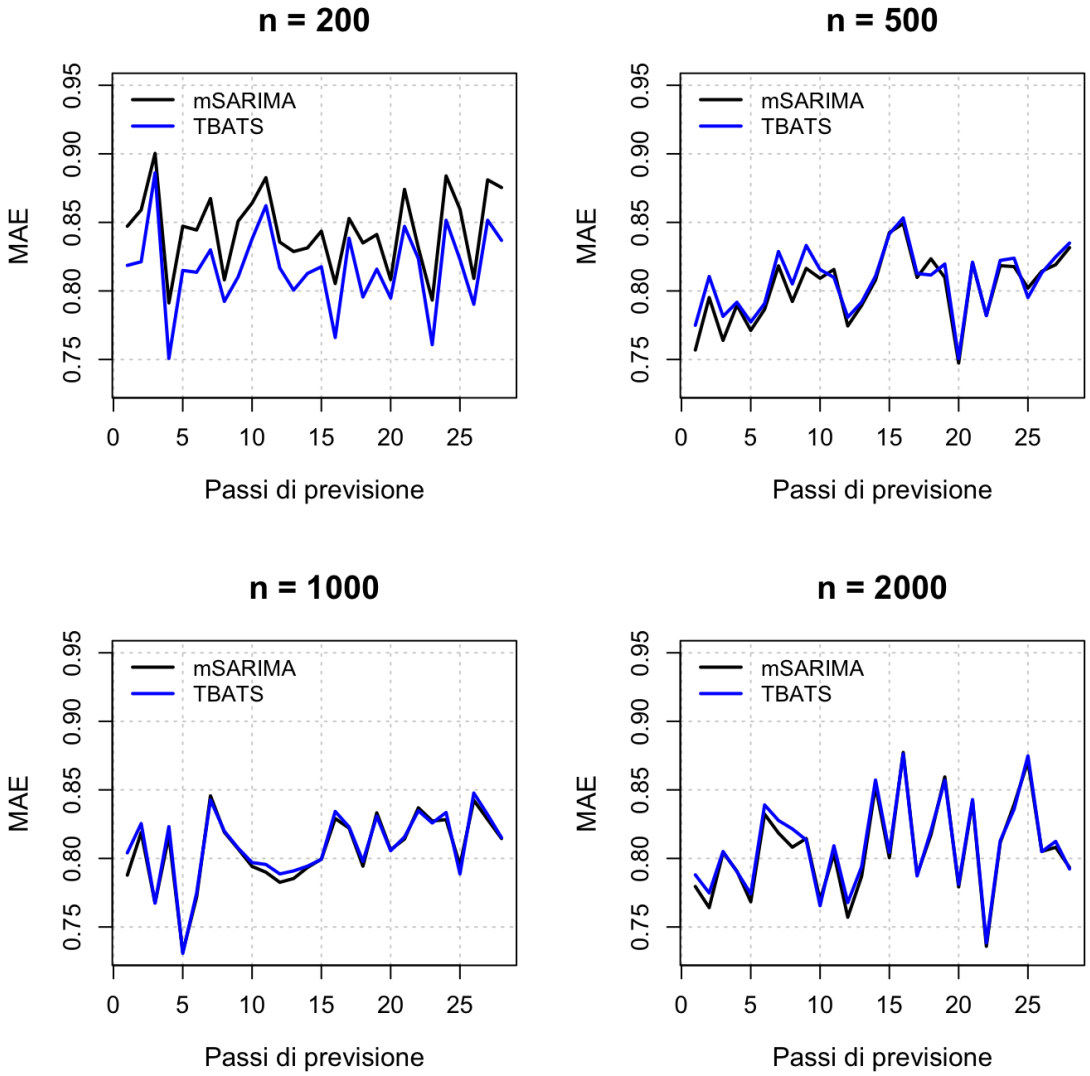


Figura B.2: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Modello 3.

I due modelli hanno prestazioni comparabili ma TBATS ha delle prestazioni leggermente superiori quando la numerosità della serie storica è molto bassa.

B.0.3 Modello 4: mSARIMA$(1,0,1)(0,1,0)_4(0,1,0)_7$

Valori dei parametri $\phi = 0.5, \theta = 0.6, \sigma_\varepsilon^2 = 1$

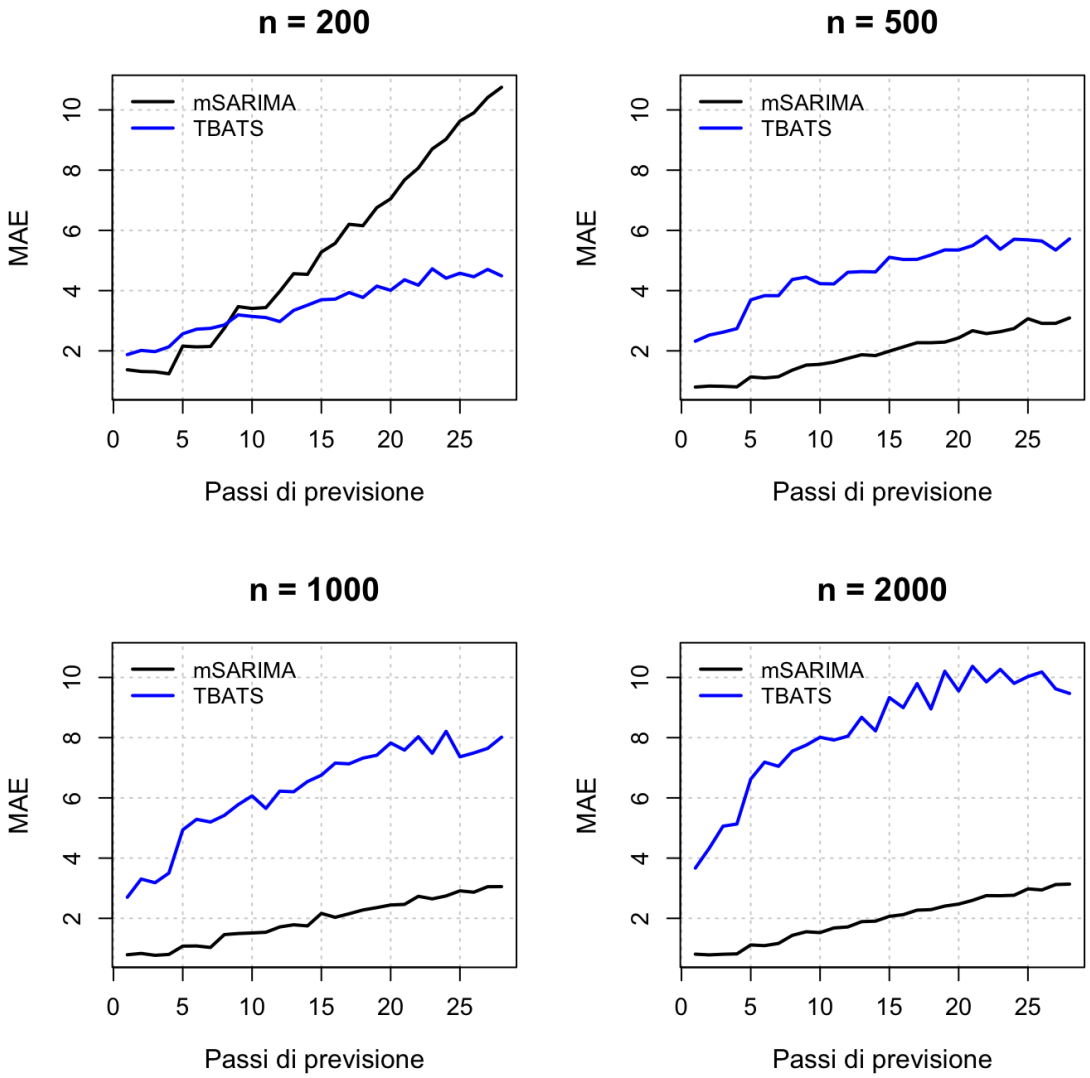


Figura B.3: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Modello 4.

La presenza di componenti integrate modifica il comportamento del MAE dei due modelli. Per serie con numerosità molto bassa il modello mSARIMA riscontra delle difficoltà; quando la numerosità aumenta, al contrario, ottiene performance molto superiori rispetto a quelle del modello TBATS.

B.0.4 Modello 5: mSARIMA$(0,0,0)(1,1,0)_4(1,1,0)_7$

Valori dei parametri $\Phi_1 = 0.5, \Phi_2 = 0.5, \sigma_\varepsilon^2 = 1$

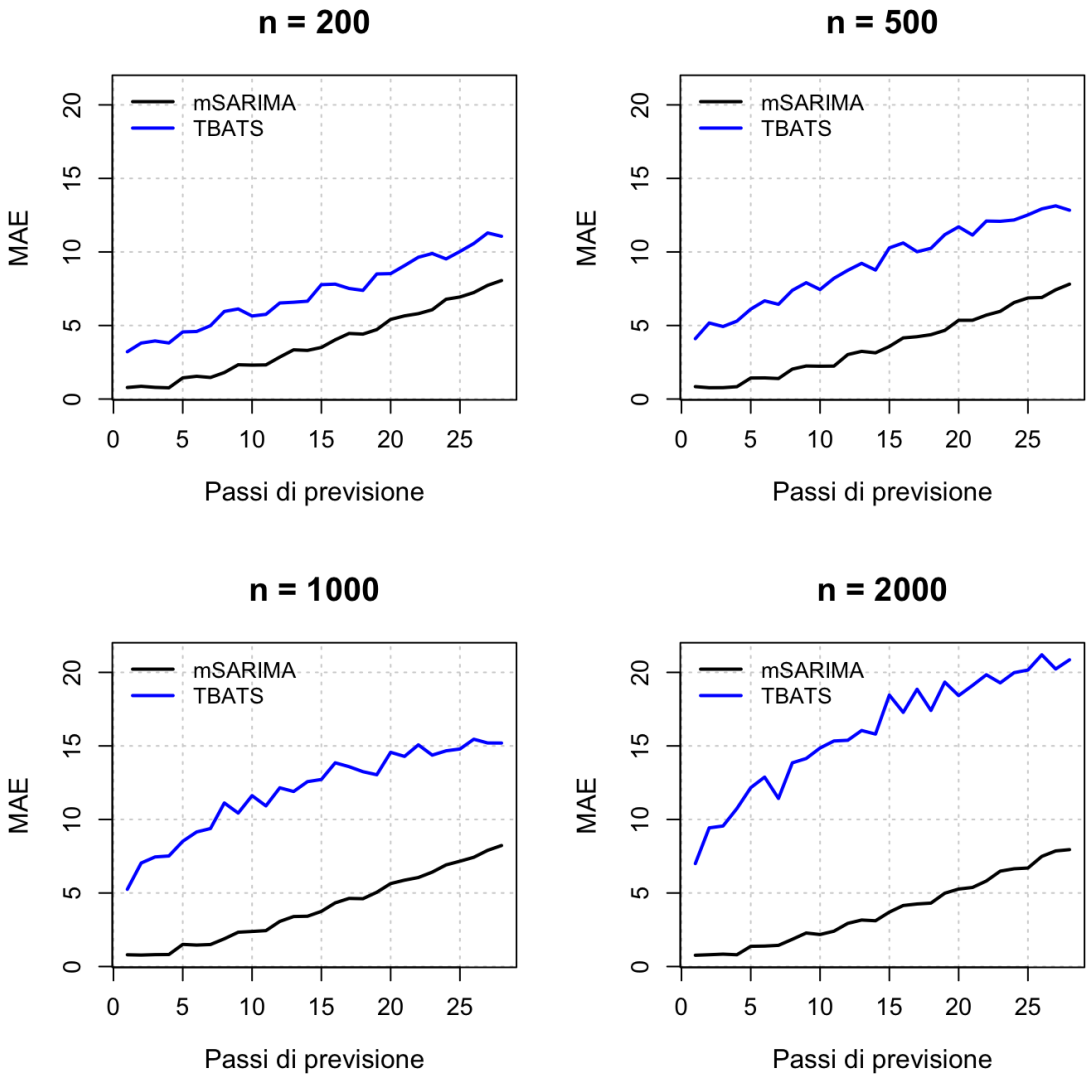


Figura B.4: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Modello 5.

Il modello mSARIMA ottiene sempre dei valori di MAE più bassi di quelli di TBATS. Le sue performance rimangono invariate all'aumentare della numerosità della serie storica, al contrario del modello TBATS che risulta particolarmente inadatto quando la numerosità aumenta.

B.0.5 Modello 6: mSARIMA$(0,0,0)(0,1,1)_4(0,1,1)_7$

Valori dei parametri $\Theta_1 = 0.5, \Theta_2 = 0.5, \sigma_\varepsilon^2 = 1$

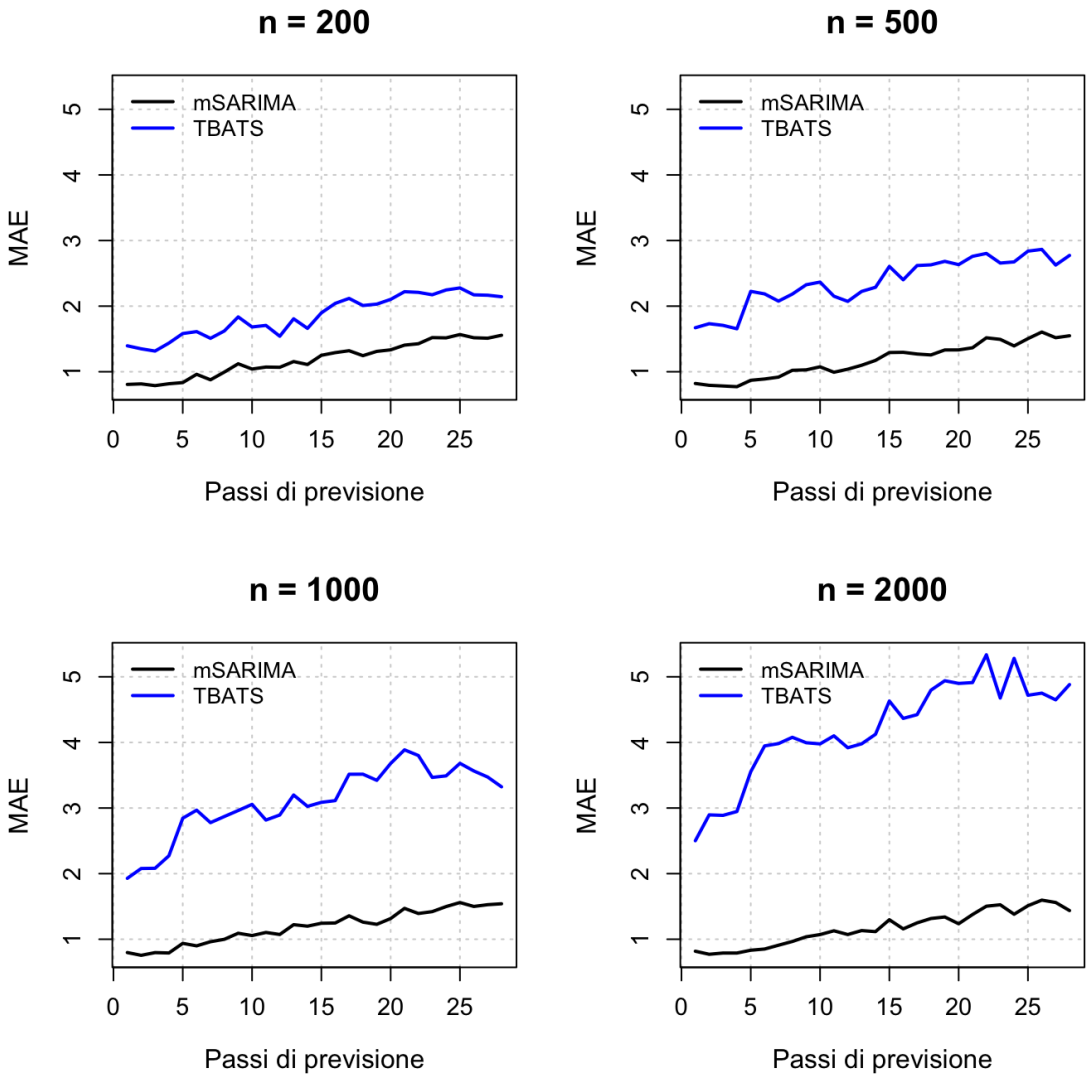


Figura B.5: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Modello 6.

Come accade nel Modello 5, mSARIMA ha buone prestazioni indipendentemente dalla numerosità della serie, mentre il MAE del modello TBATS aumenta all'aumentare di n.

B.0.6 Modello 7: mSARIMA$(0,1,1)(0,1,1)_4(0,1,1)_7$

Valori dei parametri $\theta = 0.5, \Theta_1 = 0.5, \Theta_2 = 0.5, \sigma_\varepsilon^2 = 1$

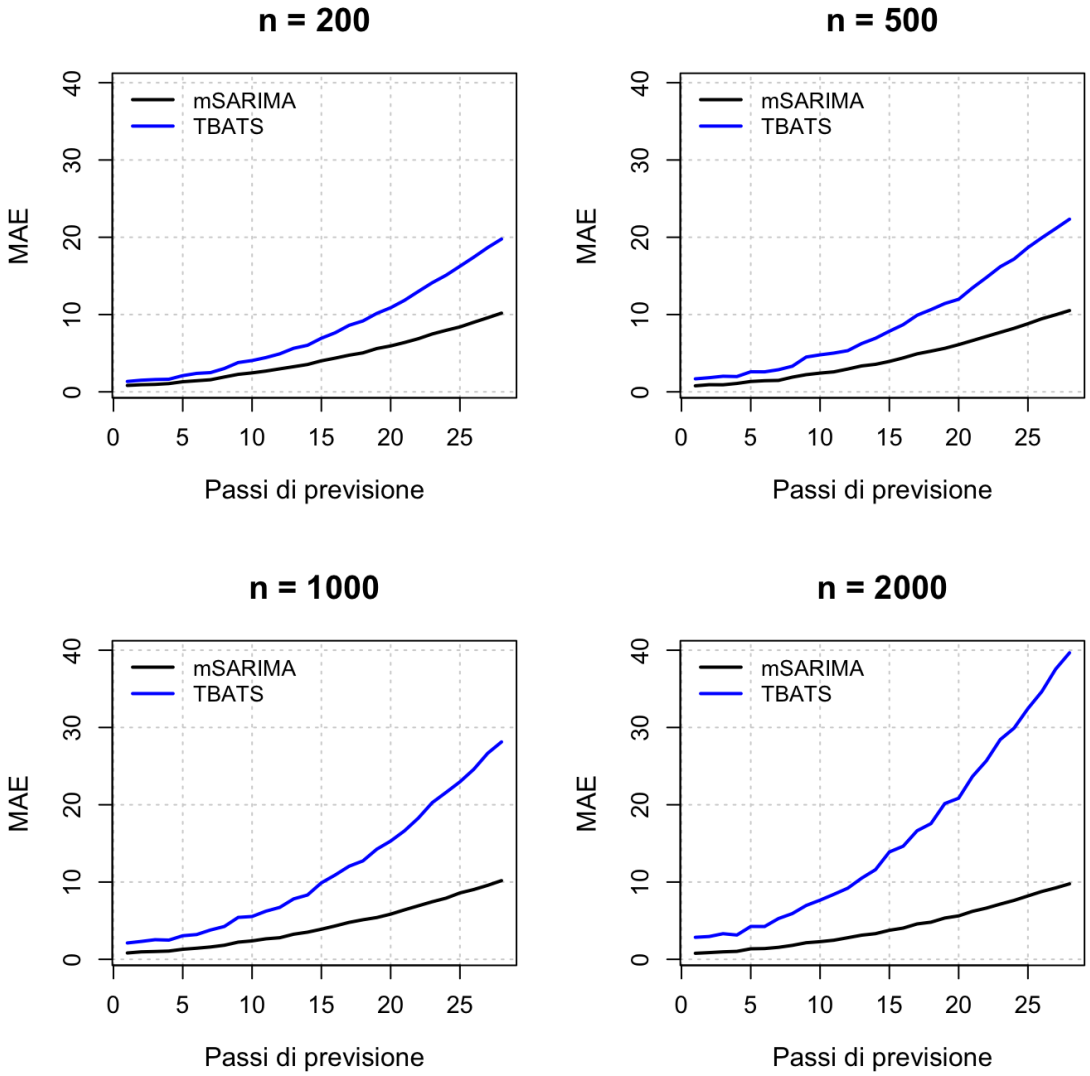


Figura B.6: Errore Assoluto Medio di previsione (MAE) per mSARIMA (linea nera) e TBATS (linea blu) per diverse numerosità della serie storica. Dati generati da Modello 7.

Ai primi passi di previsione i due modelli ottengono dei valori di MAE simili, ma con l'aumentare dei passi di previsione le performance di TBATS peggiorano più che proporzionalmente rispetto a quelle di mSARIMA, in particolare per serie storiche con numerosità più elevata.