POLITECNICO DI TORINO

Corso di apprendimento statistico

Tesina

Time Series analysis

Professore
prof. Gianluca Mastrantonio

Studente
Andrea Ruglioni

Anno Accademico 2022-2023

Contents

1 Introduzione **2**

2 Analisi preliminare **2**
2.1 Componenti della serie 3

3 Trattazione teorica **6**
3.1 ACF e PACF 6
3.1.1 ACF 6
3.1.2 PACF 6
3.2 Modello SARIMA 6
3.2.1 AR(p) 7
3.2.2 I(d) 7
3.2.3 MA(q) 7
3.2.4 SARIMA 7

4 Costruzione del modello **8**

5 Previsione **12**
5.1 Confronto tra modelli 13

6 Conclusione **15**

1 Introduzione

Lo scopo di questa tesina è analizzare la time series relativa agli arrivi internazionali in Australia dalla Nuova Zelanda dall'anno 1981 al 2012. In particolare, dopo un'analisi preliminare dei dati, costruiremo passo dopo passo il rispettivo modello SARIMA, con lo scopo di comprendere gli effetti di ciascuna componente del modello. Successivamente, metteremo alla prova tale modello facendo previsione di un arco temporale di circa 2 anni.

La trattazione teorica è basata sulla prima parte del corso di apprendimento statistico e sul libro "Forecasting: Principles and Practice"(3rd Edition), di Rob J Hyndman e George Athanasopoulos.

2 Analisi preliminare

Il dataset utilizzato è `aus_arrivals`, presente nella libreria *fpp3* e composto dai seguenti attributi:

- **Quarter**: l'indice della serie ad intervalli quadrimestrali a partire dall'anno 1981 fino al terzo quadrimestre del 2012.
- **Origin**: il paese di provenienza degli arrivi.
- **Arrivals**: il numero totale di arrivi in Australia dal paese Origin avvenuto nel quadrimestre.

I paesi di origine nel dataset sono Nuova Zelanda, Stati Uniti, Regno Unito e Giappone. Ai fini della nostra trattazione abbiamo selezionato esclusivamente gli arrivi provenienti dalla Nuova Zelanda.

Innanzitutto, svolgiamo un'analisi preliminare dei dati con un semplice plot, in modo da comprendere l'andamento generale della serie.

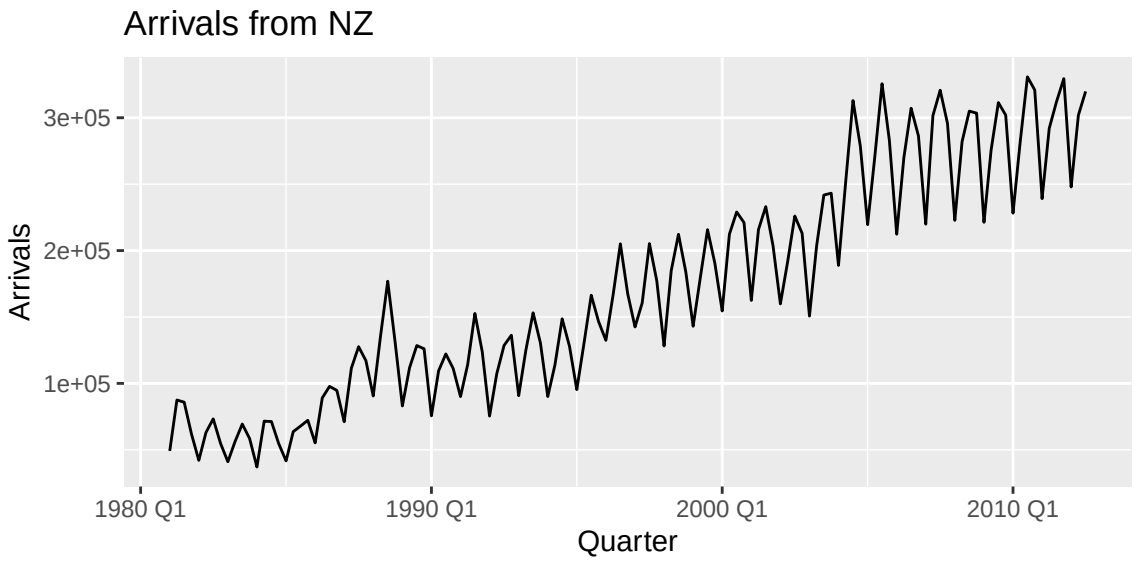

Figure 1: number of arrivals from New Zealand to Australia

Osserviamo la presenza di un trend crescente quasi-lineare durante tutto l'intervallo temporale. Inoltre, si intravede una componente stagionale dalla presenza di oscillazioni periodiche. Per evidenziarla, effettuiamo un'altro plot.

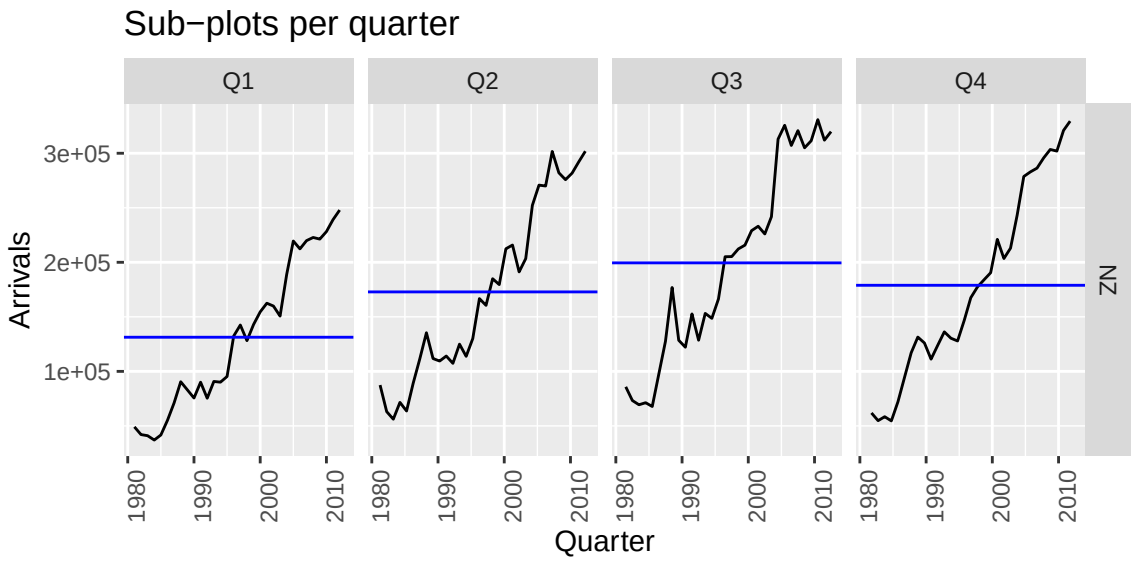

Figure 2: number of arrivals grouped by quarters

I grafici rappresentano l'andamento suddiviso per quadrimestre e le linee blu orizzontali indicano le medie per ciascun periodo. E' chiara la presenza di un effetto stagionale. Infatti, tendenzialmente, nel primo quadrimestre si ha il numero minore di arrivi nell'anno, mentre il picco si ottiene nel terzo quadrimestre.

2.1 Componenti della serie

Isoliamo più nettamente le componenti della nostra serie $\{x_t\}_{t\in T}$ tramite la suddivisione

$$x_t = S_t + T_t + R_t,$$

dove S_t è la componente stagionale, T_t rappresenta il trend, mentre R_t è il residuo. Tale decomposizione è però appropriata nel caso in cui l'ampiezza delle oscillazioni periodiche rimanga costante. A tal fine può essere utile effettuare una trasformazione *Box-Cox* in modo da stabilizzarne la varianza. Quest'ultima dipende da un parametro λ e produce una nuova serie y_t definita come:

$$y_t = \begin{cases} \log(x_t) & \text{se } \lambda = 0 \\ (\text{sign}(x_t)|x_t|^\lambda - 1)/\lambda & \text{altrimenti} \end{cases}$$

Quindi, per $\lambda = 0$ coincide con il logaritmo naturale, mentre per $\lambda \neq 0$ si ottiene una *power transformation*.

Il valore ottimale di λ è quello per cui le variazioni stagionali sono simili durante tutta la serie. Per noi, tale valore ottimale è dato per $\lambda = 0.33$, ottenuto grazie alla funzione `guerrero`, che ricava il miglior parametro con, appunto, il metodo di Guerrero.

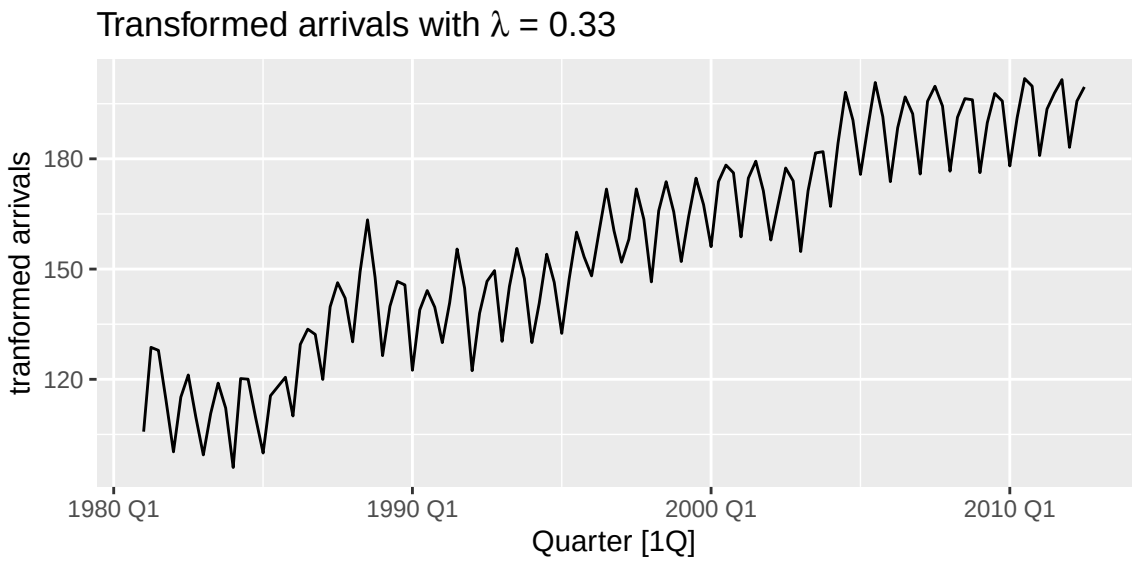

Figure 3: Box-Cox transformed number of arrivals

Con questa trasformazione la varianza della oscillazioni sembra effettivamente più stabile.

Tornando alla decomposizione della serie temporale, utilizziamo a tal fine il modello `STL`, parte della libreria *feasts*. STL è l'acronimo di *Seasonal and Trend decomposition using Loess*, dove loess è un metodo per stimare regressioni non lineari.

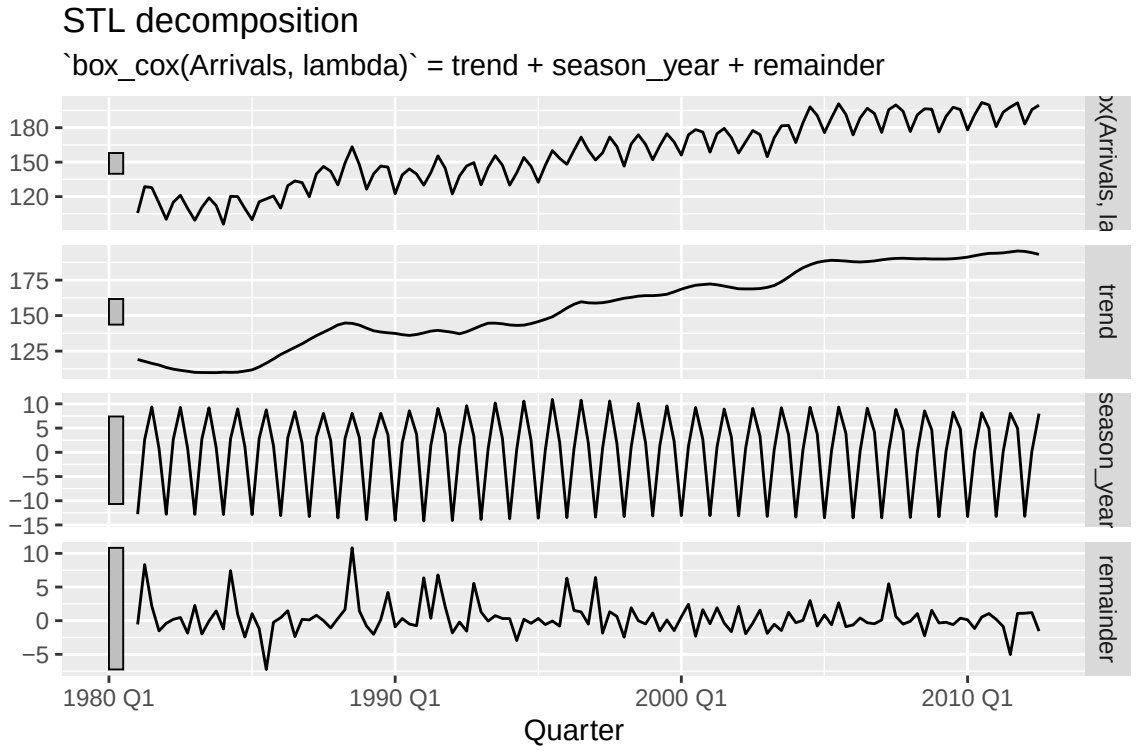

Figure 4: decomposed time series using STL function

Grazie a questo metodo, siamo in grado di confermare rigorosamente le nostre intuizioni sulla presenza del trend crescente quasi-lineare e sull'esistenza di una componente periodica annuale. E' interessante osservare che, grazie alla trasformazione Box-Cox, l'ampiezza delle oscillazioni stagionali sembra essere davvero costante.

3 Trattazione teorica

In questa sezione, ci occupiamo di definire il modello SARIMA, l'ACF e il PACF, in quanto queste nozioni saranno intensamente utilizzate in seguito. Ricordiamo prima il concetto di stazionarietà. Un processo si dice *stazionario* in senso forte se per ogni n-upla $(t_1, t_2, \ldots, t_n)$ di tempi e per ogni h, abbiamo che

$$P(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n) = P(X_{t_1+h} = x_1, X_{t_2+h} = x_2, \ldots, X_{t_n+h} = x_n),$$

ovvero la distribuzione non varia se trasliamo la serie di un tempo h. Intuitivamente, ciò significa che la serie temporale è stazionaria se le sue statistiche non dipendono dal tempo in cui la serie è osservata. Quindi time series con trend o stagionalità sono chiaramente non stazionarie, perchè la loro media varia nel tempo.

3.1 ACF e PACF

Queste quantità sono la sigla rispettivamente di *AutoCorrelation Function* e *Partial AutoCorrelation Function*, indicano quindi la correlazione della time series a tempi differenti, in modo da scoprire la presenza di regolarità nel suo comportamento.

3.1.1 ACF

Matematicamente il coefficiente di autocorrelazione è definito come

$$\rho_t(l) = \frac{\mathrm{Cov}(X_t, X_{t+l})}{\sqrt{\mathrm{Var}(X_t)\mathrm{Var}(X_{t+l})}}.$$

La dipendenza è data dal lag l e dall'istante temporale t. Se però la serie è stazionaria allora la covarianza è indipendente dal tempo t ed analogamente vale per il denominatore poichè la varianza è costante ad ogni tempo. Di conseguenza, il coefficente di autocorrelazione dipende esclusivamente dal lag e si può scrivere come

$$\rho(l) = \frac{\mathrm{Cov}(X_t, X_{t+l})}{\mathrm{Var}(X_t)},$$

con t un tempo qualsiasi. Si indica con il termine "correlogramma" il grafico di $\rho(l)$.

3.1.2 PACF

L'autocorrelazione parziale corrisponde all'autocorrelazione depurata dalle variabili che si trovano nel mezzo. Quest'ultima è data quindi da

$$\psi_t(l) = \frac{\mathrm{Cov}(X_t, X_{t+l}|X_{t+1}, \ldots, X_{t+l-1})}{\sqrt{\mathrm{Var}(X_t|X_{t+1}, \ldots, X_{t+l-1})\mathrm{Var}(X_{t+l}|X_{t+1}, \ldots, X_{t+l-1})}}.$$

Ancora, se la serie è stazionaria, l'espressione sopra si semplifica in quanto non ci sarà più dipendenza dal tempo e si ottiene

$$\psi(l) = \frac{\mathrm{Cov}(X_t, X_{t+l}|X_{t+1}, \ldots, X_{t+l-1})}{\mathrm{Var}(X_t|X_{t+1}, \ldots, X_{t+l-1})}.$$

3.2 Modello SARIMA

Il modello statistico ARIMA, acronimo di *AutoRegressive Integrated Moving Average*, è indicato come:

$$\mathrm{ARIMA}(p, d, q),$$

dove p, d, q, sono gli ordini rispettivamente della parte autoregressiva, di differenziazione e di media mobile. Introduciamo l'operatore di backshift B, per cui $Bx_t = x_{t-1}$, che ci permette di generalizzare con più facilità il modello. Le componenti dell'ARIMA sono le seguenti.

3.2.1 AR(p)

La parte autoregressiva predice la variabile utilizzando una combinazione lineare dei valori passati definendo il modello

$$x_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + w_t,$$

che si può scrivere utilizzando il backshift come

$$(1 - \alpha_1 B - \cdots - \alpha_p B^p) x_t = w_t,$$

dove $\alpha_i, i = 1, \ldots, p$ sono coefficienti e $w_t \sim \mathcal{N}(0, \sigma^2)$ è un rumore bianco. E' immediato vedere che questo processo è markoviano di ordine p, quindi l'autocorrelazione parziale si annulla per valori del lag l maggiori di p. Invece il coefficiente di correlazione diminuisce gradualmente. Per esempio, nell'AR(1) abbiamo che $\rho(l) = \alpha^l$, dove necessariamente $\alpha < 1$ per garantire la stazionarietà del processo, quindi decresce esponenzialmente.

3.2.2 I(d)

La differenziazione di ordine d è data dalla relazione

$$(1 - B)^d x_t = w_t.$$

Questo processo è spesso utilizzato per eliminare il trend da una time series e quindi renderla stazionaria. In particolare, un trend lineare, come nel nostro caso, può essere eliminato differenziando una sola volta. Infatti, ad esempio, integrando la semplice serie temporale $x_t = a + bt + w_t$, con w_t rumore bianco, che ha trend lineare, si ottiene

$$y_t = (1 - B)x_t = x_t - x_{t-1} = b + w_t - w_{t-1}.$$

Notiamo che mentre la serie $\{x_t\}_{t \geq 0}$ non è stazionaria in quanto $X_t \sim \mathcal{N}(a + bt, \sigma^2)$, la serie $\{y_t\}_{t \geq 0}$ è una moving average di ordine 1, quindi è stazionaria. Anticipiamo che nel modello SARIMA sono introdotte anche integrazioni stagionali in modo da eliminare tale componente con lo scopo di ottenere una serie stazionaria.

3.2.3 MA(q)

Il modello di moving average di ordine q è

$$x_t = w_t + \beta_1 w_{t-1} + \cdots + \beta_q w_{t-q},$$

o equivalentemente

$$x_t = (1 + \beta_1 B + \cdots + \beta_q B^q) w_t,$$

in cui $w_i, i = 0, 1, \ldots, q$ sono rumori bianchi. Inversamente al modello regressivo, si ha che in un MA(q), l'autocorrelazione con lag l maggiore di q è nulla.

3.2.4 SARIMA

Combinando questi modelli, si ottiene il modello ARIMA(p, d, q), che complessivamente è indicato dal processo

$$(1 - \alpha_1 B - \cdots - \alpha_p B^p)(1 - B)^d x_t = (1 + \beta_1 B + \cdots + \beta_q B^q) w_t.$$

Il modello SARIMA, *Seasonal ARIMA*, integra al modello anche la sua parte stagionale con analoghe metodologie. Questo si indica con

$$\text{ARIMA}(p, d, q)(P, D, Q)_m,$$

dove m è la stagionalità.

In conclusione, se volessimo costruire un modello SARIMA$(0, 1, 0)(2, 0, 3)_4$, la serie temporale si otterrebbe tramite

$$(1 - \alpha_1 B^4 - \alpha_2 B^8)(1 - B) x_t = (1 + \beta_1 B^4 + \beta_2 B^8 + \beta_3 B^{12}) w_t.$$

In cui la parte regressiva stagionale è data da $(1 - \alpha_1 B^4 - \alpha_2 B^8)$, successivamente abbiamo la differenziazione, mentre la componente di media mobile è posta a destra dell'uguaglianza.

4 Costruzione del modello

Il primo obiettivo è rendere la time series stazionaria. Cerchiamo quindi di eliminare il trend e l'andamento periodico differenziando la serie temporale, includendo anche differenziazioni stagionali.

Applichiamo un'integrazione di ordine $d = 1$, in modo da eliminare il chiaro trend quasi-lineare.

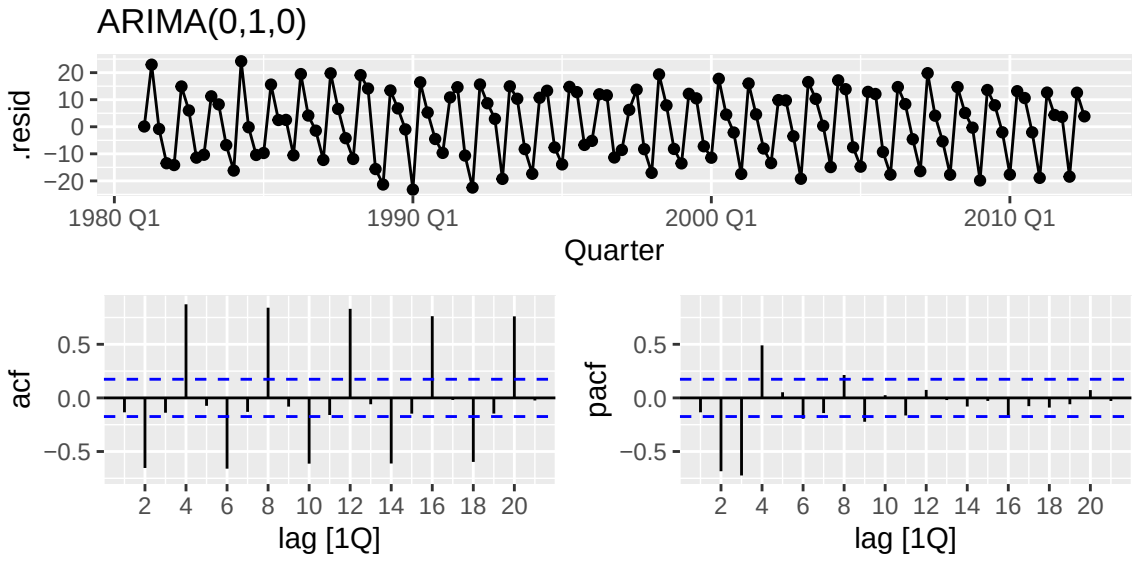

Figure 5: plots for ARIMA(0,0,0)(0,1,0)[4] model

In questo modo i residui hanno perso il loro trend e sembrano evere media approssimativamente nulla. Dalla precedente scomposizione STL della serie conosciamo la presenza della componente stagionale annuale. Questa può essere notata anche osservando l'attuale ACF, in cui sono presenti forti correlazioni ad ogni lag annuale. Differenziamo quindi nuovamente la serie temporale con un lag di 4. In questo modo i residui rimanenti dovrebbero risultare stazionari.

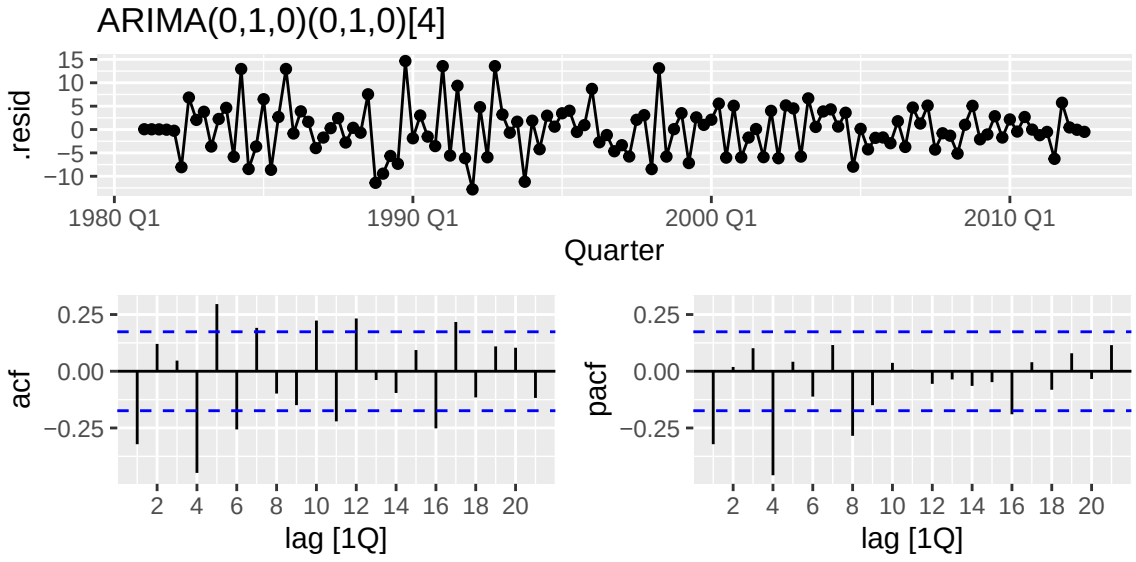

Figure 6: plots for ARIMA(0,1,0)(0,1,0)[4] model

L'effetto stagionale dei residui sembra scomparso e la serie risulta essere stazionaria.

Per confermare le precedenti intuizioni, possiamo utilizzare lo *unit root test*, in particolare noi useremo il Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test che utilizza per ipotesi nulla la stazionarietà della serie. La libreria *feast* offre le funzioni `unitroot_kpss`, `unitroot_ndiffs` e `unitroot_nsdiffs`. La prima effettua il test d'ipotesi e mostra, come nel nostro caso, se il p-value è maggiore di 0.1 e quindi la serie può essere considerata stazionaria. Le rimanenti funzioni ricavano rispettivamente gli ordini di integrazione d e D ottimali. In tabella mostriamo l'output del test kpss della serie già integrata e gli ordini di integrazioni suggeriti a partire dalla serie iniziale (che coincidono con quelli da noi applicati).

Table 1: unit root test results

kpss_pvalue	ndiffs	nsdiffs
0.1	1	1

Tornando al modello, osservando il correlogramma in Figura 6, possiamo notare che al lag 4 troviamo una forte correlazione. Invece, nel PACF è possibile vedere una correlazione decrescente nei lag stagionali (lag $= 4, 8, \dots$). Questi sono indicatori della presenza di un termine di MA(1) stagionale, che introduciamo al modello.

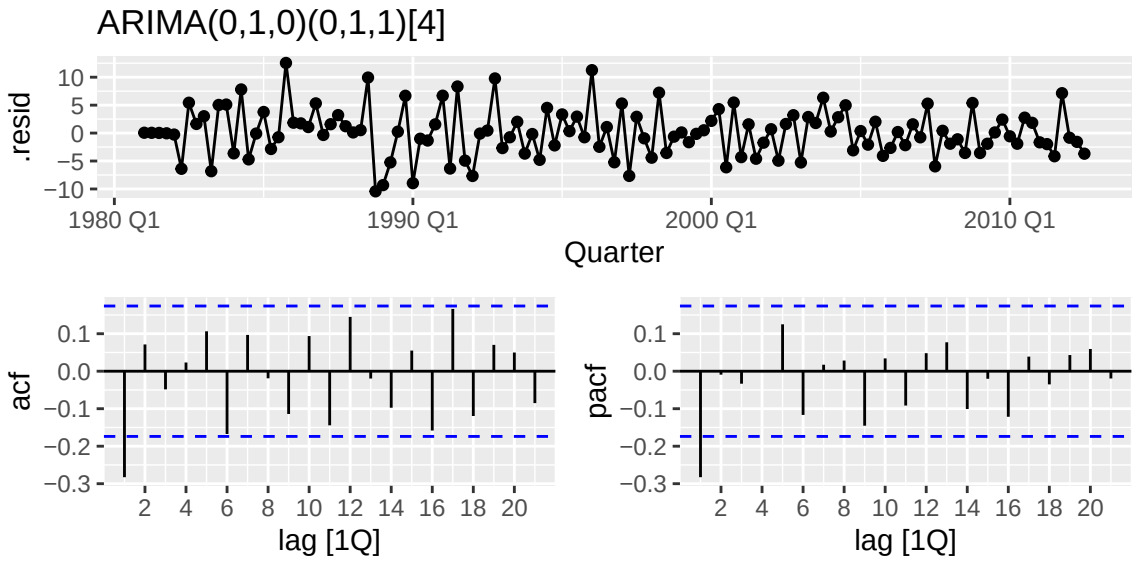

Figure 7: plots for ARIMA(0,1,0)(0,1,1)[4] model

Risulta ora una correlazione netta di lag unitario in entrambi i grafici ACF e PACF. Di conseguenza, possiamo modellizzare i residui aggiungendo una componente autoregressiva di ordine $p = 1$. Analizziamo il modello risultante con i seguenti plot.

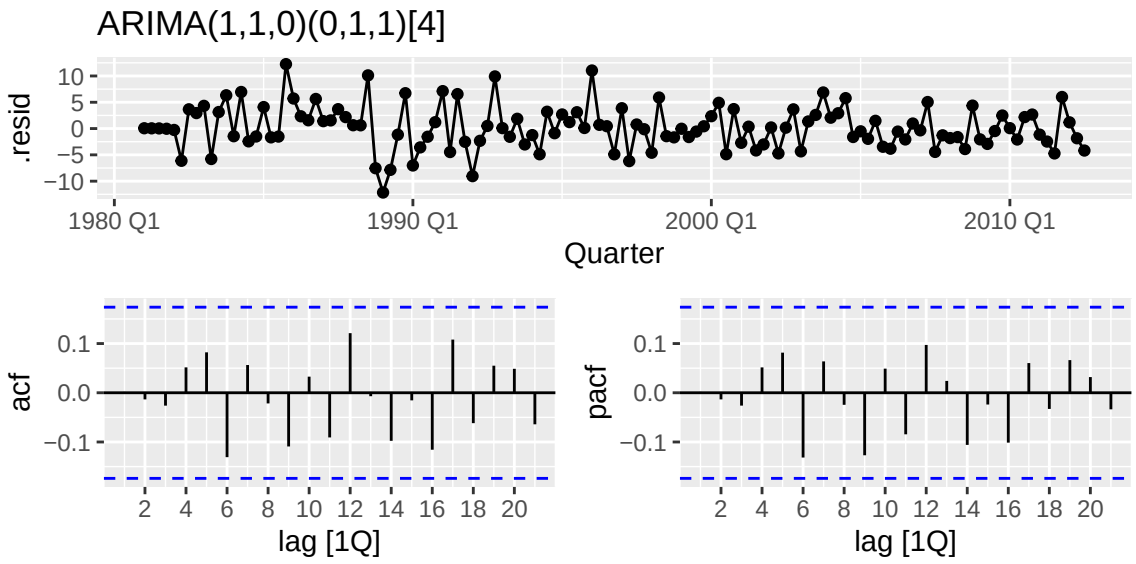

Figure 8: plots for ARIMA(1,1,0)(0,1,1)[4] model

Il modello finale sembra rappresentare esclusivamente white noise. Infatti, i grafici di autocorrelazione ed autocorrelazione parziale assumono esclusivamente valori non significativi. Verifichiamo che i residui abbiano effettivamente distribuzione normale con media pari a zero.

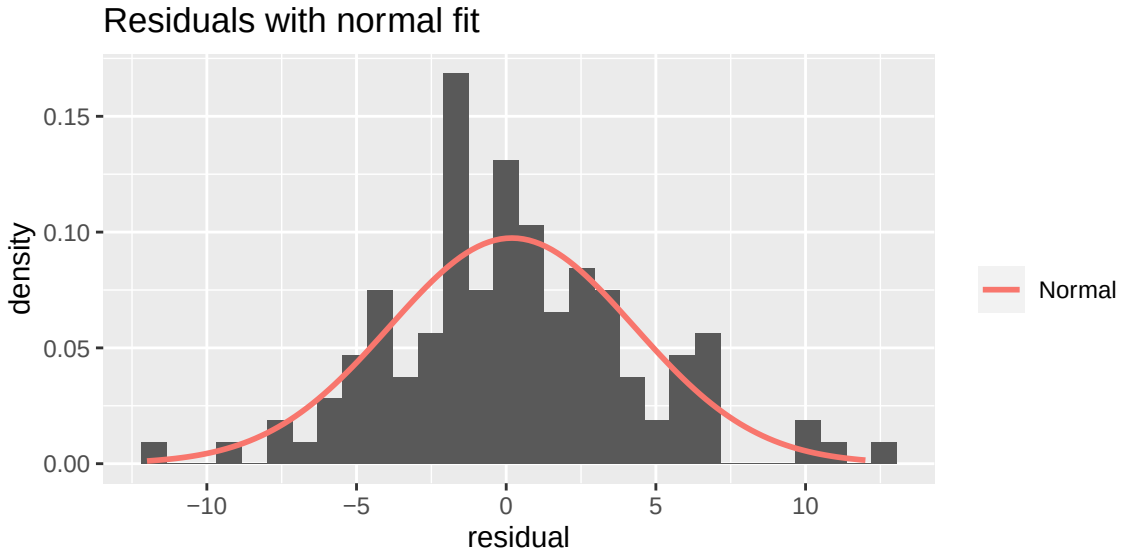

Figure 9: comparison between residuals and its normal fit

Dal grafico sembra che i residui abbiano una distribuzione tendenzialmente gaussiana centrata. Possiamo quindi utilizzare il modello ARIMA$(1,1,0)(0,1,1)_4$ per effettuare previsioni e studiare il comportamento degli arrivi quadrimestrali in Australia dalla Nuova Zelanda.

Confrontiamo ora il modello costruito con il modello generato automaticamente dalla funzione `ARIMA` della libreria *fable*, per verificare i risultati ottenuti.

Table 2: comparison between the two models

.model	AIC	p	d	q	P	D	Q	constant	period
ours	704.1146	1	1	0	0	1	1	FALSE	4
auto	705.3043	2	0	0	1	1	2	TRUE	4

Osserviamo che i due modelli differiscono per la scelta degli ordini. Utilizziamo per confrontarli l'AIC (Akaike's information criterion) che fornisce una misura della qualità della stima di un modello statistico tenendo conto sia della bontà di adattamento che della complessità del modello. Formalmente questo è dato da

$$AIC = 2k - 2\ln(L),$$

dove k è il numero di parametri ed L il valore massimizzato della verosimiglianza per il modello studiato. Notiamo che l'AIC del nostro SARIMA è leggermente minore rispetto quello automaticamentre prodotto da `ARIMA`, quindi il nostro sembra apparentemente migliore. Questo può essere dovuto al fatto che la funzione utilizza un algoritmo greedy e varie approssimazioni in modo da velocizzare la ricerca del modello. Trova quindi una combinazione di ordini ottimale localmente e non globalmente.

5 Previsione

In questa sezione effettuiamo una previsione con un'orizzonte temporale di circa 2 anni per controllare la validità del modello. Ricordiamo che il modello costruito è sulla trasformazione Box-Cox della serie temporale, di conseguenza è necessario effettuare la trasformazione inversa per essere in grado di prevedere gli arrivi. Prima, plottiamo la vera time series contro i valori fittati, in modo da vedere se il nostro modello segue l'andamento dell'effettiva serie.

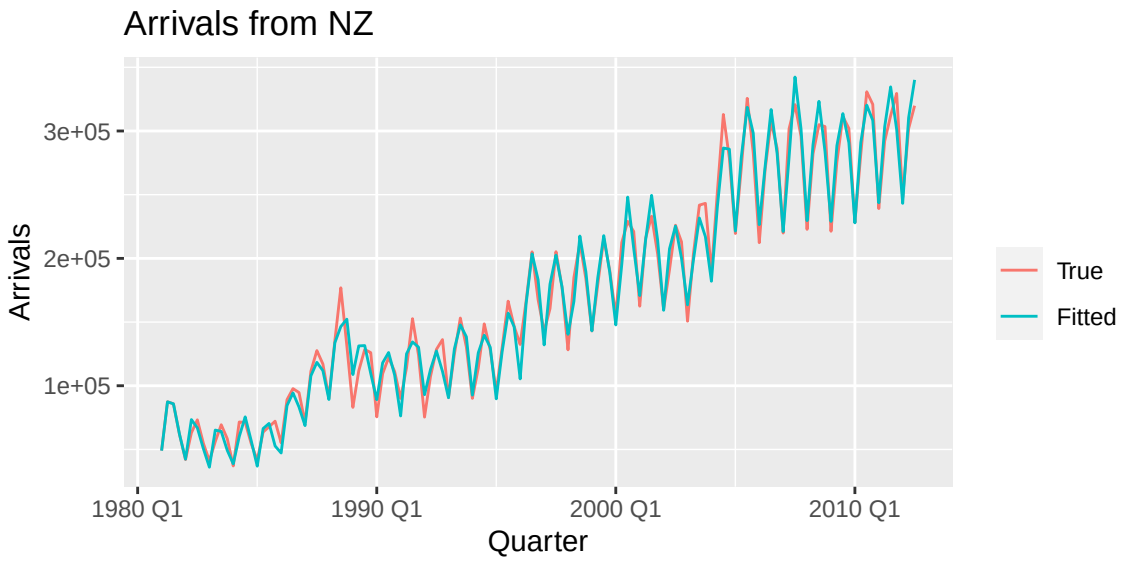

Figure 10: plots of real data and model fitted values

Possiamo notare che il modello sembra approssimare fedelmente il vero valore della serie temporale ed, in generale, il suo comportamento.

Infine, rappresentiamo l'andamento della previsione assieme ai suoi relativi intervalli di confidenza dell'80% e del 95%. Utilizziamo come training set i dati precedenti all'anno 2011, in modo da avere a disposizione gli ultimi 7 quadrimestri per confrontare la nostra previsione con il reale andamento della serie.

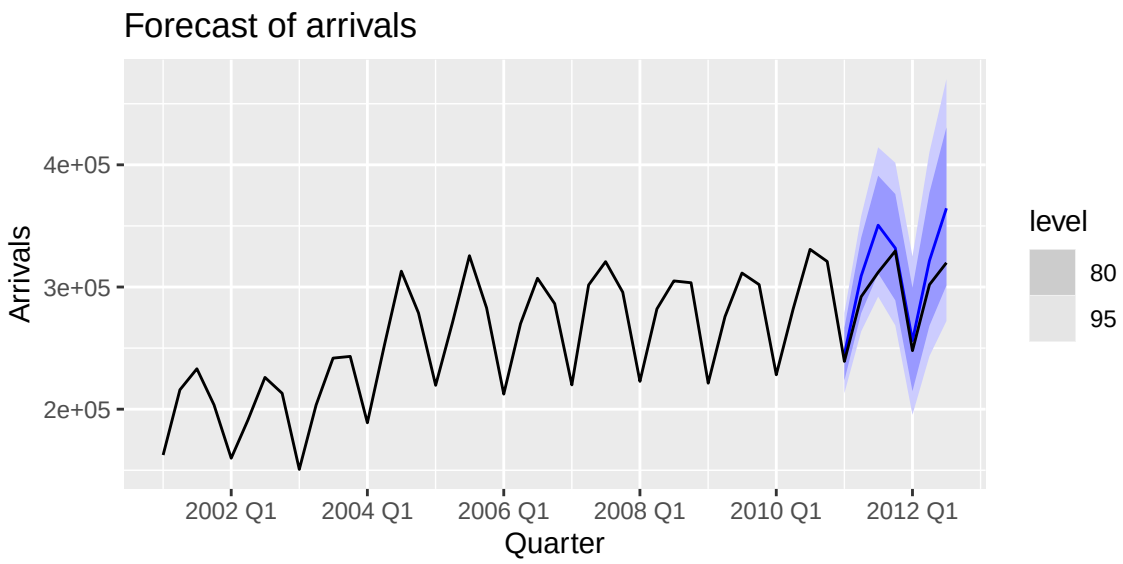

Figure 11: forecast of the last 7 quarters

Osserviamo che il modello sembra prevedere in modo coerente il successivo orizzonte temporale, seguendo l'andamento stagionale ed il trend crescente dei dati reali. In particolare, il vero andamento risiede sempre all'interno dell'intervallo di confidenza all'80%.

5.1 Confronto tra modelli

Mostriamo ora anche la previsione del modello precedentemente generato in automatico dalla funzione `ARIMA`, in modo da confrontarli in azione.

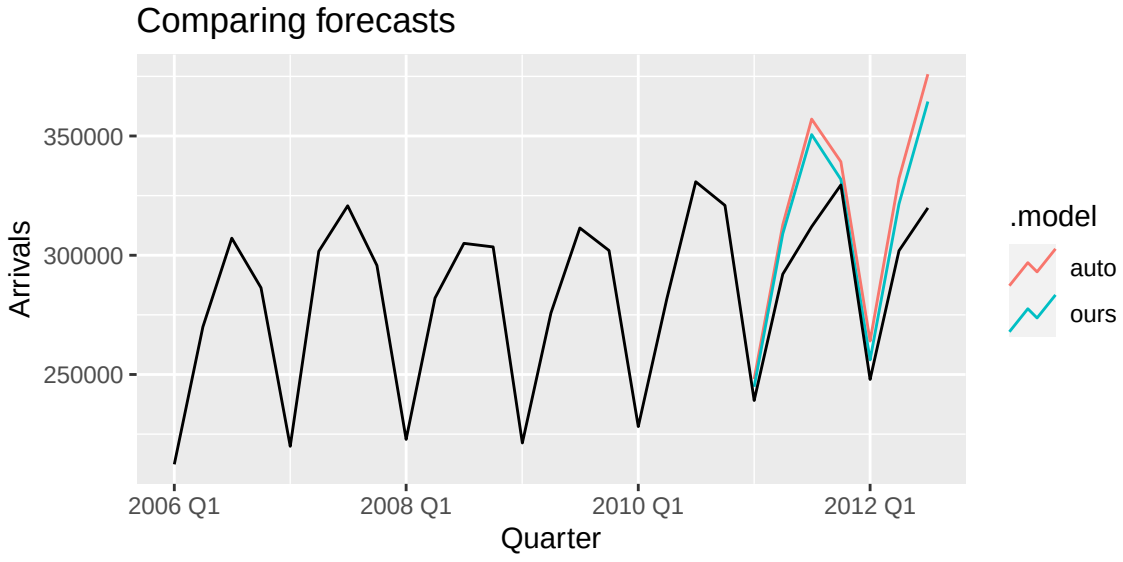

Figure 12: comparision between the two built model

Le due previsioni sono molto simili ed il modello costruito sembra essere leggermente più fedele ai dati reali, in accordo con le considerazioni svolte precedentemente. Indicando con $e_t = x_t - \hat{x}_t$ l'errore della previsione

al tempo t, confrontiamo le due previsioni in maniera oggettiva utilizzando i seguenti criteri

- *Root Mean Squared Error*: RMSE $= \sqrt{\text{mean}(e_t^2)}$.
- *Mean Absolute Error*: MAE $= \text{mean}(|e_t|)$.

Questi metodi sono molto comuni e semplici. Il loro valore dipende dalla scala di misura utilizzata, quindi è corretto il confronto nel nostro caso ma non possono essere utilizzati per confrontare modelli su data sets diversi. In tali casi è possibile utilizzare la valutazione detta

- *Mean Absolute Percentage Error*: MAPE $= \text{mean}(|p_t|)$

dove $p_t = 100e_t/x_t$. Questo criterio ha il vantaggio di essere facilmente interpretabile come errore relativo e di essere senza unità di misura. Lo svantaggio è che risulta indefinito o impraticabile se la serie assume valori vicini allo zero.

Table 3: evaluation criteria to compare the two models

.model	RMSE	MAE	MAPE
auto	31533.90	26712.84	8.90238
ours	24659.47	19434.49	6.43347

I valori ottenuti del RMSE e MAE sono molto grandi perchè l'ordine di grandezza dei dati è di 10^6. Per questo motivo è più comodo l'utilizzo dell'errore percentuale, che risalta la maggiore capacità espressiva del nostro modello.

6 Conclusione

Per concludere, abbiamo analizzato con risultati soddisfacenti la serie temporale sugli arrivi in Australia a partire dalla Nuova Zelanda, costruendone un nostro modello.

In particolare, siamo partiti studiando intuitivamente il trend e la stagionalità della serie, per poi, utilizzando la decomposizione svolta dalla funzione `STL`, analizzare in maniera più dettagliata il suo comportamento.

Dopo una trattazione teorica sul modello SARIMA e sulle funzioni ACF e PACF, abbiamo costruito passo dopo passo, il modello esplicativo della serie. Questa sezione è stata particolarmente interessante, in quanto siamo riusciti a distinguere le varie parti del modello osservando i correlogrammi ed il plot dei residui, comprendendo cosi il comportamento di ciascuna componente.

Infine, abbiamo confrontato il modello costruito con il modello automaticamente generato dalla funzione `ARIMA`. Abbiamo effettuato la previsione di un intervallo temporale di circa due anni, notando la miglior performance del nostro modello.