

L'impatto economico delle cene e pranzi fuori in Australia: un report dettagliato con analisi di previsione

ADVANCED STATISTICAL LEARNING II

Carmela Pia Senatore
Universita' degli studi di Salerno
Matricola: 0522501721



INDICE



01

Presentazione dei dati

Presentazione della serie storica con autoplot e rilevazione di valori anomali

02

Analisi descrittiva della serie storica

Analisi descrittiva della serie storica per la comprensione di : trend, stagionalità e ciclicità

03

Modelli di previsione Benchmark

Analisi dei modelli benchmark : drift method, naive, seasonal naive e average method

04

Implementazione dei modelli

Definizione, addestramento e previsione di diversi modelli: lineare autoregressivo, lineare multiplo, rete neurale autoregressiva.

05

Confronto tra modelli

06

Conclusioni

L'impatto economico delle cene e pranzi fuori in Australia: un report dettagliato con analisi di previsione

1. Presentazione dei dati

La serie storica presa in esame è contenuta all'interno del pacchetto **fpp2**. Il nome della serie storica è **auscafe** ed è una serie univariata mensile rappresentante le spese totale in cene e pranzi fuori in bar, ristoranti, servizi di cibo da asporto in Australia (bilioni di dollari). L'arco di tempo considerato varia da aprile 1982 a settembre 2017, per cui la dimensione del dataset è di 408 osservazioni.

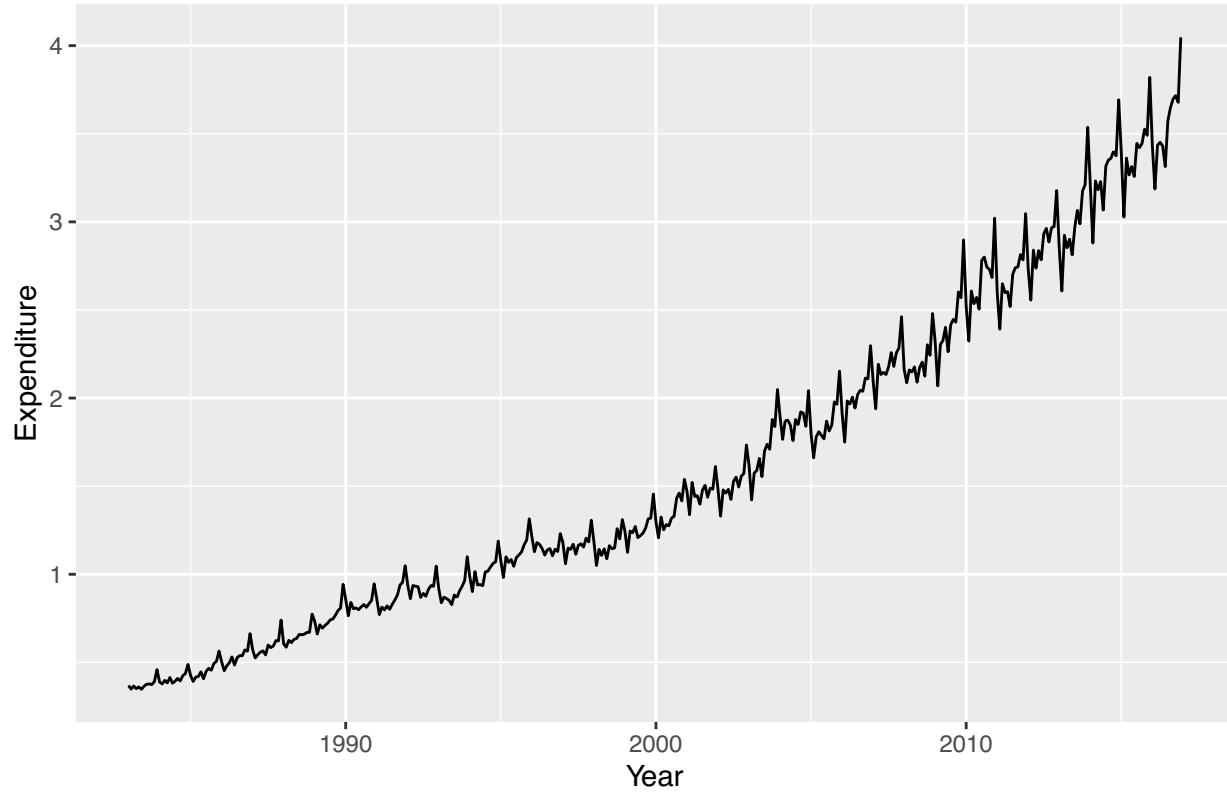
Poichè per il 1982 le osservazioni non partono da inizio anno ma dal quarto mese e per il 2017 le osservazioni non terminano con la fine dell'anno ma con settembre, i dati prenderanno in considerazione l'arco temporale degli anni completi dal 1983 al 2016.

La prima analisi effettuata riguarda la qualità dei dati. La presenza di valori mancanti o valori anomali può notevolmente influire sull'analisi. Con la funzione **statsNA** si verifica la presenza di valori mancanti. La serie non presenta missing values.

```
## [1] "Length of time series:"  
## [1] 408  
## [1] "-----"  
## [1] "Number of Missing Values:"  
## [1] 0  
## [1] "-----"  
## [1] "Percentage of Missing Values:"  
## [1] "0%"  
## [1] "-----"  
## [1] "Number of Gaps:"  
## [1] 0  
## [1] "-----"  
## [1] "Average Gap Size:"  
## [1] 0  
## [1] "-----"  
## [1] "No NAs in the time series."  
## [1] "-----"  
  
## [1] "There are no NAs in the time series"
```

Per il riconoscimento dei valori anomali la rappresentazione grafica della serie potrebbe essere d'aiuto.

Monthly expenditure on eating out in Australia



Come da evidenza grafica, non sembrano esserci valori anomali.

2. Analisi descrittiva della serie storica

L'analisi descrittiva della serie storica ha come obiettivo principale la determinazione delle principali componenti: trend, stagionalità e ciclicità, oltre che l'individuazione di irregolarità.

2.1 Trend, Stagionalità e ciclicità

Come è possibile notare dal precedente autoplot() la serie storica presenta:

- un forte trend prettamente crescente, il livello medio di spese per mangiare fuori tende a crescere nel tempo;
- una componente di stagionalità che deve essere analizzata nello specifico;
- non presenta ciclicità;
- Si nota come è presente un incremento di variabilità nel tempo; inizialmente le oscillazioni sono molto contenute poi con il passare del tempo incrementano di variabilità.

Il grafico per il season plot riporta in basso i mesi per comprendere come la serie storica si comporta mensilmente per tutti gli anni osservati , riportati sulla colonna.

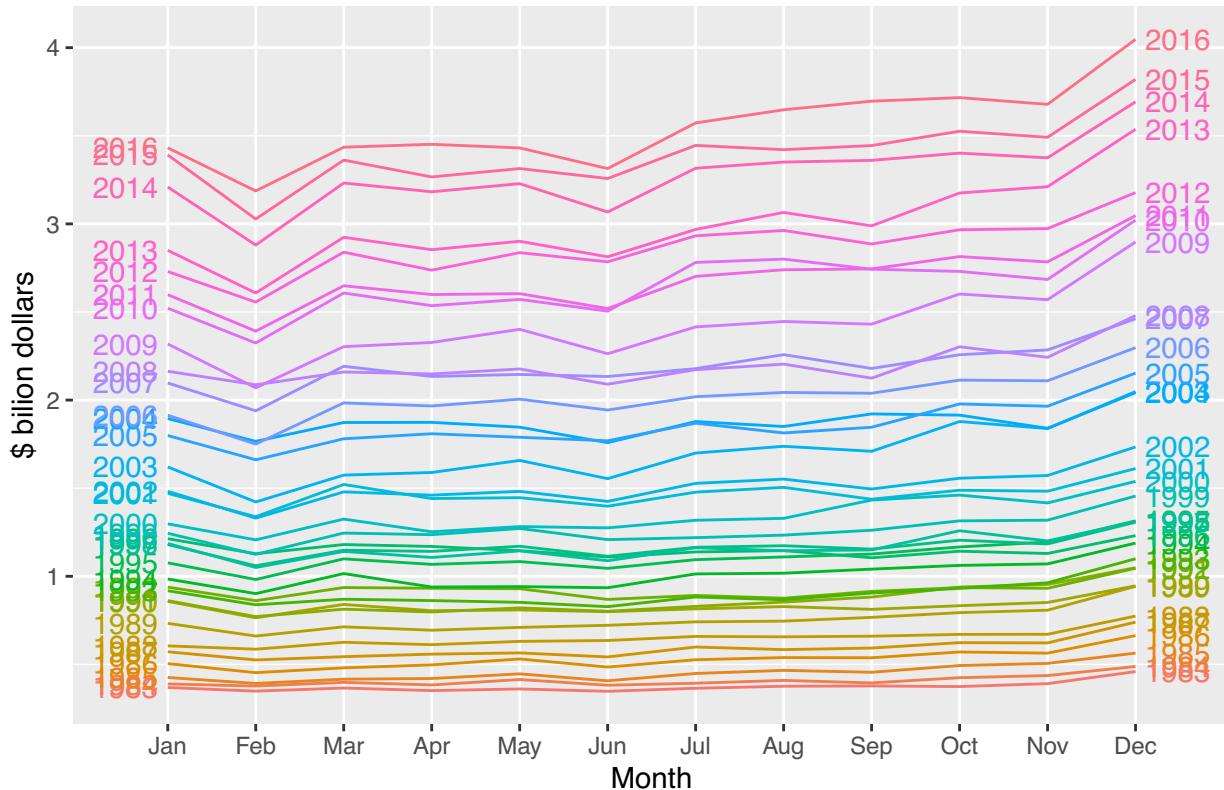
Gennaio costantemente per tutti gli anni ha un decremento di spese. Allo stesso modo conseguie un decremento a febbraio, un incremento a marzo, decremento a giugno e incremento a luglio. In particolar modo

il mese in cui l'incremento risulta essere maggiore è Dicembre. Questo comportamento nello stesso mese di ogni anno è espressione di stagionalità.

A febbraio il decremento potrebbe essere associato alla fine dell'estate e ritorno alla routine, oltre ad essere il mese più caldo dell'anno. Presumibilmente la stagionalità e gli incrementi a luglio e dicembre sono dovuti a particolari situazioni in Australia. In Australia, luglio e dicembre sono mesi di vacanza molto popolari. Luglio coincide con le vacanze invernali delle scuole, mentre dicembre è il periodo delle vacanze estive. D'altra parte, dicembre è il mese delle festività natalizie, e molte persone festeggiano spesso mangiando fuori o cenando in ristoranti. In ogni caso, la componente di stagionalità è netta.

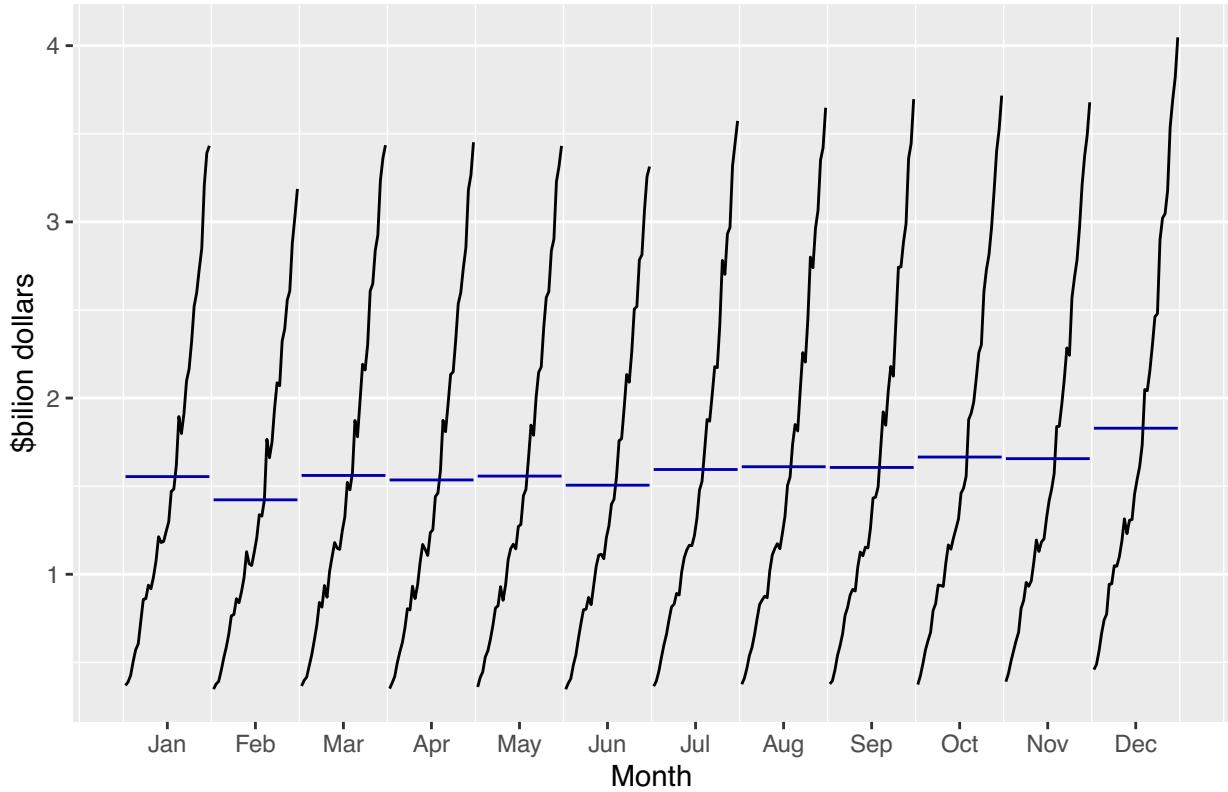
E' possibile anche analizzare la componente di trend, focalizzando l'attenzione sul lato: il valore iniziale di spese in pranzi e cene fuori parte da 0,33 bilioni di dollari nel 1982 e si completa nel 2016 con oltre 4 bilioni. Il livello medio è sempre aumentato nel tempo esprimendo un trend crescente.

Seasonal plot: expenditure in eating out



La verifica della stagionalità può essere anche svolta attraverso l'analisi del seasonal subseries plot. Il seasonal subseries plot comprende dei minigrafici non più contenenti punto per punto per ogni anno, piuttosto dati relativi a interi mesi per ogni anno. La differenza delle medie di febbraio e dicembre è notevole, considerando che l'unità di misura è bilioni di dollari. La crescita verso l'alto per ciascun mese conferma la presenza di trend positivo e crescente.

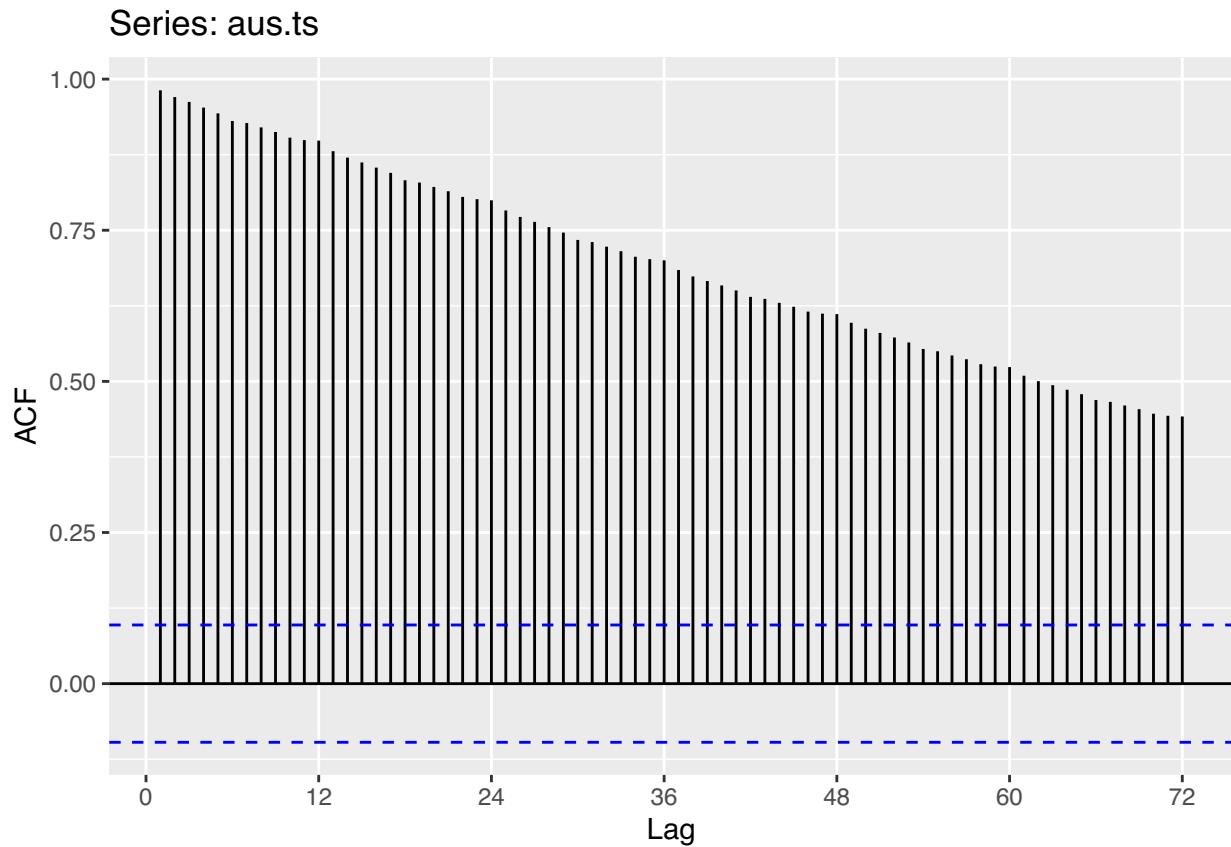
Subseries plot: expenditure in eating out



La serie storica analizzata non è stazionaria, in quanto le sue proprietà statistiche, influenzate da Trend e Stagionalità, dipendono dal tempo in cui la serie viene osservata.

2.2 ACF:

Il correlogramma mostra le autocorrelazioni della serie storica in funzione del ritardo di riferimento. Il valore di autocorrelazione più alto lo si osserva al ritardo uno e poi tende lentamente verso zero. I picchi tendono a ripetersi nel dodicesimo mese. Nelle serie white noise ogni correlazione si presenta prossima allo zero. Nel caso specifico la serie non è white noise in quanto tutti i picchi sono al di fuori delle bande di confidenza, individuate dalle linee blu tratteggiate indicanti le correlazioni come significativamente diverse da zero, la presenza di dati stagionali comporta autocorrelazioni più grandi per i ritardi stagionali. Il correlogramma, inoltre, presenta il tipico modello delle serie storiche caratterizzate da una componente di trend: le autocorrelazioni hanno valori positivi che diminuiscono lentamente all'aumentare del ritardo.



2.3 Analisi

La spesa varia leggermente da mese a mese. Il mese con la spesa più bassa è febbraio con 1.47 bilioni, mentre il mese con la spesa più alta è agosto con in media 1.63 bilioni. La spesa sembra aumentare leggermente nei mesi estivi e autunnali. Ad esempio, agosto e settembre mostrano valori elevati.

```
## # A tsibble: 12 x 2 [1]
##   Mese month_expenditure
##   <dbl>          <dbl>
## 1     1          1.61
## 2     2          1.47
## 3     3          1.62
## 4     4          1.56
## 5     5          1.58
## 6     6          1.53
## 7     7          1.62
## 8     8          1.63
## 9     9          1.63
## 10    10         1.63
## 11    11         1.62
## 12    12         1.79
```

3. Modelli di previsione Benchmark

Un metodo di previsione deve tener conto dei fattori dominanti della serie, nel caso particolare Trend e Stagionalità. Vengono calcolati i modelli di benchmark per dimostrare che effettivamente modelli più complessi

riescono a spiegare meglio il fenomeno oggetto di studio.

In questa sede verranno confrontate le prestazioni di quattro semplici modelli di previsione:

1. Average Method.

2. Naive Method.

3. Seasonal Naive Method.

4. Drift Method.

Per misurare la Bontà di Adattamento/Previsione verranno calcolati i seguenti parametri di errore:

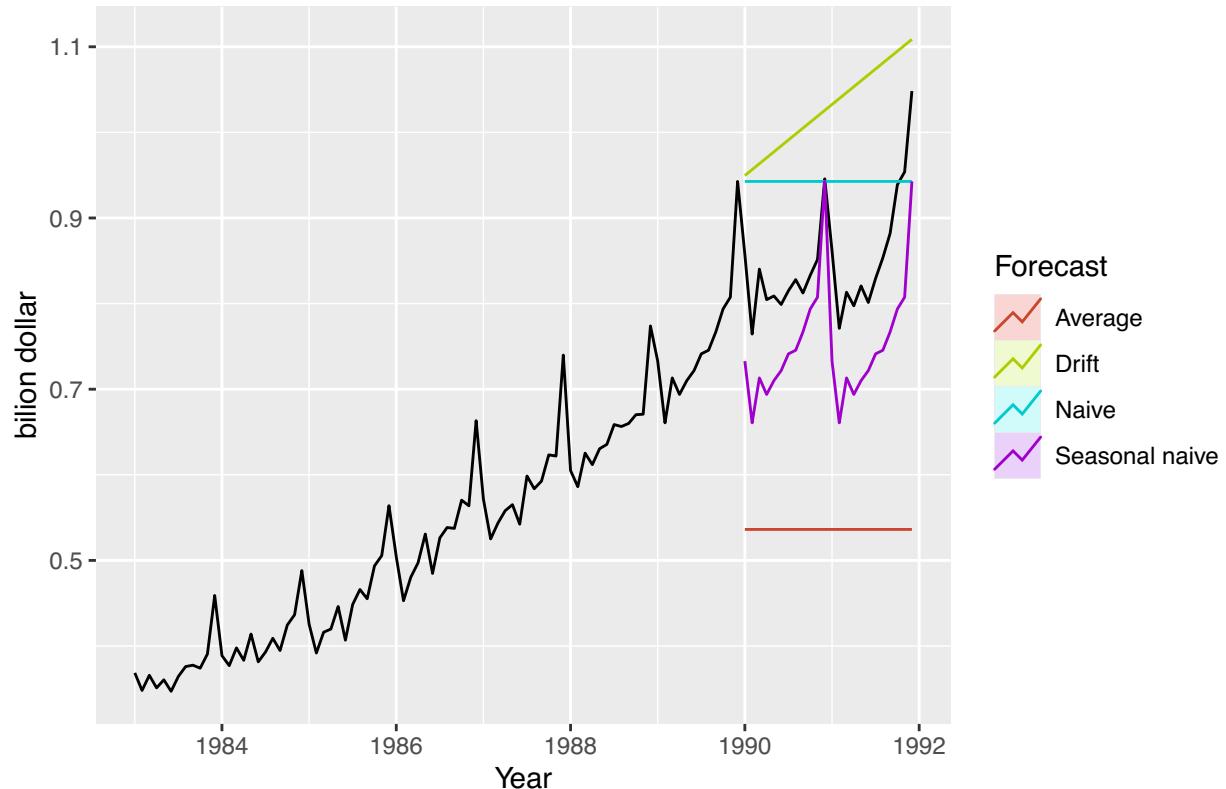
- **errore medio assoluto** (mean absolute error: MAE): media aritmetica degli errori presi in valore assoluto;
- **radice errore quadratico medio** (root mean squared error: RMSE): radice quadrata della media aritmetica dei quadrati degli errori;
- **errore medio assoluto percentuale** (mean absolute percentage error: MAPE): media aritmetica degli errori relativi, presi in valore assoluto e moltiplicati per 100. In realtà non si tratta di un indice simmetrico in quanto viene dato un peso maggiore agli errori negativi rispetto agli errori positivi, per questo viene considerato un indice corretto noto come *symmetric mean absolute percentage error, SMAPE*.

Per verificare e comprendere i modelli di previsione, viene effettuata una divisione in **Training set e Test set**. Per far sì che poi successivamente sia possibile confrontare i risultati con i modelli implementati per serie multivariante, gli archi temporali di train e split saranno i seguenti: Il training test conterrà le osservazioni dal 1983 fino al 1989, mentre il test set conterrà le osservazioni dal 1990 al 1991.

Nel grafico, i vari metodi di previsione vengono confrontati e, considerando la linea nera come il test set, emerge una netta differenza tra di essi, soprattutto per quanto riguarda la componente stagionale. Infatti, solo il metodo Seasonal Naive sembra essere in grado di cogliere questa componente, mentre il Drift method sembra essere il metodo che consente di ottenere le previsioni leggermente migliori. Tuttavia, nessuno dei quattro modelli riesce a prevedere con precisione le osservazioni della serie storica in esame:

- il Drift method, ad esempio, riesce a catturare il trend, ma non tiene conto della stagionalità;
- il Seasonal naive, invece, è in grado di catturare la componente di stagionalità ma non riesce a cogliere l'andamento del trend;
- Il naive individua il livello di andamento della serie;
- l'Average method non coglie né il trend e né la stagionalità.

Previsioni con modelli benchmark



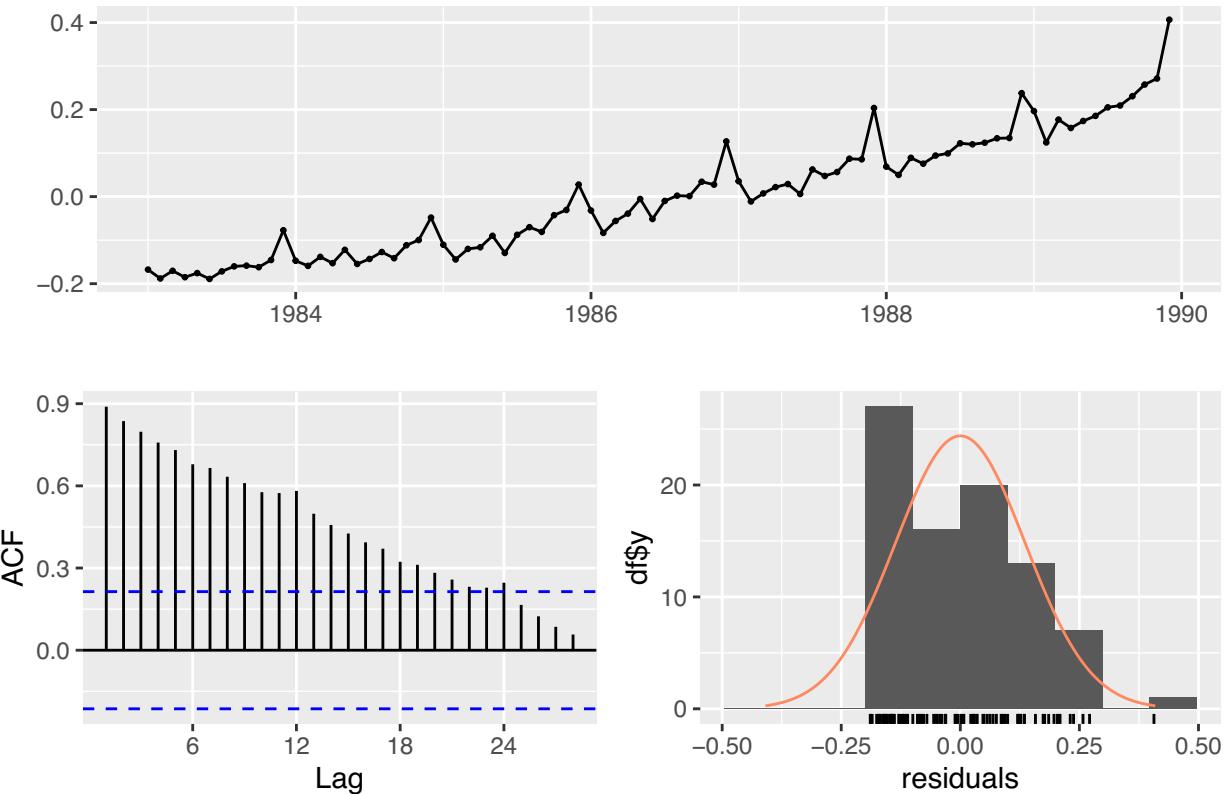
Residual Diagnostics

I residui rappresentano l'errore di previsione del modello statistico utilizzato per prevedere i valori di una serie storica al tempo t . Essi sono calcolati come la differenza tra il valore osservato al tempo t e il valore stimato dal modello al tempo t .

Nel contesto delle serie storiche, l'analisi dei residui è particolarmente importante perché permette di valutare l'accuratezza delle previsioni e la bontà del modello utilizzato. In particolare, l'analisi dei residui consente di:

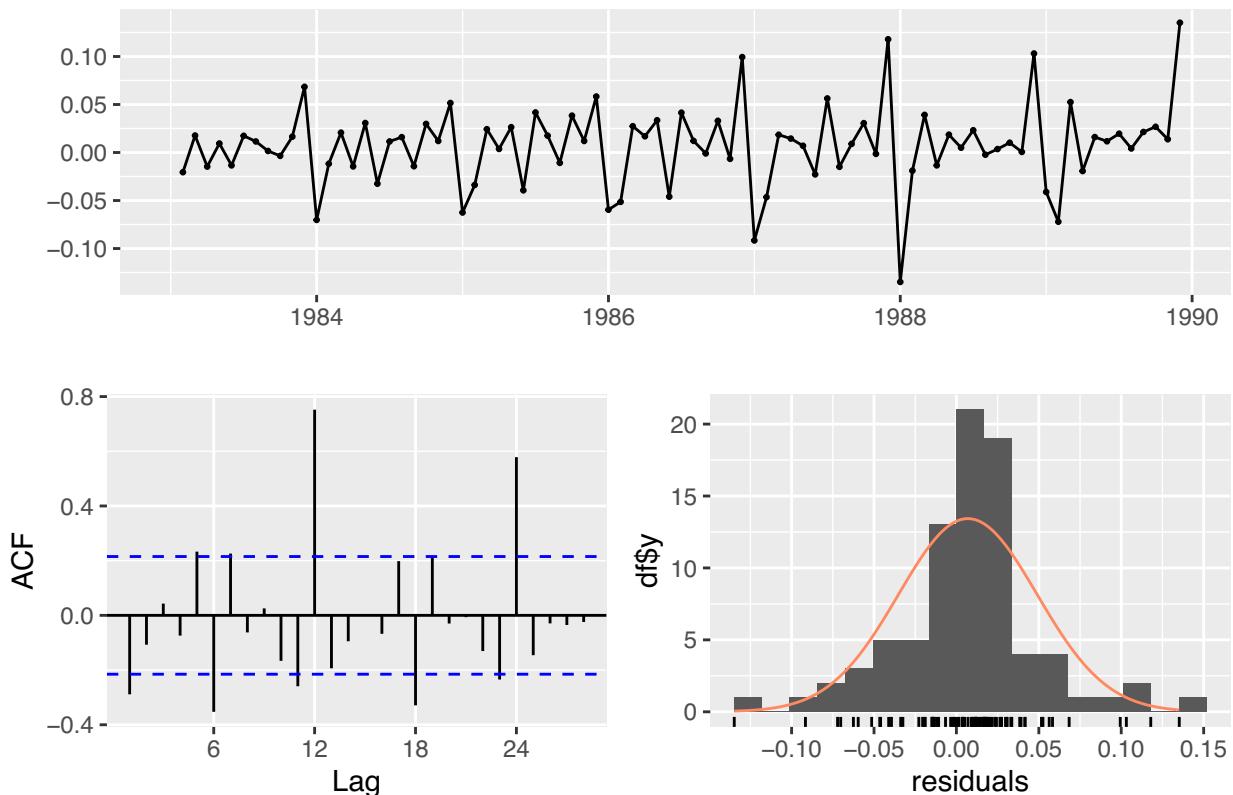
- 1. Verificare l'ipotesi di stazionarietà della serie storica: se i residui sono casuali e non presentano alcun tipo di pattern o tendenza, ciò indica che la serie è stazionaria.
- 2. Identificare eventuali pattern o tendenze nei residui: se i residui mostrano pattern o tendenze, ciò suggerisce che il modello utilizzato non tiene conto di tutte le variazioni presenti nella serie storica e che pertanto è necessario introdurre ulteriori variabili esplicative o considerare modelli più complessi.
- 3. Valutare l'effetto di eventuali outlier o dati anomali: se i residui presentano valori molto elevati o molto bassi, ciò può indicare la presenza di dati anomali o outlier nella serie storica, che potrebbero influenzare in modo significativo le previsioni del modello.

Residuals from Mean



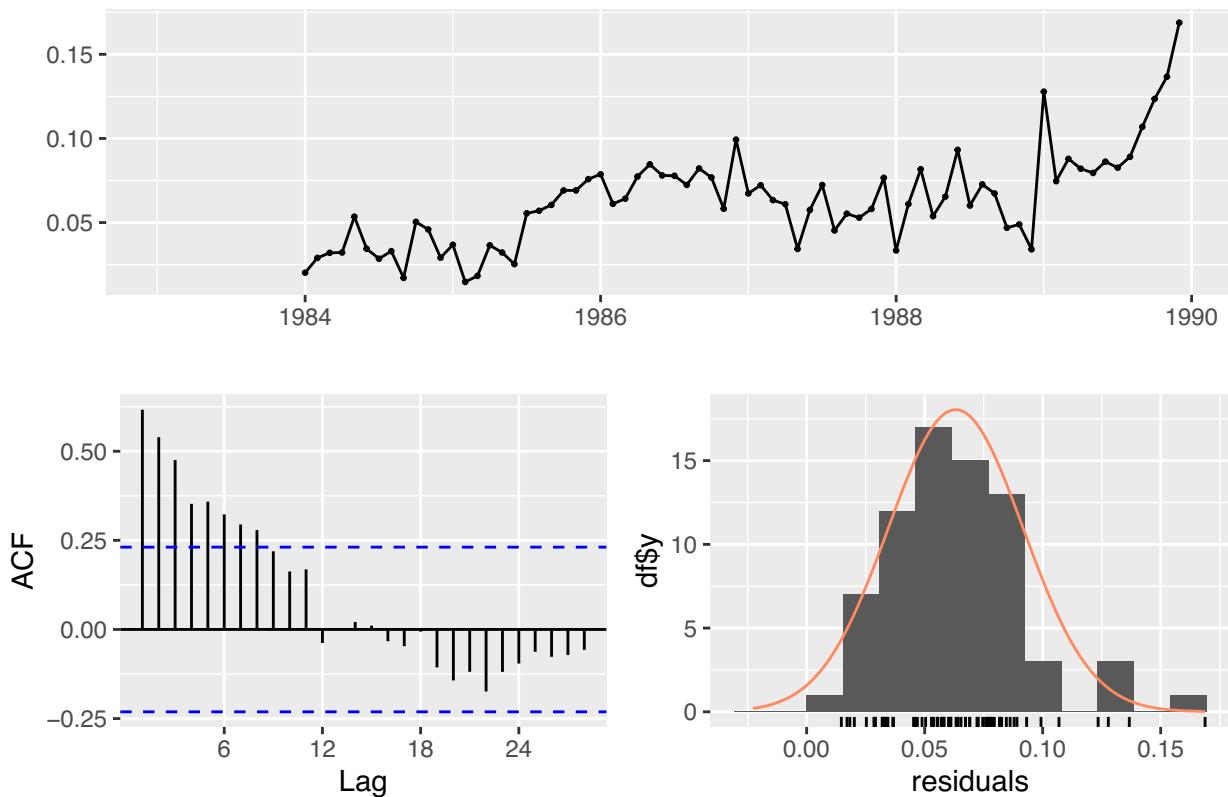
```
##  
## Ljung-Box test  
##  
## data: Residuals from Mean  
## Q* = 642.2, df = 17, p-value < 2.2e-16  
##  
## Model df: 0. Total lags used: 17
```

Residuals from Naive method



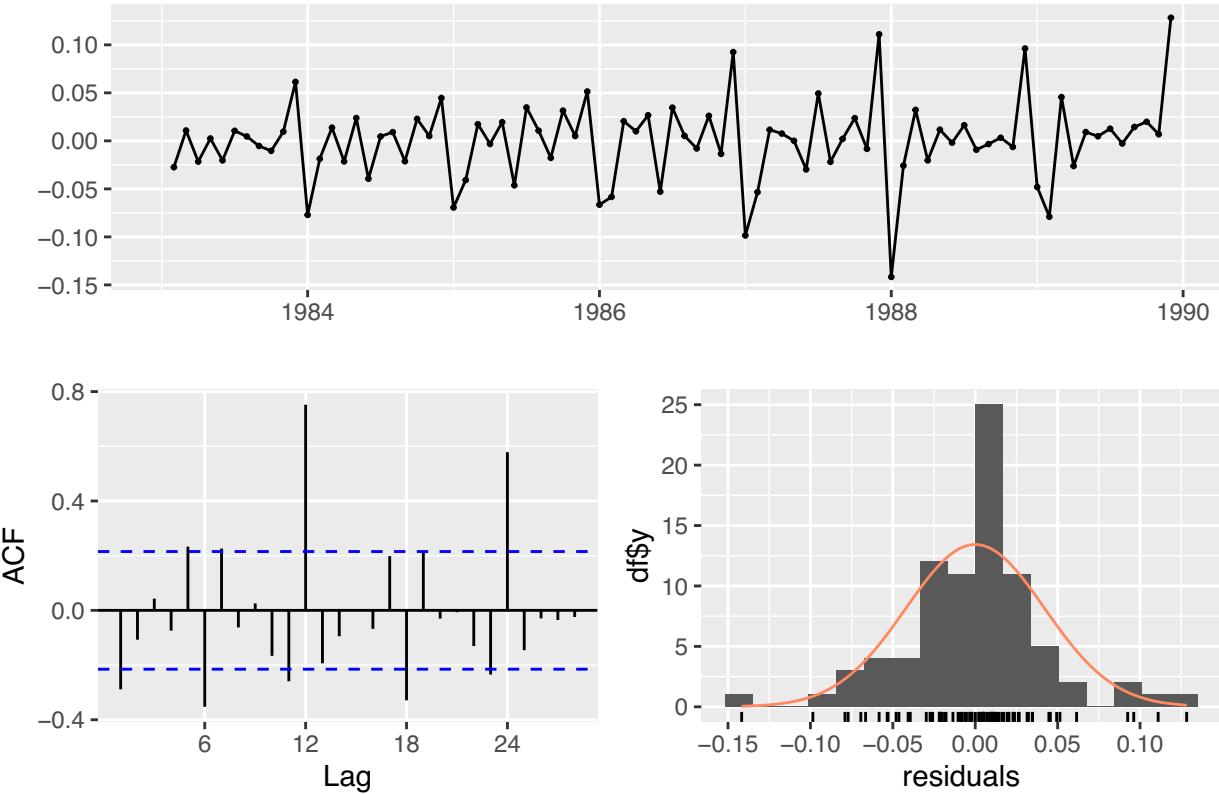
```
##  
## Ljung-Box test  
##  
## data: Residuals from Naive method  
## Q* = 105.1, df = 17, p-value = 9.992e-15  
##  
## Model df: 0. Total lags used: 17
```

Residuals from Seasonal naive method



```
##  
## Ljung-Box test  
##  
## data: Residuals from Seasonal naive method  
## Q* = 119.36, df = 17, p-value < 2.2e-16  
##  
## Model df: 0. Total lags used: 17
```

Residuals from Random walk with drift



```
##
## Ljung-Box test
##
## data: Residuals from Random walk with drift
## Q* = 105.1, df = 17, p-value = 9.992e-15
##
## Model df: 0. Total lags used: 17
```

L'analisi dei residua evidenzia che:

- Per l' **average method** il timeplot dei residui è decisamente non stazionario in media oltre che caratterizzato da una forte presenza di trend. La distribuzione non è normale, piuttosto è caratterizzata da una forte asimmetria positiva che implica una massiccia presenza di osservazioni sulla coda destra. Nessuna delle correlazioni risulta essere presente all'interno delle bande di confidenza, il valore di autocorrelazione più elevato si presenta al tempo 1 a cui successivamente si assiste un calo lento verso 0 con valori massimi nei multipli dei 12esimi mesi. Il Ljung-Box test permette di accettare l'ipotesi di correlazione tra i residui.
- Per il **naive method** il timeplot dei residui mostrerebbe un 'andamento costante inizialmente a cui succede una crescente variabilità. La distribuzione dei residui evidenzia un'andamento leptocurtico oltre che la presenza di valori anomali che fiancheggia la distribuzione verso il termine delle code. Nell'ACF alcuni ritardi rientrano all'interno delle bande di confidenza, altri invece sono fuori dalle bande. In particolar modo si ripetono sistematicamente gli andamenti al ritardi 1 e 6 negativi e 12 assume il valore massimo positivo. I valori osservati della serie storica sono correlati ai valori precedenti a questi ritardi specifici per cui potrebbe suggerire l'esistenza di una qualche struttura temporale nella serie storica.

- Per il **seasonal naive method** le autocorrelazioni sono molto forti e positive fino al ritardo 10, poi negative fino al ritardo 20, poi di nuovo positive. È quindi possibile individuare una struttura non lineare dei residui. La distribuzione potrebbe essere approssimativamente normale se non fosse per una concentrazione di valori nei pressi del valore 0.1.
- Per il **random walk with drift** il timeplot identifica un'andamento caratterizzato da incremento di variabilità con il passare del tempo, rispetto al metodo naive, la distribuzione risulta essere più leptocurtica.

Ovviamente, nessuno dei residui analizzati mostra un'andamento White noise.

Valutazione dell'accuratezza

Dalla valutazione del fitting sul training set emerge che fra i metodi il migliore è il Random Walk con Drift perchè ritorna un valore di fit più basso.

```
##                                     RMSE        MAE        MPE        MAPE
## Average                  0.13556635 0.11447840 -6.4051407 22.563044
## Naive                     0.04242624 0.03066386  0.8589356  5.691446
## Seasonal Naive           0.06925295 0.06315139 10.9307316 10.930732
## Random Walk con Drift  0.04185860 0.02929656 -0.5076156  5.484218
```

Valutando l'accuratezza previsiva sul test set si conferma l'impressione grafica: il miglior modello previsivo tra i modelli benchmark presi in esame è il Seasonal Naive in quanto presenta i valori più bassi per RMSE, MAE e MAPE. Quel che sorprende maggiormente, focalizzando l'attenzione sull'output della tabella e il plot precedentemente analizzato, l'accuratezza previsiva del metodo Naive e del Seasonal Naive è molto simile.

```
##                                     RMSE        MAE        MPE        MAPE
## Average                  0.3175455 0.31095714 36.37360 36.37360
## Naive                     0.1151982 0.10552500 -11.86388 12.82675
## Seasonal Naive           0.1005573 0.09465833 11.20340 11.20340
## Random Walk con Drift  0.1901035 0.18201084 -21.90710 21.90710
```

4. Modelli ML

4.1 Test di linearità

Molte serie storiche non possono essere descritte adeguatamente da modelli lineari a causa della loro natura complessa e non lineare. Per valutare la linearità di una serie storica, si può utilizzare il test di Terasvirta e il test di White. Questi test pongono come ipotesi nulla che la DGP è lineare, mentre l'ipotesi alternativa è che la DGP non è lineare. I risultati di entrambi i test per mostrano un p-value inferiore alla soglia $\alpha = 0.05$. Ciò fornisce una solida evidenza empirica contro l'ipotesi nulla, costringendo a rifiutarla. Pertanto, è possibile concludere che in media expenditure monthly non è lineare.

Alla luce di questi risultati, ci si propone di implementare reti neurali auto-regressive single layer per modellare la serie storica.

```
##
##  Teraesvirta Neural Network Test
##
## data: auscafe
## X-squared = 10.823, df = 2, p-value = 0.004464
```

```

## 
##  White Neural Network Test
## 
## data: auscafe
## X-squared = 8.7567, df = 2, p-value = 0.01255

```

4.2 Implementazione dei modelli

Verranno implementati sia modelli di regressione lineari autoregressivi e multipli che reti neurali. La motivazione alla base è la possibilità di comprendere quanto effettivamente le reti sono in grado di ottenere buone previsioni rispetto ai modelli più semplici (benchmark) e lineari (multipli lineari)

4.2.1 Modello con trend e stagionalità.

Poichè la serie storica risulta essere fortemente contraddistinta da trend e stagionalità è possibile stimare un modello in cui gli unici due regressori sono proprio le due componenti. La possibile applicazione vale esplicitamente e espressamente per le serie storiche con andamento prettamente crescente o decrescente, mentre non rimane valida per serie che mostrano cambi di direzione.

Il primo modello specificato comprende solo il trend. Il trend è statisticamente significativo ,a un livello di alpha pari a 0.05, e la sua pendenza è 0.005. Per ogni anno che passa il valore di spesa per mangiare fuori in Australia aumenta di 0.005 bilioni.

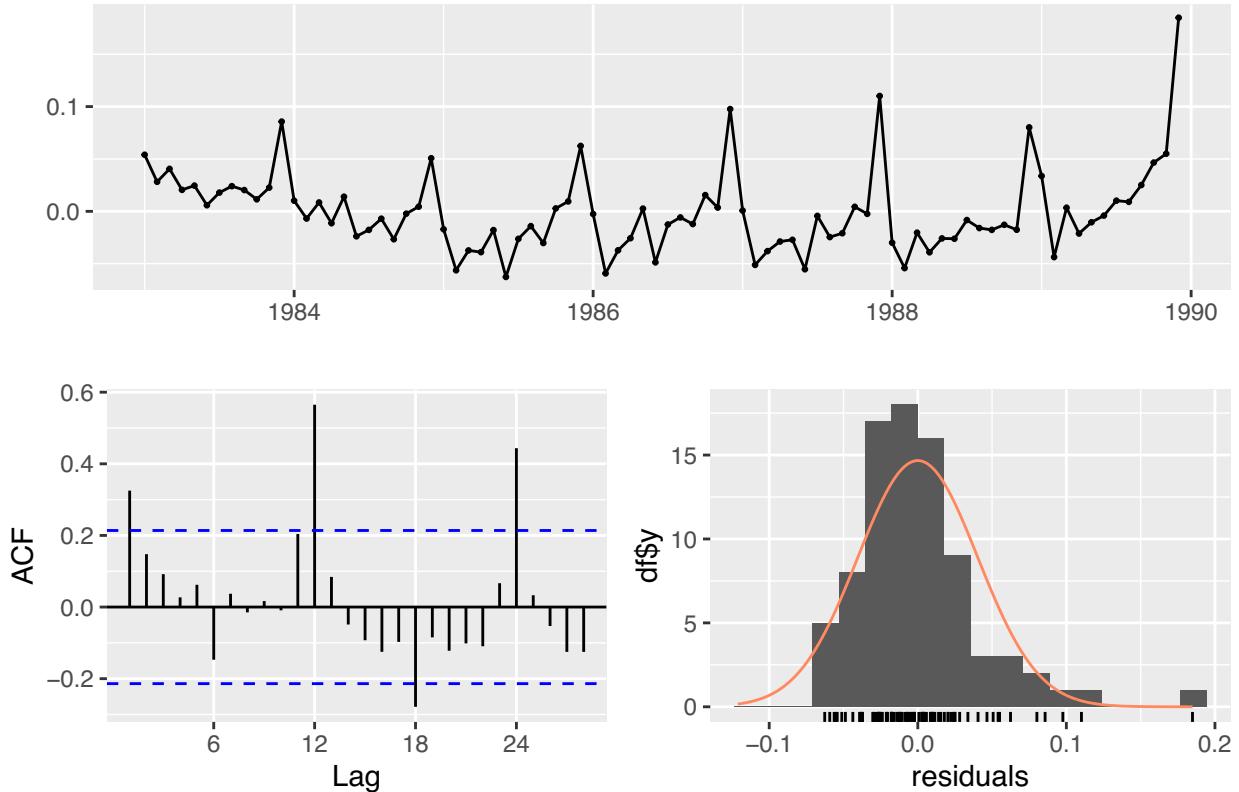
```

## 
## Call:
## tslm(formula = aus.ts_train ~ trend)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.062640 -0.025694 -0.006349  0.014367  0.184886
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.3092318  0.0089468  34.56   <2e-16 ***
## trend       0.0053403  0.0001828   29.21   <2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.04063 on 82 degrees of freedom
## Multiple R-squared:  0.9123, Adjusted R-squared:  0.9112 
## F-statistic:  853 on 1 and 82 DF,  p-value: < 2.2e-16

```

Effettuando l'analisi dei residui si evince che il modello così stimato viola l'ipotesi di autocorrelazione nei residui. Le autocorrelazioni sono molto forti e positive e vanno a 0 molto lentamente. La struttura con trend decrescente mostra picchi maggiori nel dodicesimo mese. Anche la distribuzione non è per nulla normale, piuttosto da segni di una bimodalità. Il fatto che il trend sia statisticamente positivo ma poi non spiega la struttura dei dati insegnna che non bisogna letteralmente affidarsi in maniera completa alla lettura dei p-value.

Residuals from Linear regression model



```
##  
## Breusch-Godfrey test for serial correlation of order up to 17  
##  
## data: Residuals from Linear regression model  
## LM test = 60.328, df = 17, p-value = 9.276e-07
```

Per questo motivo, il secondo modello considerato contiene sia trend che stagionalità, con l'obiettivo di creare un modello in grado di catturare entrambe le componenti. Le dummy individuate fanno riferimento ai diversi mesi dell'anno: l'informazione contenuta in *yt* fa riferimento all'osservazione appartenente al primo mese dell'anno. Dal summary del modello appena implementato si deduce che:

- Non tutte le variabili sono statisticamente significative, considerando un livello di alpha pari a 0.10 le variabili stat. significative sono l'intercetta, il trend, season2, season 6 e season 12; tuttavia, la non significatività di una variabile non ha grande influenza nel campo delle previsioni.
- C'è una tendenza media crescente di 0.0073 bilioni di dollari per mese, per cui come era stato già evidenziato precedentemente è presente trend crescente nella time serie; Ricordando che è stata evidenziato stagionalità in febbraio e dicembre:
 - In media, a Febbraio c'è una spesa inferiore di 0.04 bilioni di dollari rispetto a Gennaio;
 - Giugno ha una spesa di 0.037 bilioni di dollari inferiore rispetto al primo mese;
 - Dicembre ha una spesa di 0.10 bilioni di dollari superiore al primo mese, il suo coefficiente associato è il più elevato in assoluto.

Valutando l'Adjusted R-squared, rappresentante della quota parte di variabilità della forecast variabile catturata dal modello, assume un valore molto alto pari a 0.96. Ad ogni modo l'indice R^2 non è un indice che ha rilievo nella valutazione della performance previsiva del modello.

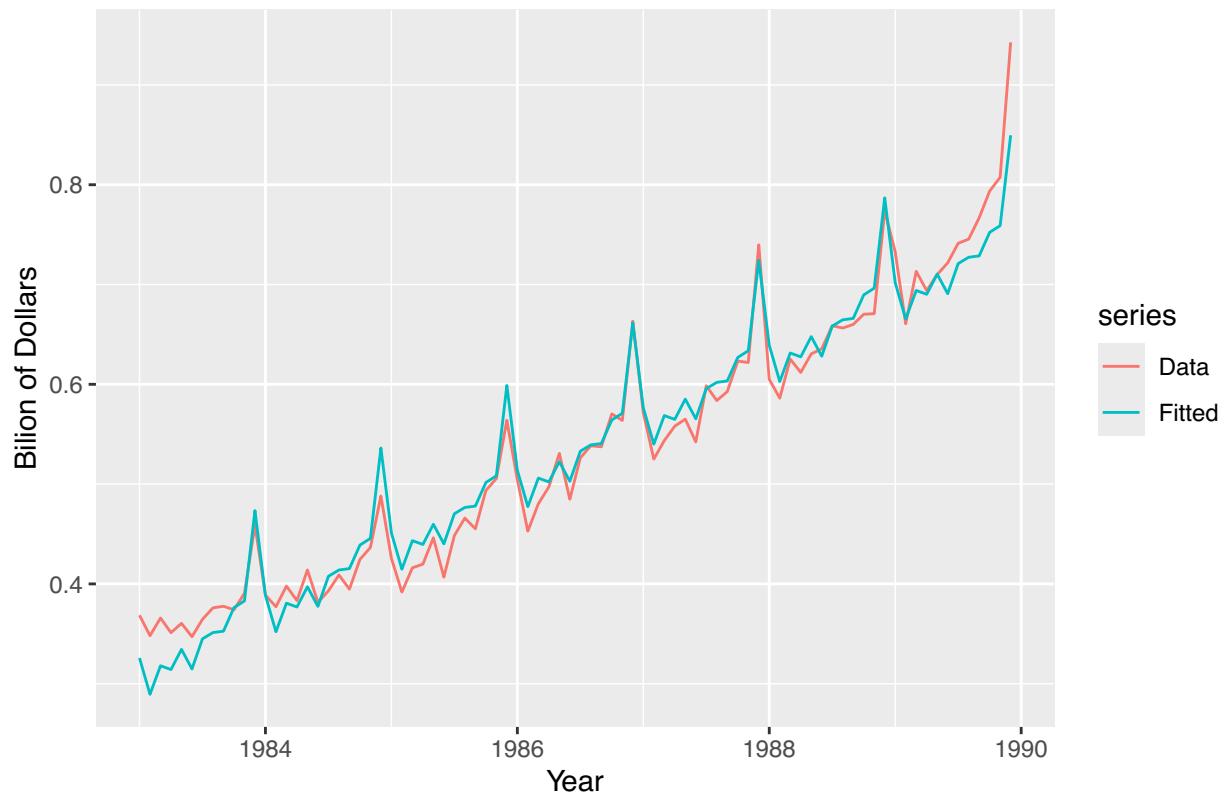
```

## 
## Call:
## tslm(formula = aus.ts_train ~ trend + season)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.048061 -0.016814 -0.004959  0.015934  0.093155 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.3205609  0.0107962 29.692 < 2e-16 ***
## trend       0.0052231  0.0001186 44.025 < 2e-16 ***
## season2     -0.0416374  0.0139496 -2.985 0.00389 **  
## season3     -0.0182461  0.0139511 -1.308 0.19514    
## season4     -0.0272978  0.0139537 -1.956 0.05436 .    
## season5     -0.0122637  0.0139572 -0.879 0.38255    
## season6     -0.0371153  0.0139617 -2.658 0.00970 **  
## season7     -0.0121670  0.0139673 -0.871 0.38663    
## season8     -0.0110900  0.0139738 -0.794 0.43006    
## season9     -0.0149560  0.0139814 -1.070 0.28837    
## season10    0.0034353  0.0139899  0.246 0.80674    
## season11    0.0049122  0.0139995  0.351 0.72671    
## season12    0.0902463  0.0140100  6.442 1.22e-08 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.0261 on 71 degrees of freedom
## Multiple R-squared:  0.9687, Adjusted R-squared:  0.9634 
## F-statistic: 183 on 12 and 71 DF, p-value: < 2.2e-16

```

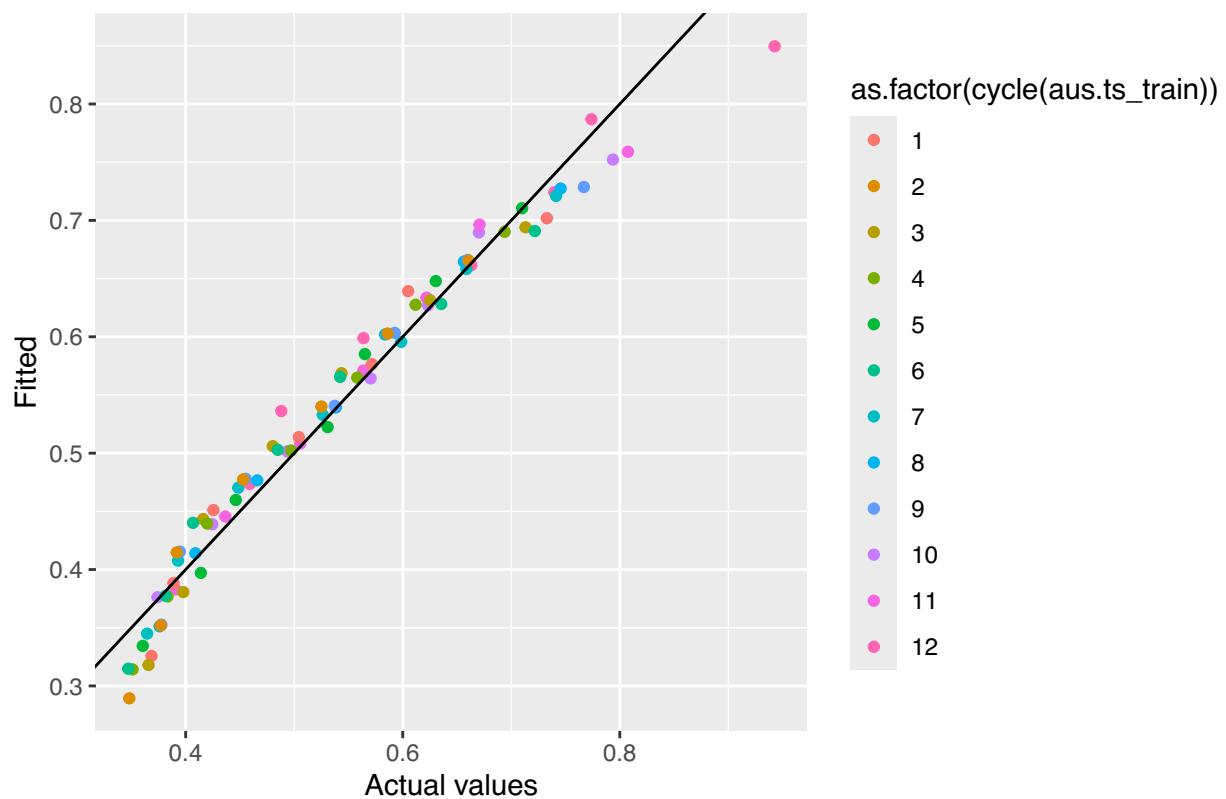
Il modello stimato cattura gran parte della stagionalità e trend, tuttavia sovrastima i dati per l'arco temporale che va dal 1984 al 1986 e sottostima dal 1988 in poi.

Monthly expenditure in eating out



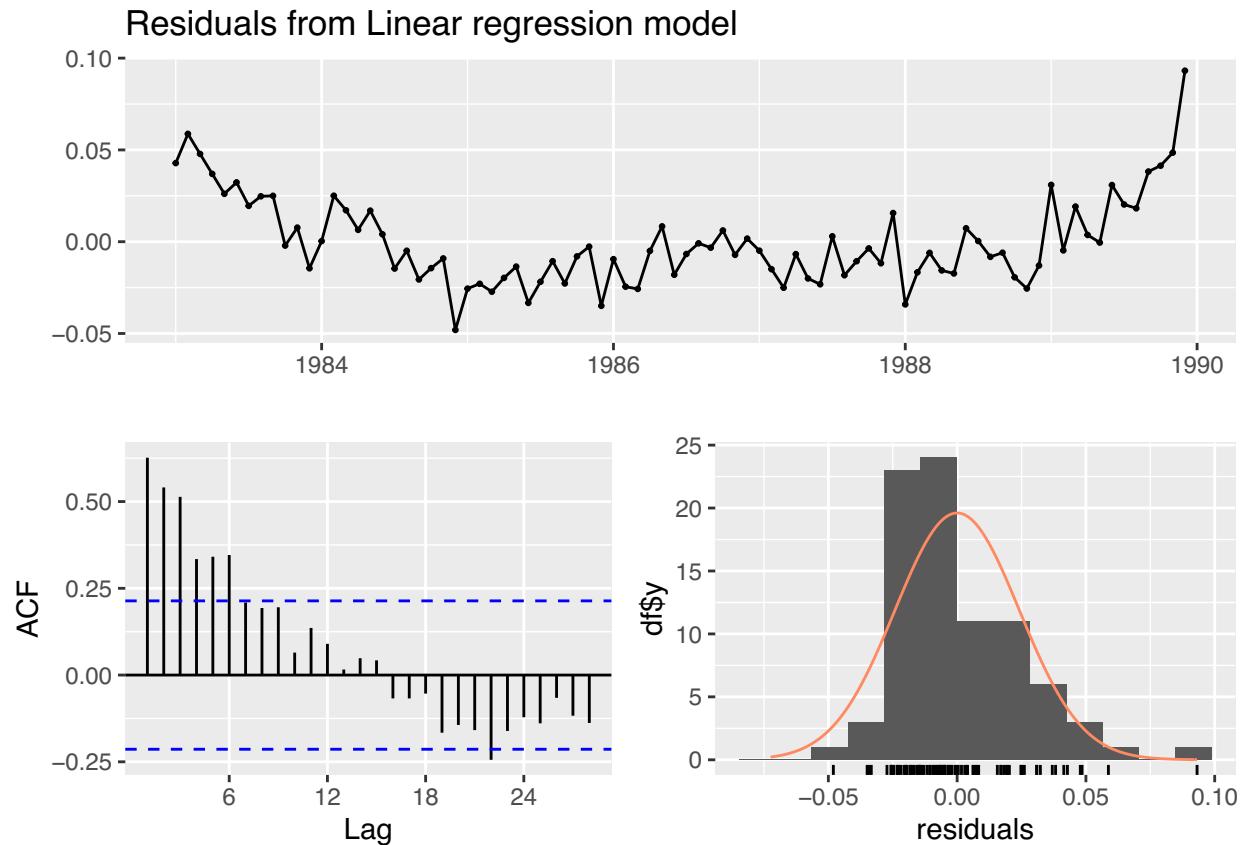
Per la verifica dell'adattamento del modello ai dati è possibile analizzare , tramite un confronto grafico, i valori reali e quelli stimati. I punti sono abbastanza vicini ai reali per certi versi, ma molto distanti per altri. I punti sono stati colorati per ogni mese per comprenderne l'effetto, tuttavia, l'elevata numerosità non rende chiara la distinzione.

Monthly expenditure in eating out: fittes vs real



Analisi dei residui del modello con trend e stagionalità

Valutando il test e i grafici implementati sui residui, questi non sono White Noise. Il test implementato è quello di Breusch-Godfrey per i modelli di regressione per testare se i residui risultano essere incorrelati. Come prima, i residui non mostrano un andamento White Noise: sono correlati, hanno distribuzione diversa da quella normale e non sembrano avere una stazionarietà in media. Il modello con solo trend e stagionalità non è per nulla ottimo.



```
##
## Breusch-Godfrey test for serial correlation of order up to 17
##
## data: Residuals from Linear regression model
## LM test = 55.231, df = 17, p-value = 6.291e-06
```

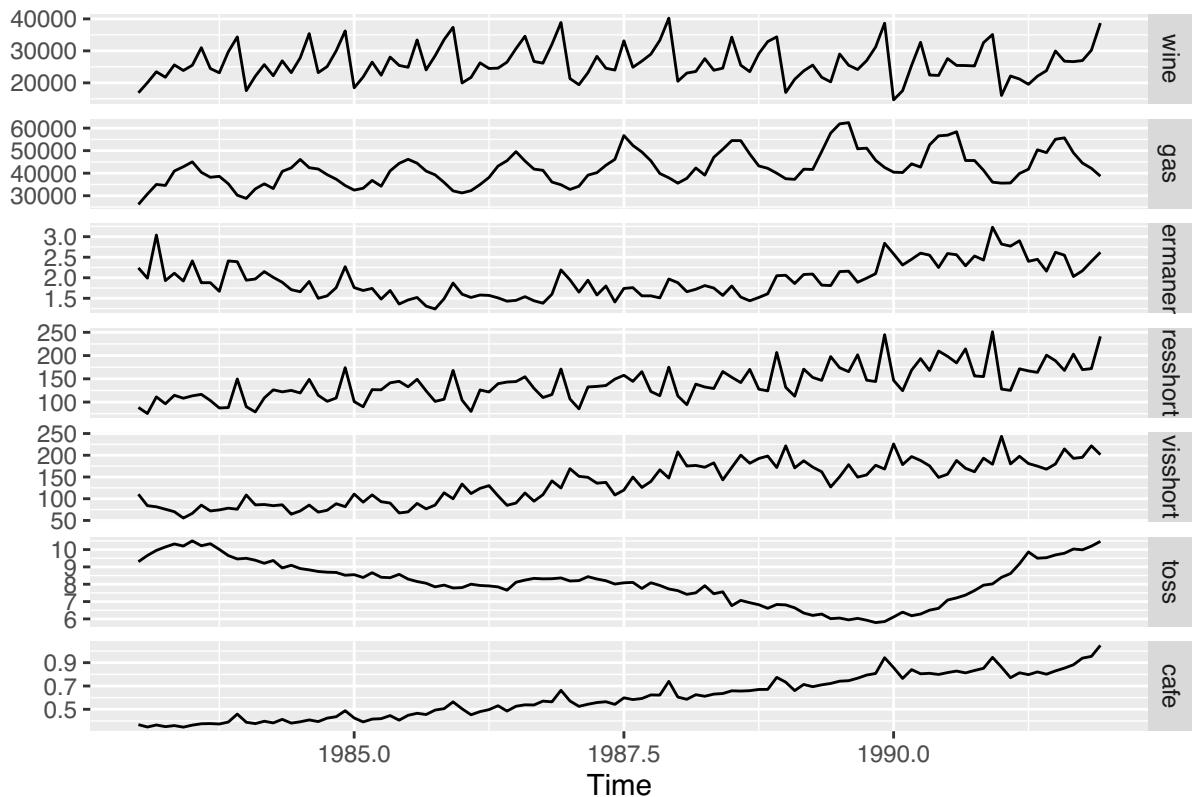
4.2.2 Modello di regressione con altre variabili

Il modello di regressione multipla permette di tener conto delle relazioni presenti tra la variabile dipendente e quelle indipendenti. In particolar modo indagare le associazioni e l'influenza delle covariate sul valore medio della variabile di risposta. In questo lavoro verranno osservate le relazioni tra variabili attraverso l'analisi di un'aggregazione di serie storiche composte dalle seguenti variabili:

- **GAS**: produzione mensile di gas in Australia dal Gennaio 1956 -Agosto 1995;
- **WINE**: vendite totali di vino . Gennaio 1980 - Agosto 1994.
- **PERMANENT**: partenze permanenti dall'Australia, Gennaio 1976-Novembre 2016.
- **RESSHORT**: partenze di residenti a breve termine dall'Australia, Gennaio 1976-Novembre 2016.
- **VISSHORT**: partenze di visitatori dall'Australia, Gennaio 1976-Novembre 2016.
- **TASSOCC**: tasso di disoccupazione in Australia, Febbraio 1978- Gennaio 2023. Questo il sito di riferimento <https://fred.stlouisfed.org/series/LRUNTTTAUM156S>

Poichè le osservazioni provengono da serie storiche con tempi diversi è stato individuato un arco temporale comune a tutte. L'arco temporale ricopre la distanza di 10 anni tra il 1983 e il 1991.

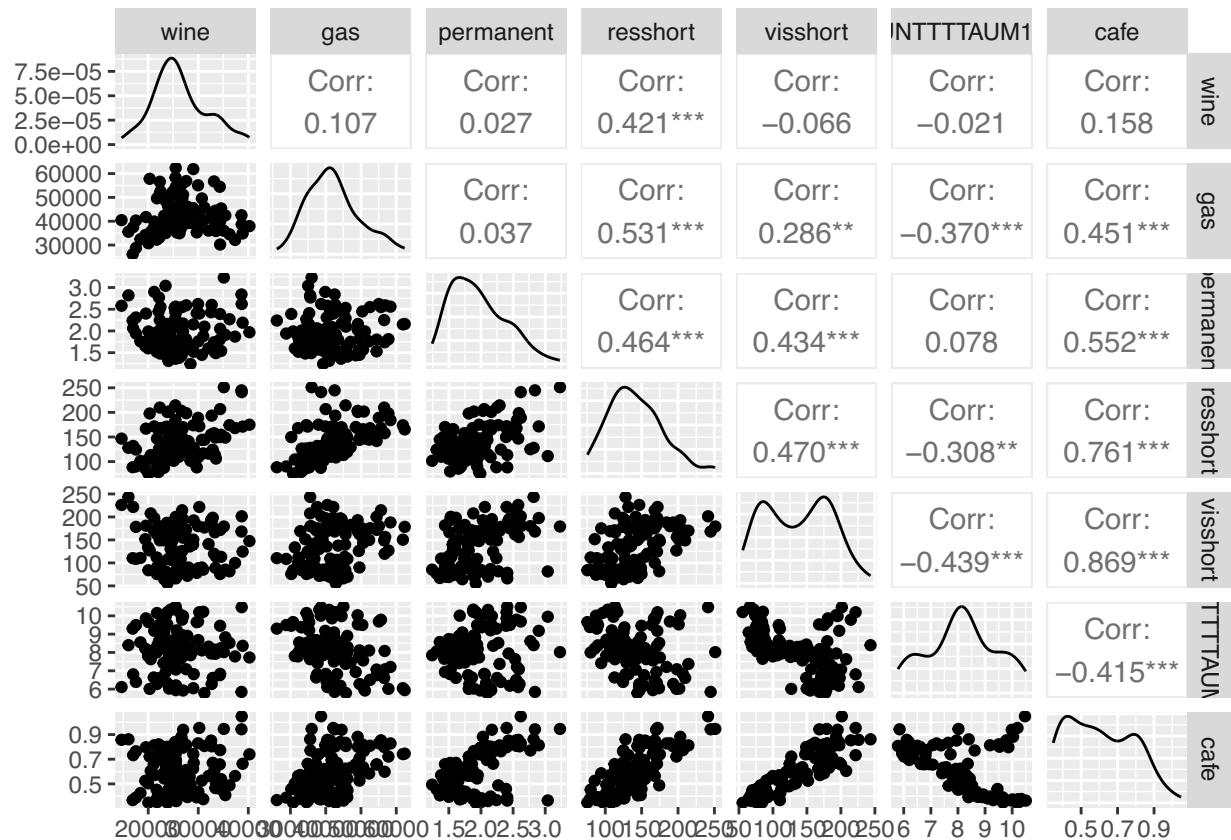
Il plot della serie multivariata è un grafico che rappresenta l'andamento delle variabili nel tempo. In questo tipo di grafico, ogni variabile della serie è rappresentata su un asse (l'asse y), mentre il tempo è rappresentato sull'altro asse (l'asse x). Le diverse variabili della serie sono rappresentate come linee, ognuna delle quali mostra l'andamento della variabile specifica nel tempo. Il grafico è utile per analizzare e confrontare l'andamento delle diverse variabili della serie, e per individuare eventuali relazioni tra di esse. Ad esempio, **vissshort** e **resshort** mostrano un andamento simile nel tempo, ciò può indicare una possibile correlazione tra di esse. Al contrario, **wine** e **cafe** mostrano un andamento opposto nel tempo, ciò può indicare una relazione inversa tra di esse.



Un'ulteriore analisi preliminare alla costruzione del modello di regressione è la rappresentazione, tramite una matrice, delle relazioni tra le variabili. In particolare, la diagonale principale è rappresentata dalla visualizzazione della distribuzione della variabile in questione, da questo si evidenzia come le variabili **wine** e **gas** abbiano una distribuzione approssimativamente normale con code destre più pesanti, ovvero asimmetria positiva e forma leptocurtica. Diversamente **permanent** e **resshort** mostrano una distribuzione per nulla normale con presenza massiccia di dati intorno alla media; le restanti tre variabili (**vissshort**, **cafe** e **toss**) presentano un andamento bimodale, la bimodalità è probabilmente dovuta al fattore di stagionalità evidente (in seguito un approfondimento su queste variabili). La parte destra superiore alla diagonale principale, riporta i valori delle correlazioni. La variabile **cafe**, che sarà dipendente nel modello, risulta essere fortemente correlata positivamente con Vissshort, Resshort permanent e gas; debolmente correlata positivamente con toss e wine. La forza di tali correlazioni (tranne che per wine) è mostrata dal coefficiente che si trova nella prima riga: il coefficiente risulta essere statisticamente diverso da 0 in quanto gli asterischi rappresentano un rifiuto dell'ipotesi nulla d'incorrelazione tra le serie in esame.

Un problema comune con le serie storiche è quello di fare l'errore di implementare regressioni spurie: ovvero, mettere in relazione delle serie storiche che risultano essere correlate ma che in realtà misurano fenomeni completamente differenti. Non è questo il caso in quanto le variabili sono state scelte con cura: la presenza di turisti potrebbe far aumentare la spesa totale di cene a pranzi fuori così come l'aumento del tasso di

disoccupazione potrebbe far diminuire il numero di persone che cenano fuori casa.

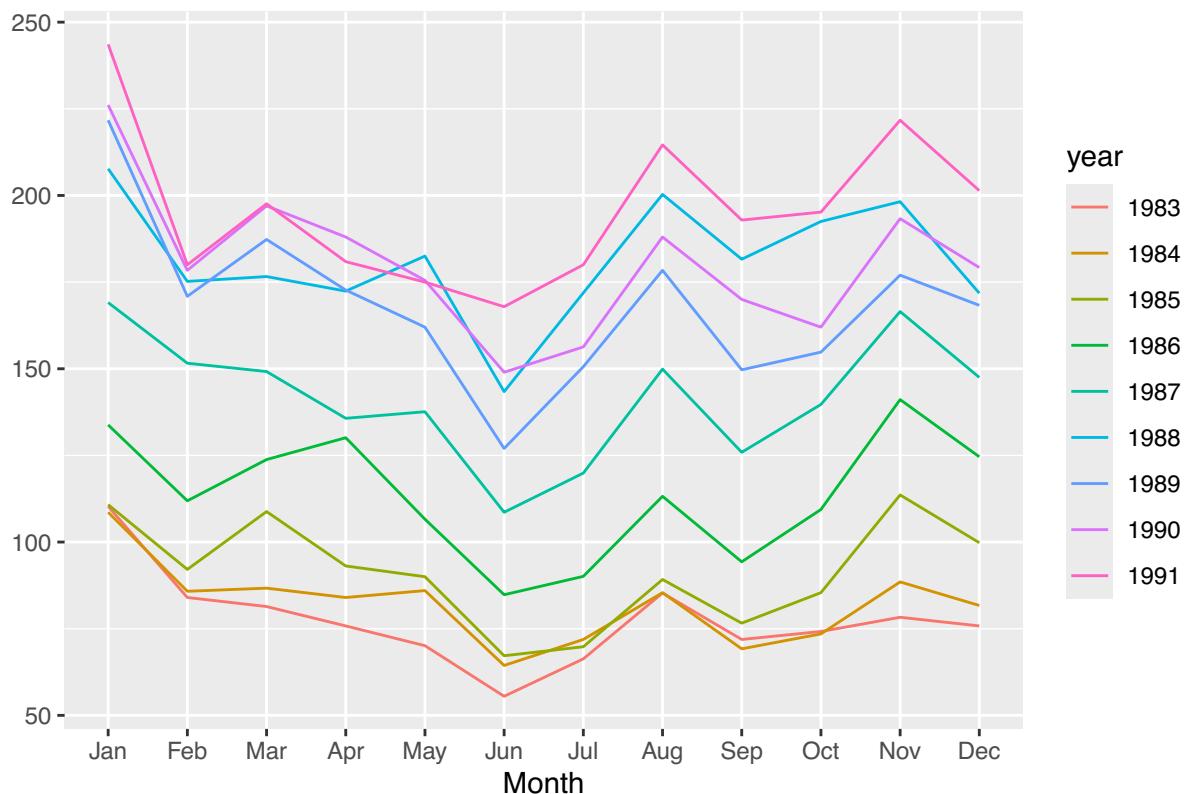


Tuttavia ci si potrebbe imbattere nella situazione durante la quale le variabili potrebbero avere un coefficiente di correlazione statisticamente significativo per il semplice motivo che potrebbero essere legate al tempo influendo sul valore di correlazione rendendola spuria. Per cui, per esplorare le relazioni tra le serie temporali e comprendere se le variazioni di una serie sono realmente correlate alle variazioni di un'altra serie è stato necessario rimuovere il trend. Sono state confrontate tutte le variabili detrendizzate rispetto alla variabile dipendente e stabilito un livello di alpha pari a 0.05, considerando l'ipotesi nulla che i valori delle coppie sono indipendenti, contro l'ipotesi alternativa H1 i valori sono legati da una relazione lineare, tutte risultano essere statisticamente correlate ad eccezione della variabile visshort.

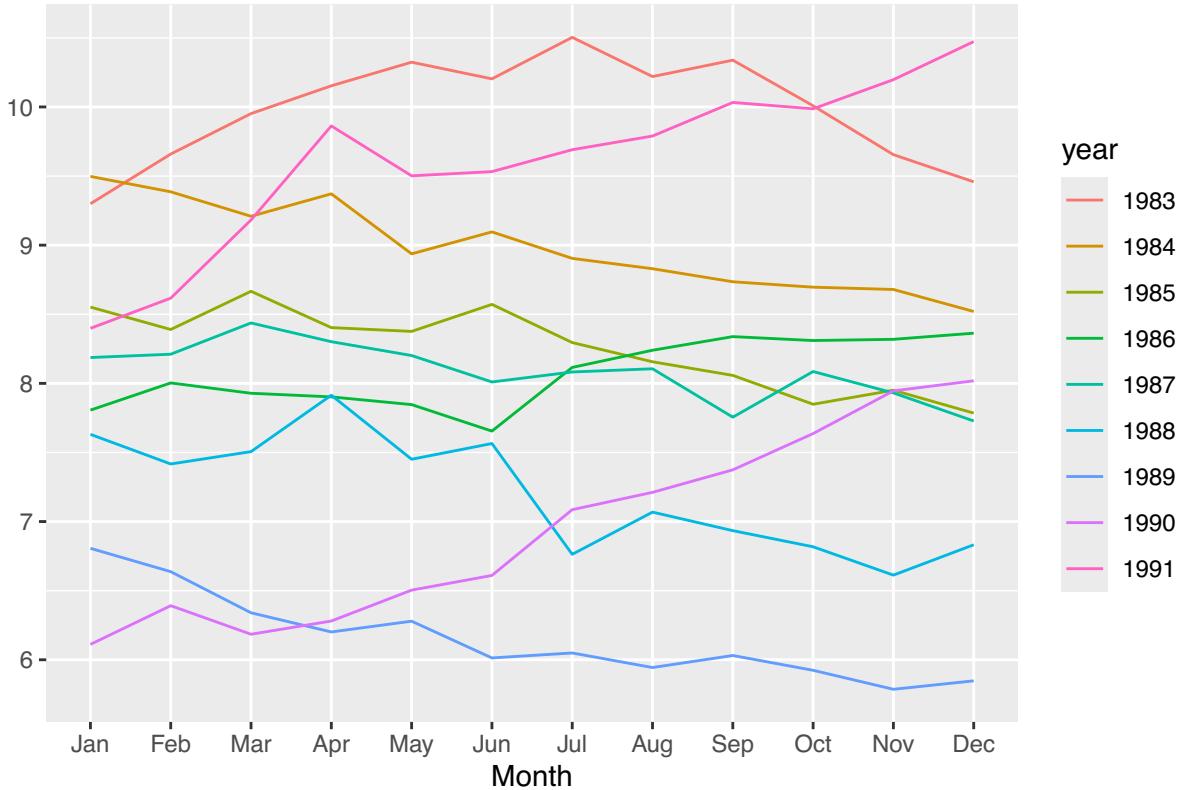
Analisi sulla distribuzione bimodale di toss e resshort

- **Visshort** evidenzia stagionalità con ripetizioni annue di decrementi a febbraio giugno e settembre incrementi nel mese di agosto;
- **Toss** non sembra avere un'evidente stagionalità.

Seasonal plot: aus_mult[, 5]



Seasonal plot: aus_mult[, 6]



La variabile da prevedere è quella di risposta rappresentante le spese totali mensili per mangiare fuori in Australia. Effettuare un modello di regressione con i predittori precedenti permette di ottenere delle previsioni più accurate. Il motivo può essere ricercato nel fatto che il numero di visitatori, il tasso di occupazione potrebbero influire sulle spese totali di consumi in cene e pranzi.

È possibile dividere la serie multivariata **in training set e test set**. Il training set contiene le informazioni da gennaio 1983 al dicembre 1989. Il test set , invece, dal 1990 in poi.

Stima dei modelli

Vengono considerati diversi modelli e i relativi residui per comprendere quale modello, tra quelli lineare, sia in grado di spiegare meglio la variabile dipendente.

Modello 1: Il primo modello considerato è quello contenente tutti i regressori presi in analisi. Il summary del modello permette di avere una colonna dedicata ai coefficienti stimati. Non tutte le variabili sono statisticamente significative. Fissato un livello alpha pari a 0.10:

- I mesi statisticamente significativi sono febbraio, marzo, aprile e giugno;
- Gas presenta un coefficiente stimato pari a 0.00000456, tenendo conto dell'effetto marginale della produzione di gas rispetto alle spese di mangiare fuori. A seguito di una variazione di produzione di gas ci si aspetta un'incremento delle spese di mangiare fuori del 0.00000456 in media;
- Permanent, rappresentante il numero di partenze dall'Australia ha un impatto positivo sulla variabile dipendente. Per ogni incremento unitario di partenze il valore speso per mangiare fuori aumenta di 0.059 bilioni di dollari.

- Resshort, rappresentante le partenze a breve termine di residenti in australia , ha anch'esso un impatto positivo, anche se minore rispetto a permanent. Per ogni incremento unitario di partenze il valore speso in cene e pranzi fuori aumenta di 0.000445;
- Visshort, rappresentante le partenze di visitatori a breve termine, comporta un incremento della variabile dipendente. Per ogni partenza il valore speso in pranzi e cene fuori aumenta di 0.0005136. I residui variano in un range che va da -0.05 a 0.04. Media e mediana approssimativamente coincidono il che potrebbe essere indice di una distribuzione simmetrica. Confrontando i residui di questo modello con quello relativo al solo trend e stagionalità in cui il campo di variazione è più ampio permette di comprendere che la scelta di un numero maggiore di regressori che effettivamente impattano sulla variabile dipendente offrono la possibilità di fare stime e previsioni vicini a quelli reali. L R^2 quadro corretto , con valore uguale a 0.98 è indice di un buon adattamento ai dati. La decisione di affidare i risultati all'r^2 corretto piuttosto che all'r^2 sta nel fatto che a differenza del normale R^2 corretto non esplode dal crescere delle variabili oltre che aumenta al aumentare dei regressori anche se le variabili non hanno nessun impatto, perchè i regressori vengono calcolati con l obiettivo di minimizzare gli errori e quindi tende verso 1. Un altro indice utile per la comprensione dell'adattamento ai dati è il residual standard error, questo pari a 0.01 implica che il modello potrebbe essere ottimo in quanto i valori mediamente si discostano poco dalla retta di regressione. Il test sull'insieme dei coefficienti, rappresentato dalla statistica Test F a cui si affianca il valore del p-value, in cui l ipotesi h0 equivale a considerare tutti i coefficienti del modello sono statisticamente uguali a 0 contro h1 in cui almeno un coefficiente è statisticamente diverso da 0, aiuta nella comprensione del modello. Se accettassimo h0, il modello con solo intercetta funziona meglio rispetto al modello con i regressori. In questo caso il p-value prossimo allo 0 indica che almeno uno dei regressori ha un impatto sulla variabile dipendente.

```
##
## Call:
## tslm(formula = cafe ~ trend + season + wine + gas + permanent +
##       resshort + visshort + toss, data = aus_mul_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.052566 -0.010883 -0.000580  0.009656  0.040684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.487e-01 7.053e-02 2.108  0.0389 *
## trend        4.795e-03 5.074e-04 9.450 8.05e-14 ***
## season2     -2.508e-02 1.287e-02 -1.949  0.0557 .
## season3     -4.000e-02 1.417e-02 -2.823  0.0063 **
## season4     -3.132e-02 1.447e-02 -2.165  0.0341 *
## season5     -2.219e-02 1.945e-02 -1.141  0.2581
## season6     -4.143e-02 2.436e-02 -1.700  0.0938 .
## season7     -3.113e-02 2.874e-02 -1.083  0.2827
## season8     -2.860e-02 2.622e-02 -1.091  0.2794
## season9     -1.495e-02 2.120e-02 -0.705  0.4831
## season10    1.749e-02 1.975e-02  0.885  0.3793
## season11    4.203e-03 1.879e-02  0.224  0.8237
## season12    3.219e-02 2.365e-02  1.361  0.1782
## wine         4.807e-07 9.850e-07  0.488  0.6272
## gas          4.560e-07 1.162e-06  0.392  0.6960
## permanent   5.900e-02 9.635e-03  6.124 5.93e-08 ***
## resshort    4.458e-04 2.390e-04  1.866  0.0666 .
## visshort    5.511e-05 1.749e-04  0.315  0.7537
## toss        -5.251e-04 5.933e-03 -0.089  0.9297
## ---
```

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01835 on 65 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9819
## F-statistic: 251.1 on 18 and 65 DF,  p-value: < 2.2e-16

```

Si procede alla stima di altri modelli (non vengono inseriti nel report).

Selezione del modello

Per tutti i modelli considerati verranno calcolati gli indici AIC, BIC, AICc, CV e ADJR2. Il modello che presenterà valori migliori per gli indici verà selezionato, analizzato nel dettaglio e confrontato rispetto ai modelli di rete neruale. Dal confronto emerge che il terzo modello (mod5) è il migliore in quanto è quello che minimizza AIC BIC e che massimizza AdjR2.

```

##          CV      AIC      AICc      BIC      AdjR2
## mod3 0.0004207073 -657.0984 -646.5753 -613.3437 0.9824102
## mod6 0.0004319699 -654.6098 -646.4904 -615.7168 0.9815484
## mod1 0.0004398457 -653.2266 -639.8933 -604.6103 0.9818966
## mod5 0.0004407692 -652.5054 -641.9823 -608.7507 0.9814216
## mod4 0.0004493845 -651.7862 -641.2631 -608.0315 0.9812619
## mod2 0.0004520609 -650.8448 -638.9698 -604.6593 0.9812162

```

Analizzando nello specifico, il modello implementato ha molte variabili statisticamente significative e molte non. Tuttavia, anche se alcune delle variabili come **gas**, e **wine** non sono statisticamente significative, la non significatività statistica nel campo previsivo non è molto importante.

- Per ogni incremento unitario di partenze di abitanti il valore speso in spese fuori casa aumenta di 0.0551 bilioni di dollari.
- Per ogni incremento unitario di gruppi di partenze di visitatori a breve termine il valore di spesa diminuisce di 0.000047 bilioni di dollari;
- Per ogni incremento unitario di consumo di gas i dollari spesi in cene e pranzi fuori aumenta di 4.347e-07.
- Per ogni incremento unitario di vendite di vino il valore speso in cene fuori casa aumenta di 4.698e-07.

```

##
## Call:
## tslm(formula = cafe ~ trend + season + gas + permanent + resshort +
##       wine, data = aus_mul_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.052481 -0.011163 -0.000300  0.009588  0.041866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.499e-01  3.761e-02   3.987 0.000168 ***
## trend        4.925e-03  2.369e-04  20.790 < 2e-16 ***
## season2     -2.693e-02  1.120e-02  -2.405 0.018961 *
## season3     -4.075e-02  1.351e-02  -3.016 0.003619 **
## season4     -3.255e-02  1.336e-02  -2.436 0.017524 *
## season5     -2.347e-02  1.838e-02  -1.277 0.206129

```

```

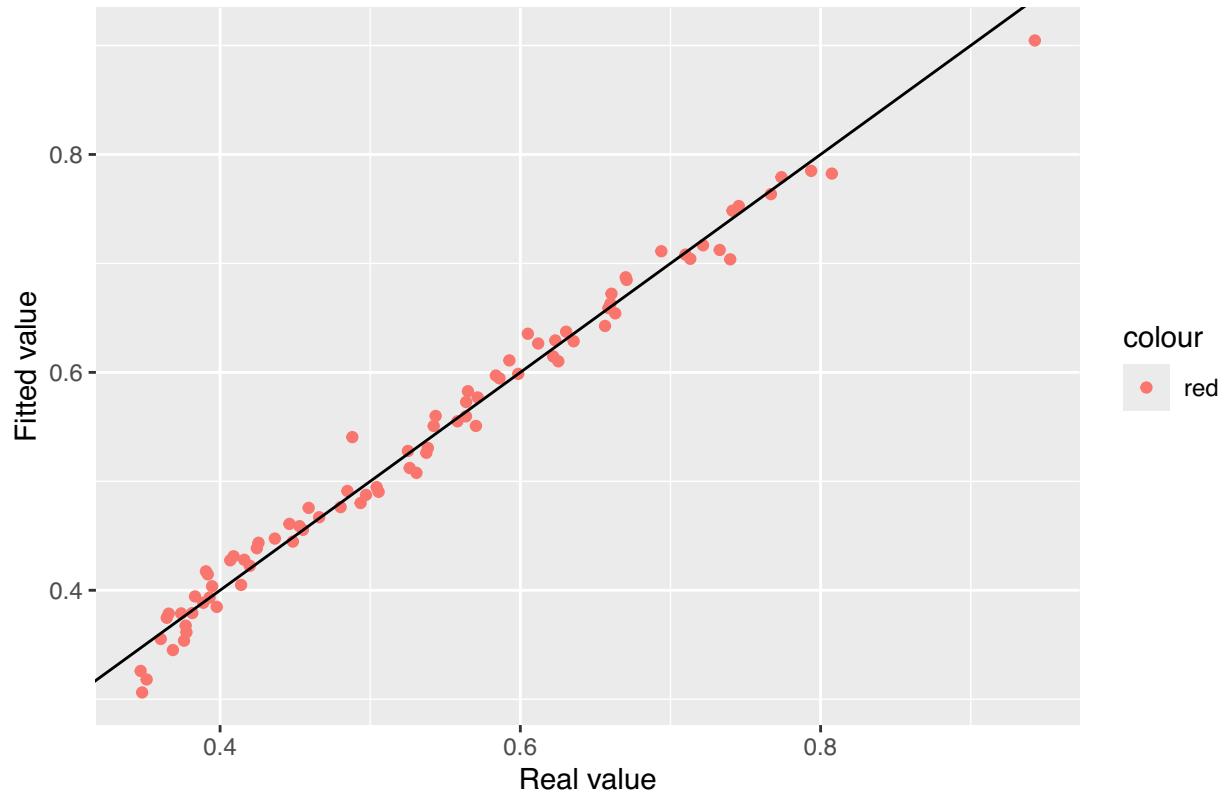
## season6      -4.374e-02  2.207e-02 -1.982 0.051624 .
## season7     -3.296e-02  2.738e-02 -1.204 0.232937
## season8     -2.920e-02  2.535e-02 -1.152 0.253527
## season9     -1.676e-02  1.943e-02 -0.863 0.391372
## season10    1.536e-02  1.824e-02  0.842 0.402659
## season11    2.913e-03  1.808e-02  0.161 0.872452
## season12    3.141e-02  2.274e-02  1.381 0.171852
## gas          4.374e-07  1.141e-06  0.383 0.702626
## permanent   5.962e-02  8.981e-03  6.638 6.69e-09 ***
## resshort    4.226e-04  1.999e-04  2.114 0.038249 *
## wine         4.698e-07  9.610e-07  0.489 0.626550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01809 on 67 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9824
## F-statistic: 290.7 on 16 and 67 DF,  p-value: < 2.2e-16

```

Confronto tra valori stimati e valori osservati

Un primo controllo sull'accuratezza del modello può essere svolto attraverso lo scatter plot dei valori osservati e stimati. Tra questi è presente una correlazione lineare positiva abbastanza forte, le stime sono abbastanza accurate; La maggior parte dei punti non sono direttamente sulla retta ma intorno per cui questo indica che il modello può essere sicuramente migliorato. L'ultimo valore potrebbe essere simbolo di anomalia o errore di misurazione. Ad ogni modo sul training set ha un ottimo fitting.

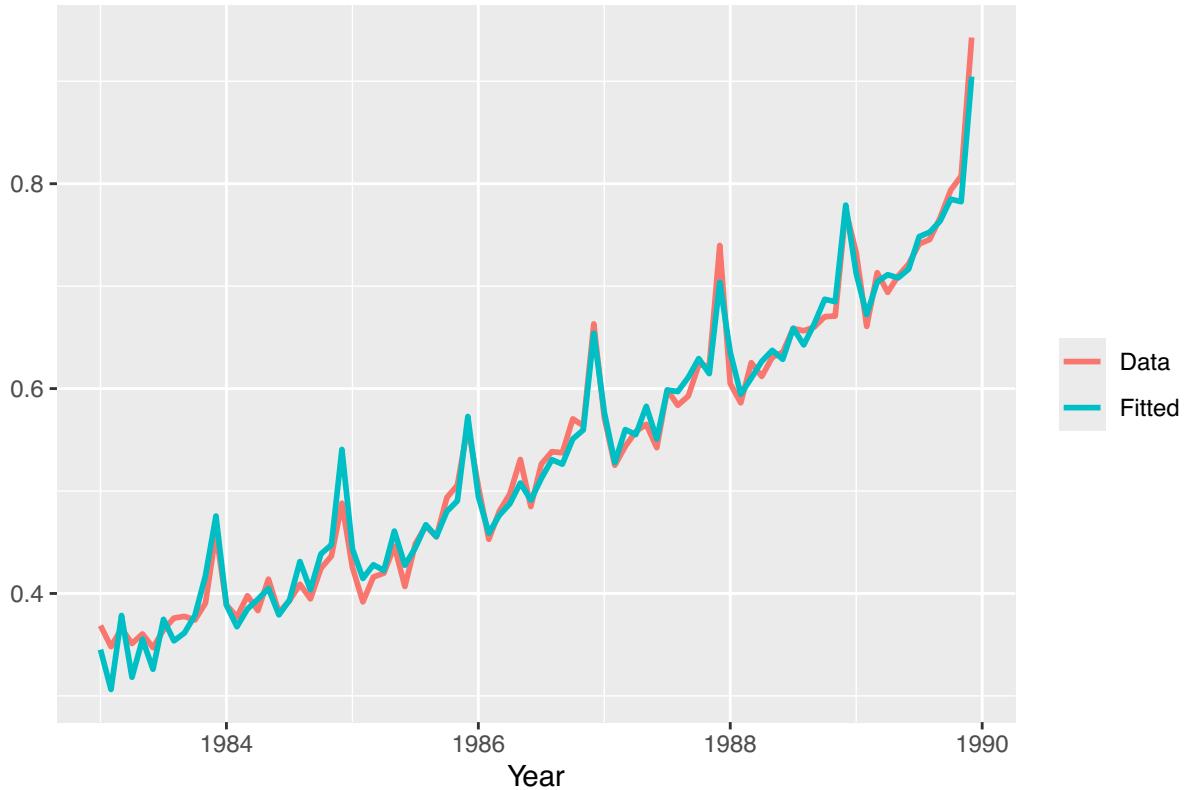
Monthly expenditure in eating out in Australia : real value vs fitted value



Anche la rappresentazione grafica conferma che il modello si adatta abbastanza bene ai dati, con partico-

lare attenzione alla situazione presente nell'ultimo anno che non riesce a cogliere bene il cambio di trend sovrastimando leggermente i dati.

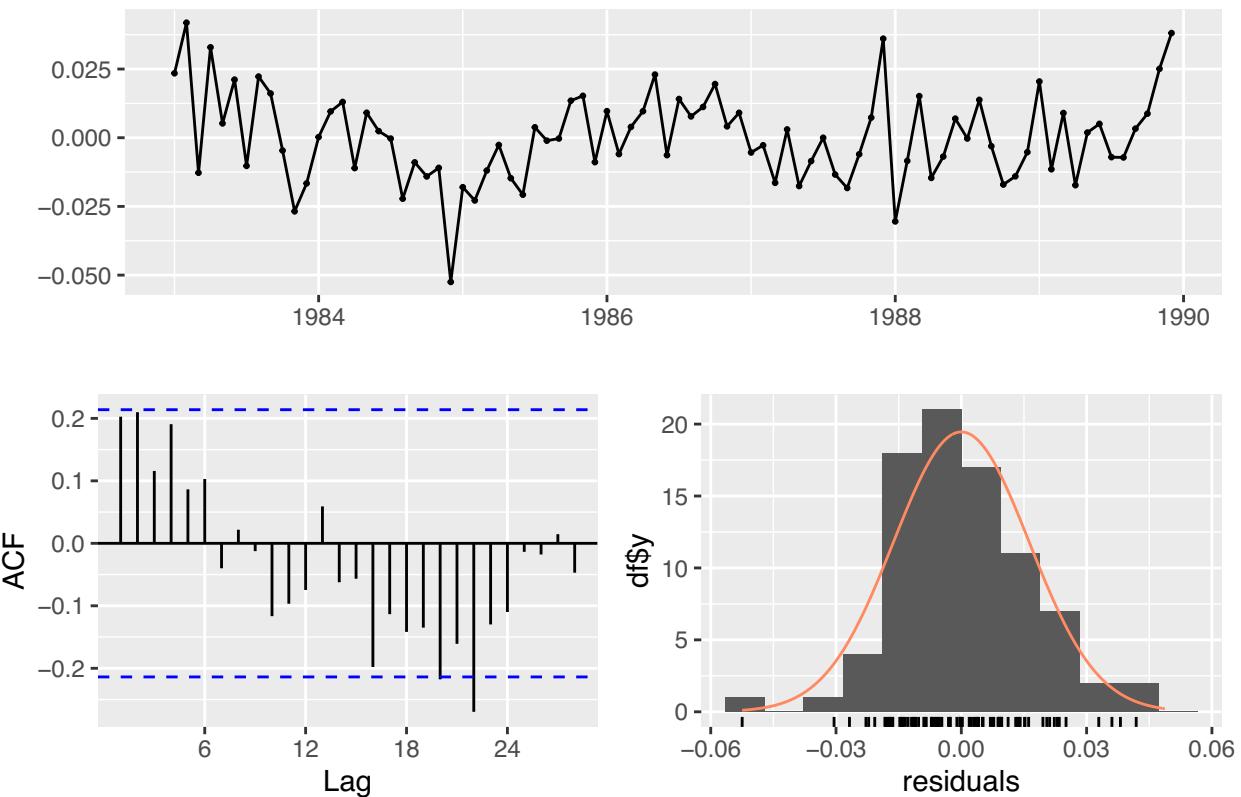
Monthly expenditure in eating out in Australia



Analisi dei residui

Dall'analisi dei residui, emerge che essi mostrano un comportamento simile al white noise. Questo suggerisce che il modello è stato in grado di catturare efficacemente le informazioni sulla variabile prevista, rendendo le previsioni potenzialmente accurate. Inoltre, tutti i ritardi dell'autocorrelazione (ACF) rientrano all'interno delle bande di confidenza, indicando l'assenza di una correlazione significativa tra la serie e i suoi ritardi. Il test di Breusch-Godfrey fornisce un valore di p-value pari a 0.30, il che suggerisce che l'ipotesi nulla di incorrelazione tra i residui può essere accettata. Il time plot dei residui mostra una media intorno allo zero, ma con una variabilità che non è costante, suggerendo la presenza di una leggera eteroschedasticità. Si nota una possibile anomalia nell'ultimo anno di osservazioni, con un cambio di pendenza che porta il modello a sovrastimare la prima metà dell'anno e a sottostimare la seconda metà dell'anno. Questo si riflette anche nella struttura del grafico temporale dei residui, tuttavia non risulta essere così significativa da incidere sulla correlazione. Infine, la distribuzione dei residui mostra una leggera asimmetria positiva, ma può essere approssimata a una distribuzione normale. In conclusione, l'analisi dei residui suggerisce che il modello potrebbe essere appropriato per le previsioni, ma potrebbe essere necessario affrontare l'eteroschedasticità e prendere in considerazione l'anomalia nell'ultimo anno di osservazioni implementando un modello diverso.

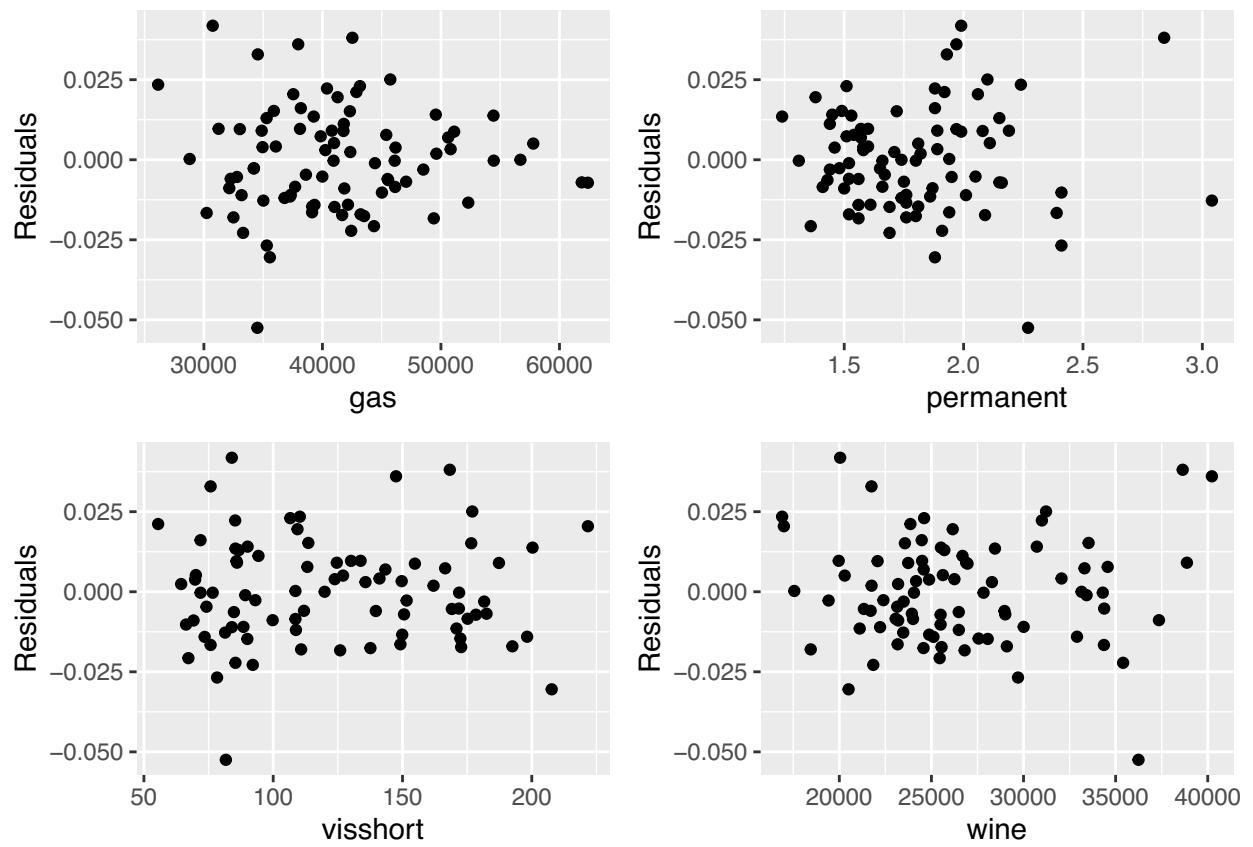
Residuals from Linear regression model



```
##  
## Breusch-Godfrey test for serial correlation of order up to 20  
##  
## data: Residuals from Linear regression model  
## LM test = 22.723, df = 20, p-value = 0.3026
```

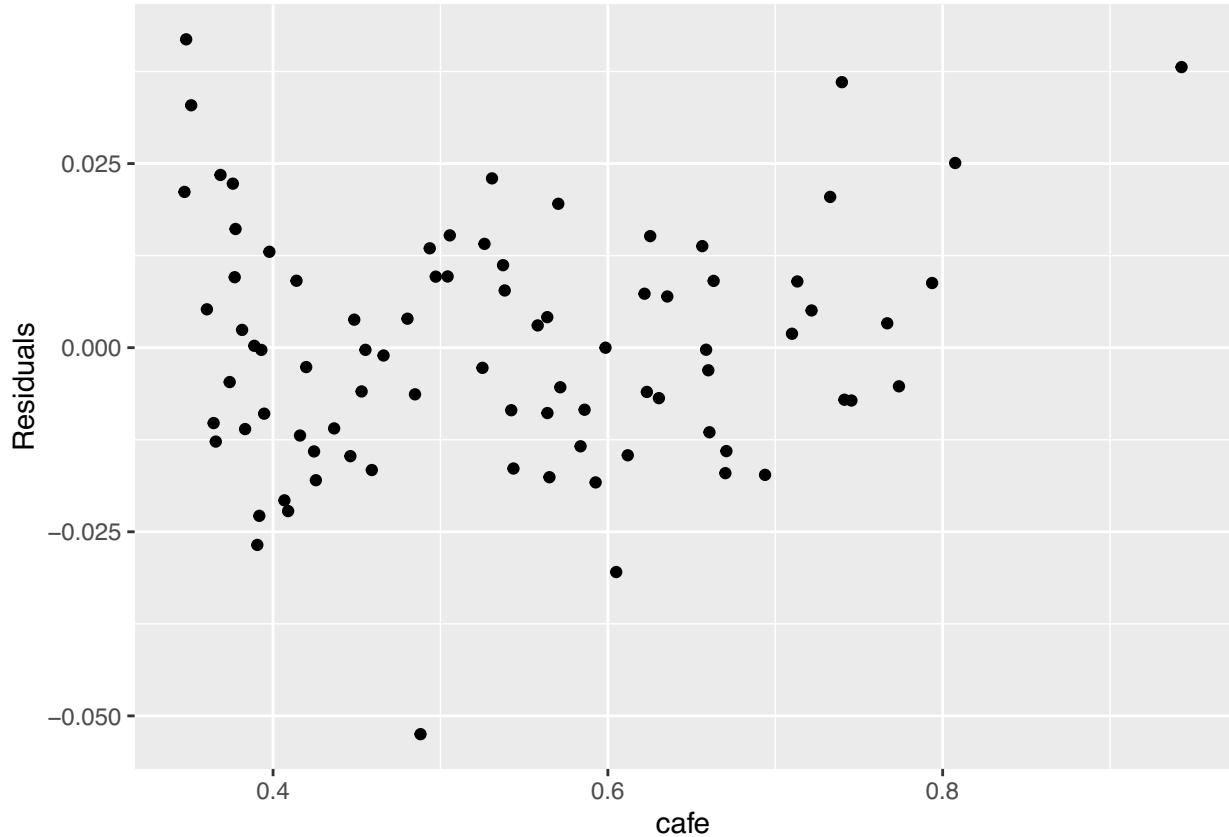
Confronto residui e regressori

I residui del modello di regressione multipla per la previsione delle spese in eating out in Australia, plottati rispetto a ciascun predittore sembrano non seguire una struttura lineare ma piuttosto casuale, suggerendo incorrelazione rispetto ai regressori.



Confronto residui e variabile dipendente

Lo stesso accade per i valori della variabile dipendente e i residui.



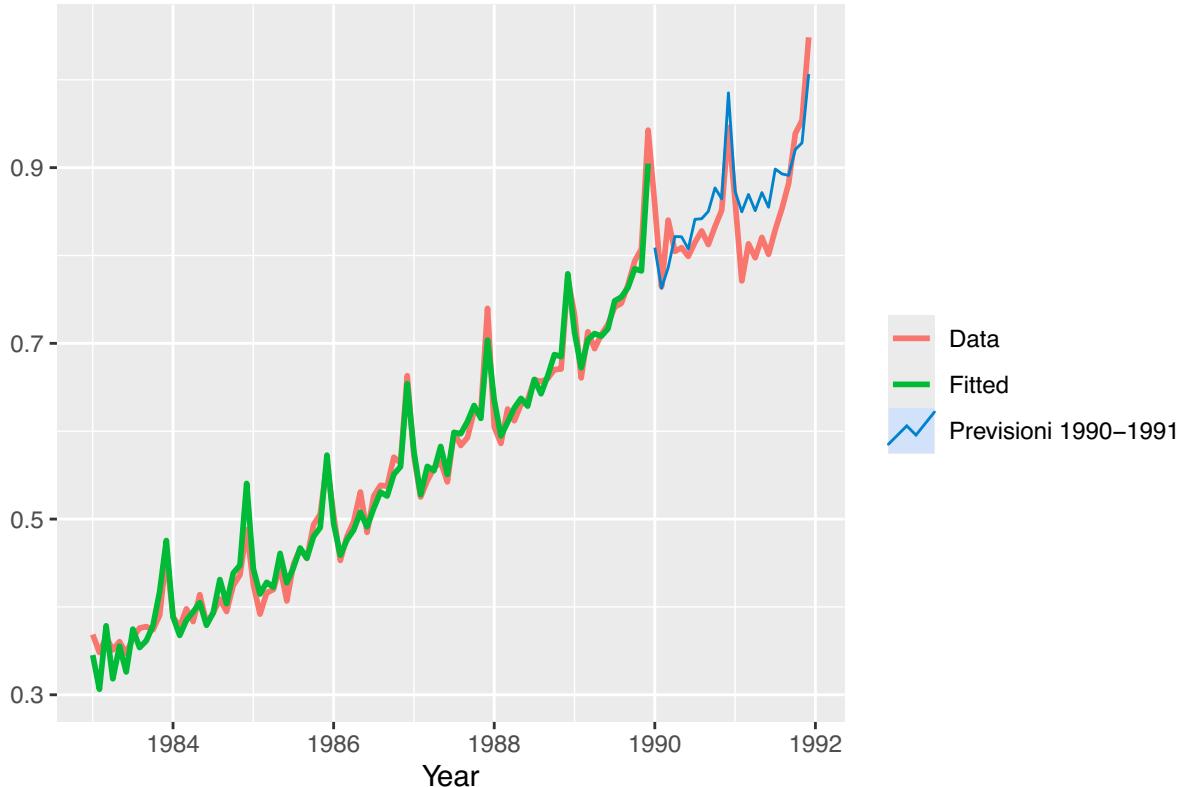
Previsioni con il modello di regressione

Il training set viene utilizzato per addestrare il modello: ovvero, il modello impara dalle osservazioni nel training set e cerca di trovare i pattern o le relazioni tra le variabili che possano aiutarlo a fare previsioni su nuovi dati. Una volta che il modello è stato addestrato, viene testato utilizzando il test set. Il test set rappresenta una porzione separata del dataset complessivo, che non è stata utilizzata per addestrare il modello. Il modello utilizza i dati del test set per fare previsioni e valutare la sua capacità di generalizzazione, ovvero di fare previsioni accurate su nuovi dati. Per cui dopo aver stimato e addestrato il modello sul training set, è possibile utilizzare il test set per fare **previsioni ex-post**. La previsione ex-post implica l'utilizzo di dati per i predittori che sono stati osservati e che non sono stati utilizzati per la stima del modello.

```
##                   RMSE        MAE       MAPE
## Training set 0.01615403 0.01275834 2.578137
## Test set      0.04015193 0.03431902 4.088907
```

Il modello riesce a stimare abbastanza bene i valori nel campione di training mentre mostra una leggera difficoltà nelle previsioni. In particolare, rispetto al test set, sovrastima pesantemente i valori (soprattutto nell'ultimo periodo)

Monthly expenditure in eating out in Australia



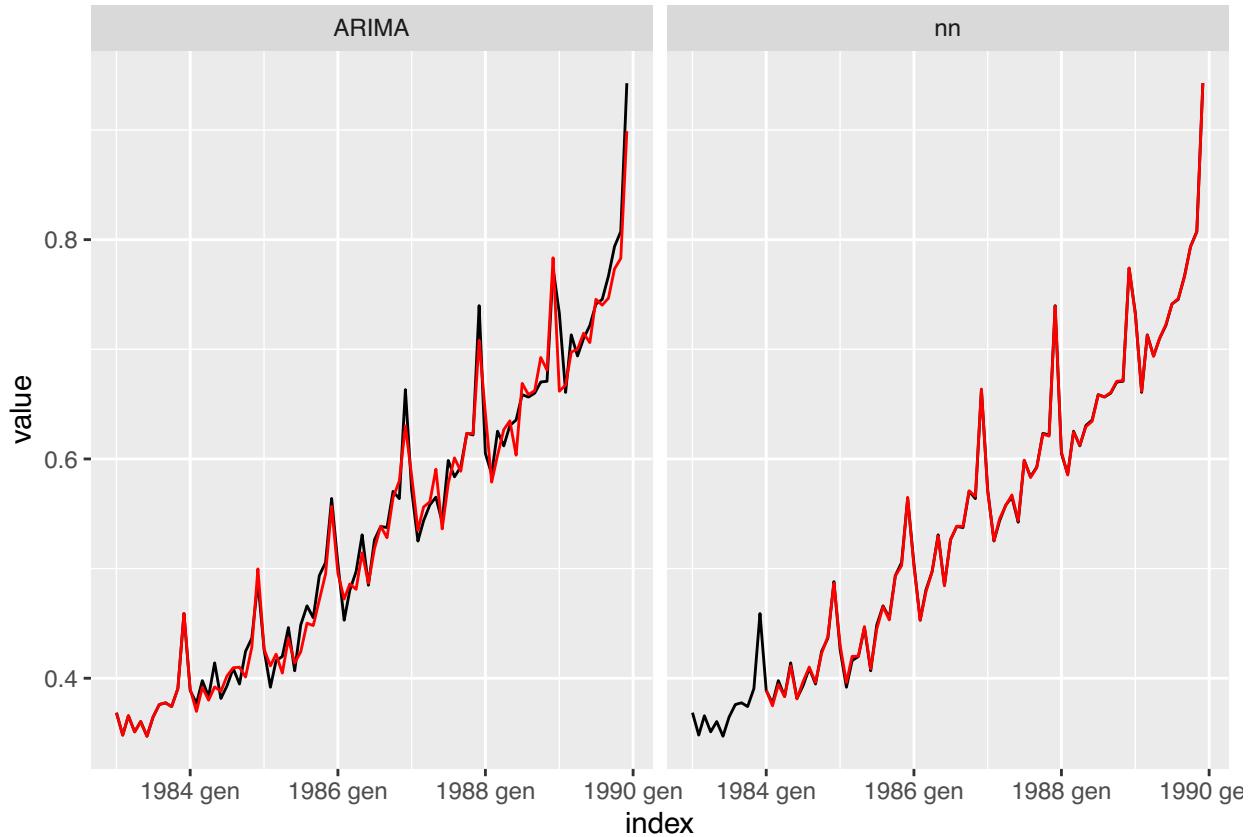
4.2.3 Rete Neurale 1

Per prevedere l'andamento delle spese in cene e pranzi fuori casa in Australia, si costruiscono due reti neurali con i seguenti parametri:

- p: il numero di ritardi utilizzati come input, lo poniamo uguale a 9 e 12.
- P: il numero di ritardi stagionali utilizzati come input, lo poniamo pari a 12.
- size: il numero di neuroni che compongono il singolo strato nascosto della rete neurale, lo poniamo pari a 6 e 10.
- repeats: il numero di reti neurali addestrate, ciascuna con pesi iniziali casuali, lo poniamo uguale a 20.

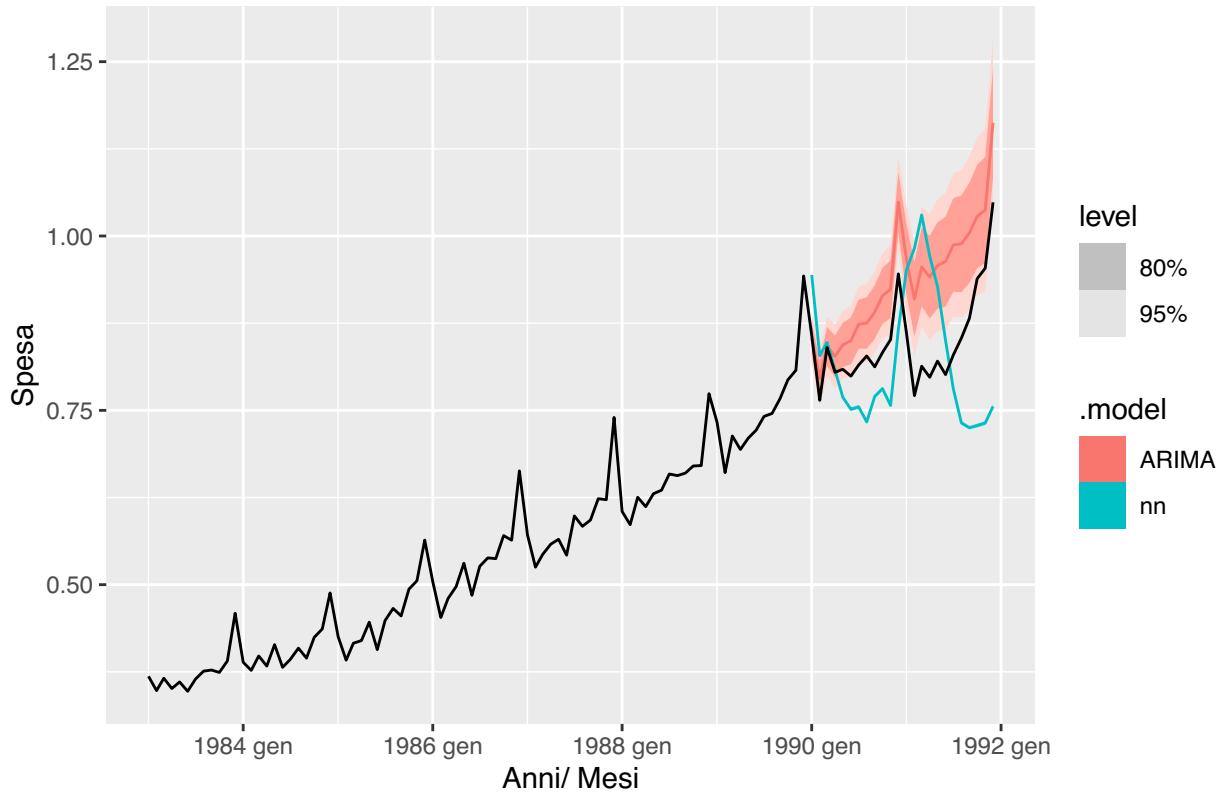
Si confrontano sia le previsioni con il modello ARIMA che con rete neurale.

Il grafico di sinistra, etichettato come “ARIMA”, mostra i valori osservati (in nero) e i valori previsti dal modello ARIMA (in rosso) nel periodo che va dal 1984 fino a poco dopo il 1990. Si può notare che il modello ARIMA segue abbastanza bene l'andamento dei dati osservati, con picchi e tendenze simili a quelli dei dati reali. Il grafico di destra, etichettato come “nn”, rappresenta un confronto simile tra i valori osservati e i valori previsti dalla rete neurale. Anche esso sembra catturare le principali tendenze e variazioni dei dati osservati, mostrando una buona capacità di adattamento ai dati storici.



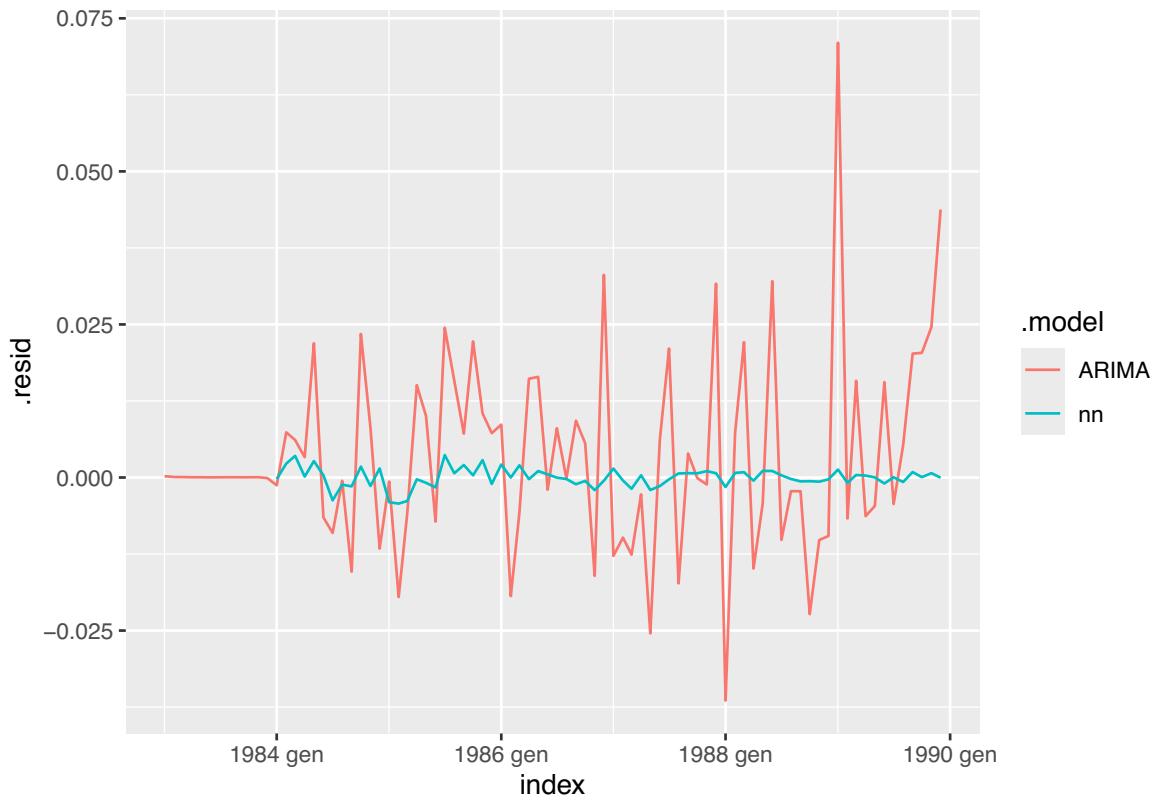
Le previsioni del modello ARIMA sono indicate con una banda rossa, che rappresenta l'intervallo di confidenza all'80% e al 95%, mentre le previsioni della rete neurale sono rappresentate in azzurro. Si osserva che entrambi i modelli prevedono un aumento nella spesa per pranzi e cene fuori casa, ma il modello ARIMA mostra una grande incertezza rispetto alla rete neurale, come indicato dall'ampiezza delle bande di confidenza. Oltretutto sembrerebbe che la rete neurale tende a sottostimare i dati mentre arima tende a sovrastimarli.

Previsione della spesa mensile per pranzi e cene fuori casa in Australia



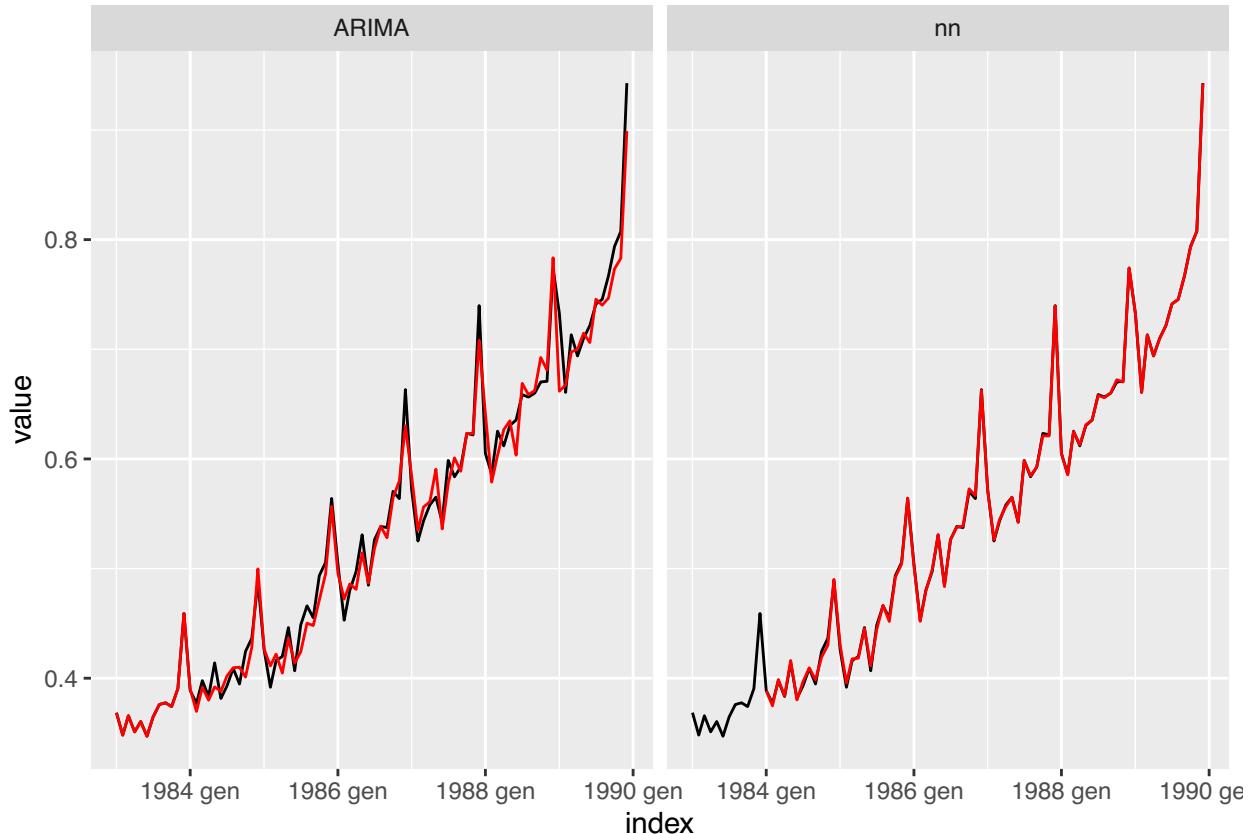
Analizzando i residui:

- **Per il modello ARIMA**, i residui mostrano una notevole variabilità nel tempo. Ci sono periodi con residui positivi e negativi di ampia entità, specialmente verso la fine del periodo analizzato. Sono presenti dei picchi che indicano che il modello ha difficoltà a prevedere con precisione alcuni punti specifici nel tempo, con errori che superano anche lo 0.05 in valore assoluto.
- **Per la rete neurale**, i residui sono molto più stabili e vicini allo zero rispetto a quella del modello ARIMA. Questo indica che la rete neurale riesce a prevedere i valori con maggiore precisione, avendo residui molto più contenuti. La stabilità dei residui suggerisce che la rete neurale cattura meglio le dinamiche sottostanti dei dati storici, risultando in previsioni più accurate.



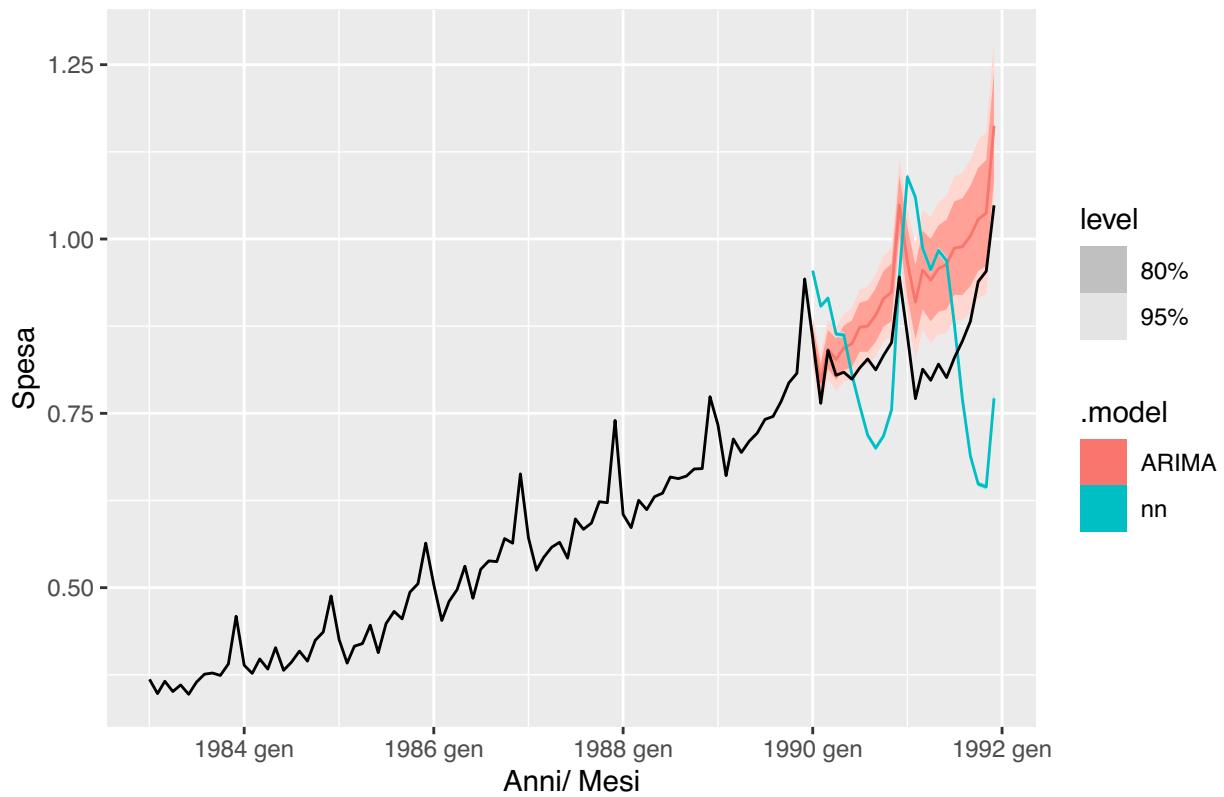
4.2.4 Rete neurale 2

In seguito, si aumentano il numero di neuroni al livello intermedio da 6 a 10, in quanto diminuirli non produce risultati utili. In questo caso aumentando il numero di neuroni al livello intermedio, è chiaro che la rete neurale riesce a cogliere meglio la variabilità dei dati e fittare la relazione. Ad ogni modo la possibilità della rete di interpolare così bene i dati del training potrebbe essere sintomo di overfitting.

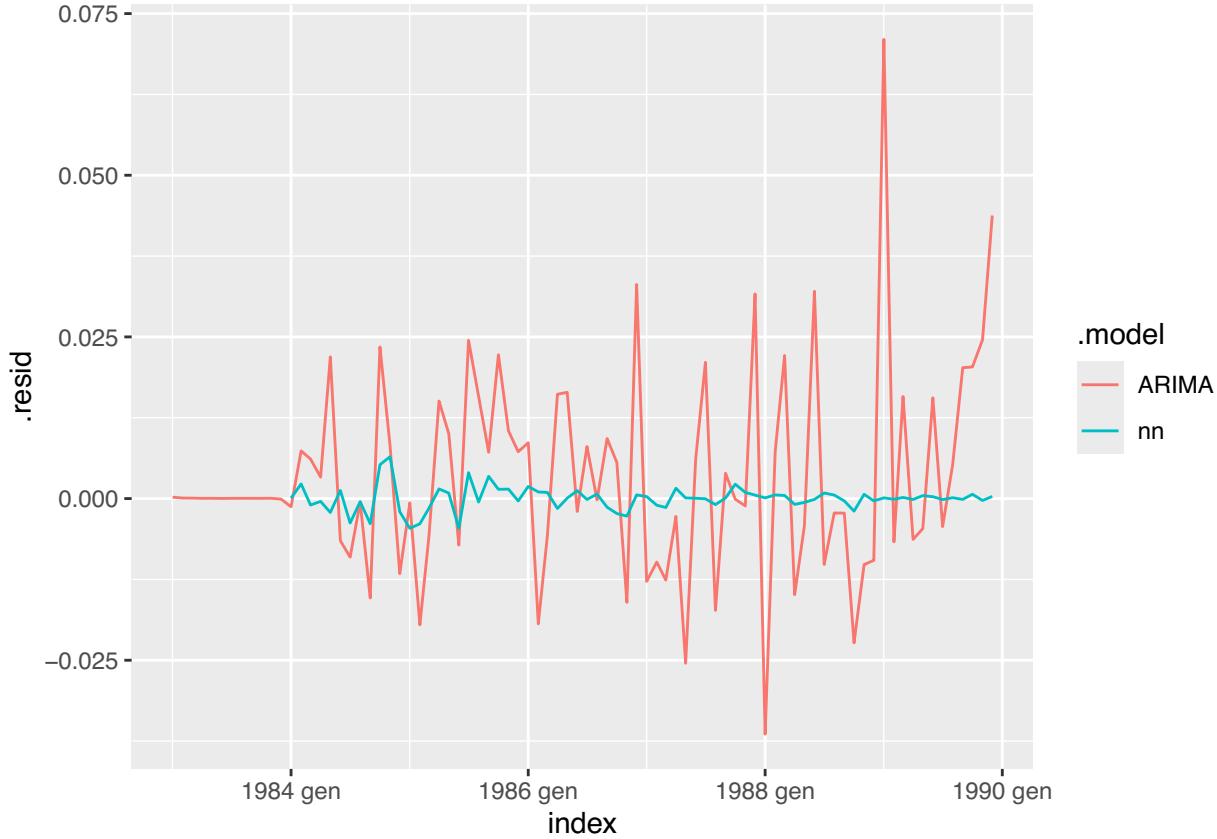


Infatti, il modello sottostima pesantemente i valori previsti per le spese di cene e pranzi fuori in australia.

Previsione della spesa mensile per pranzi e cene fuori casa in Australia



L'analisi dei residui si mostra più contenuta e stabile della precedente analisi.



4.2.5 Rete neurale con CV

E' possibile incrementare le performance della rete neurale implementando la expanded window cross-validation per selezionare i parametri di ottimizzazione. Nella cross-validation con expanded window, la finestra di addestramento aumenta gradualmente man mano che si avanza nella serie temporale. Ogni nuovo fold di addestramento include tutte le osservazioni precedenti, incrementando così il periodo di addestramento. I parametri utilizzati per il resampling sono:

- initial: il numero di osservazioni che compongono il training set, 60 mesi.
- assess: il numero di osservazioni che compongono il test set, 48 mesi.
- cumulative: indica se il ricampionamento deve avvenire su una finestra fissa (definita da initial), lo si pone uguale a FALSE.

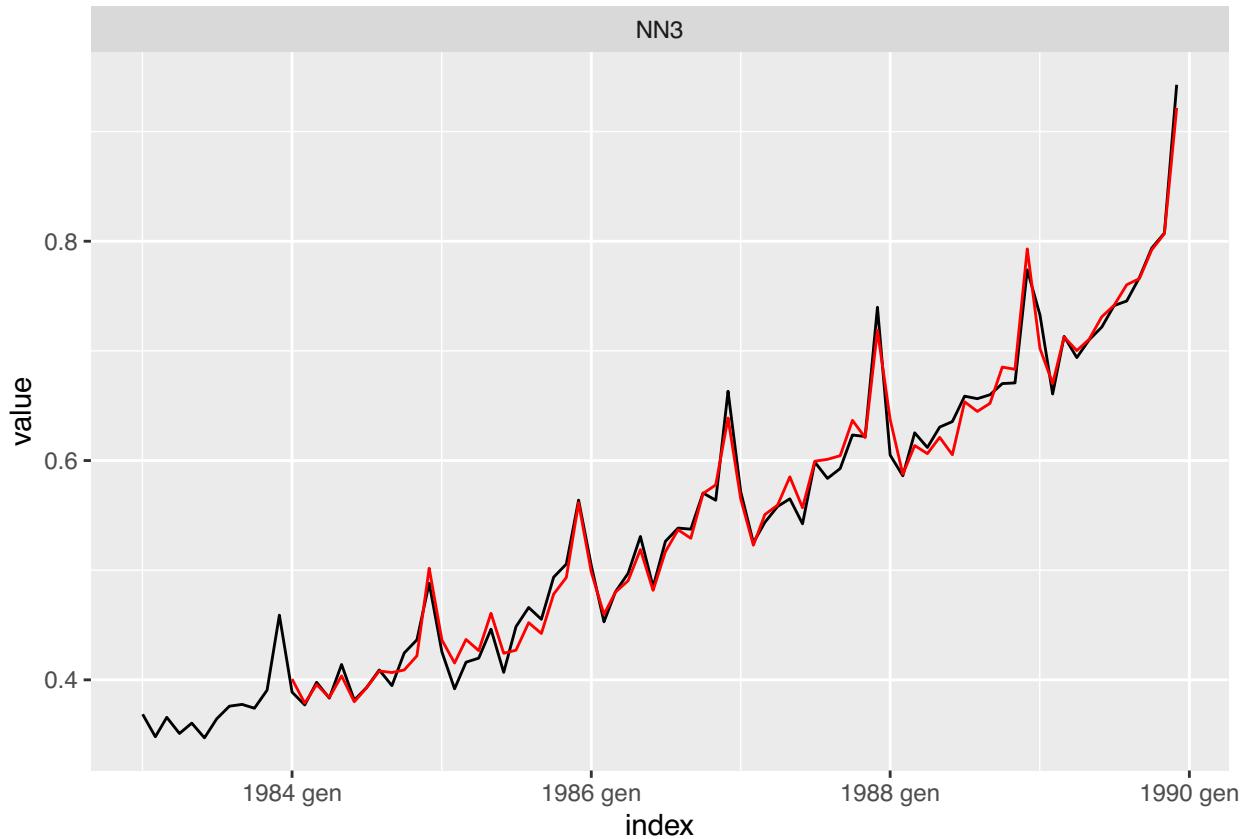
Inoltre, si considerino:

p: il numero di ritardi utilizzati come input, consideriamo i valori da 1 a 10. size: il numero di neuroni intermedi, consideriamo i valori da 1 a 10. λ : il parametro di weight-decay. Per ogni combinazione degli iperparametri considerati verrà stimata una rete neurale. Alla fine verrà scelta la combinazione di iperparametri che garantisce il miglior RMSE in cross-validation.

Osservando la tabella è possibile notare come varia l'RMSE in funzione dei 3 iperparametri considerati. Per il problema in questione il miglior modello, con l'RMSE più basso in cross-validation, è quello che presenta un numero di ritardi utilizzati come input pari a 5, un numero di neuroni intermedi pari a 6, e decay pari a 0.001.

p	size	decay	rmse
5	6	0.01	0.05758879
10	1	0.0001	0.07509626
5	1	0.0001	0.07901228
4	1	0	0.08317925
9	1	0	0.08334621

I parametri selezionati come migliori verranno adesso utilizzati per addestrare i modelli sui rispettivi training set, e i pesi stimati utilizzati per prevedere l'andamento delle spese in cene e pranzi fuori in Australia sul test set.

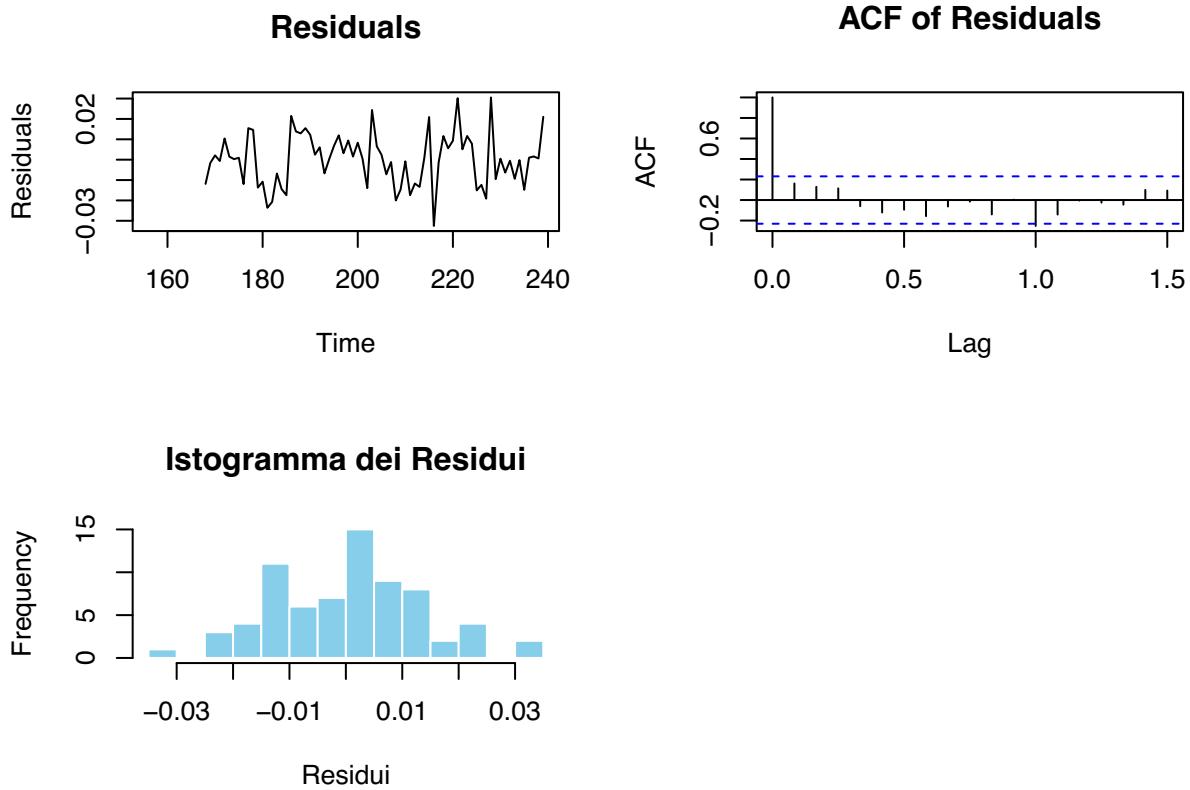


Analisi dei residui

Dall'analisi dei residui, emerge che essi mostrano un comportamento simile al white noise. Questo suggerisce che il modello è stato in grado di catturare efficacemente le informazioni sulla variabile prevista, rendendo le previsioni potenzialmente accurate. Inoltre, quasi tutti i ritardi dell'autocorrelazione (ACF) rientrano all'interno delle bande di confidenza (a eccezione del primo ritardo), indicando l'assenza di una correlazione significativa tra la serie e i suoi ritardi. Il test di Ljung-Box fornisce un valore di p-value pari a 0.18, il che suggerisce che l'ipotesi nulla di incorrelazione tra i residui può essere accettata. Il time plot dei residui mostra una media intorno allo zero, ma con una variabilità che non è costante, suggerendo la presenza di una leggera eteroschedasticità. Infine, la distribuzione dei residui non sembra avere una distribuzione normale.

```
##
## Box-Ljung test
##
## data: residuals_fit3[, 3]
```

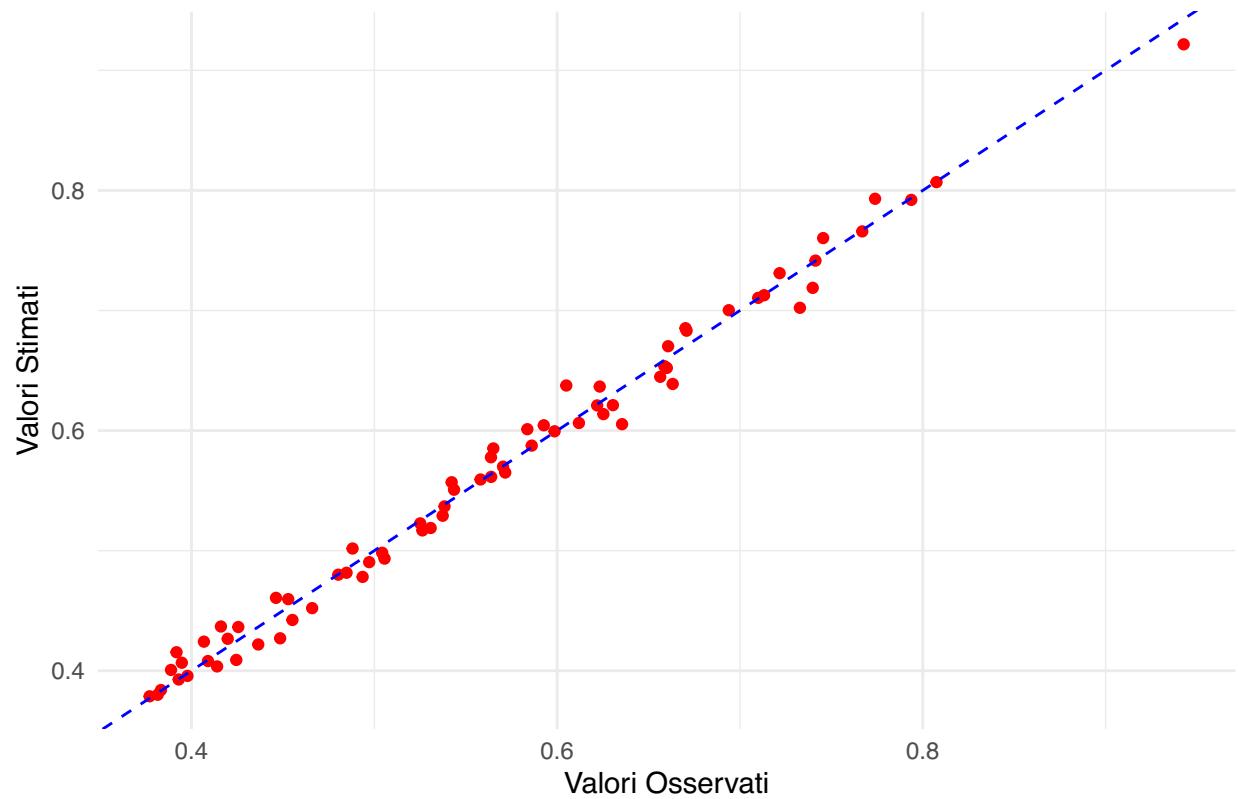
```
## X-squared = 16.131, df = 12, p-value = 0.1853
```



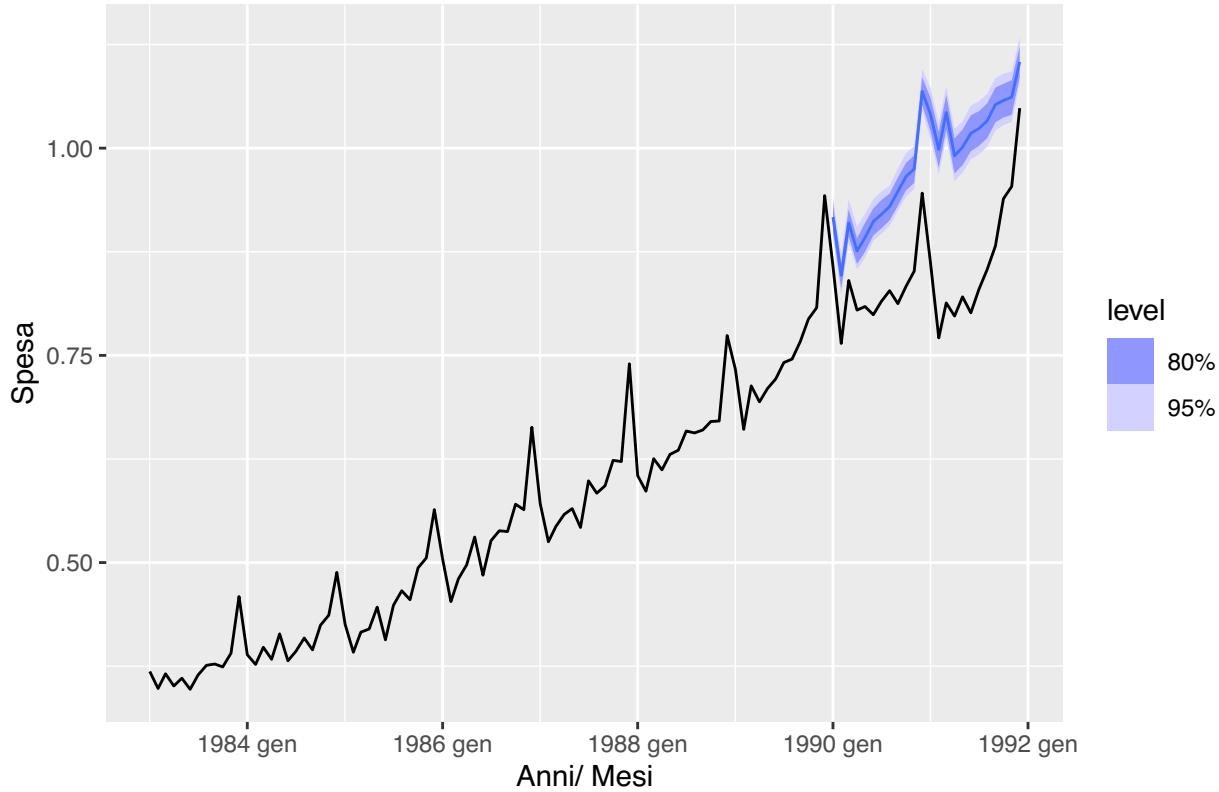
Confronto tra valori stimati e valori osservati

Il plot dei valori osservati e stimati mostra la presenza di una correlazione lineare positiva abbastanza forte, le stime sono abbastanza accurate; La maggior parte dei punti non sono direttamente sulla retta ma intorno per cui questo indica che il modello può essere sicuramente migliorato. L'ultimo valore potrebbe essere simbolo di anomalia o errore di misurazione. Ad ogni modo sul training set ha un ottimo fitting.

Confronto tra Valori Osservati e Stimati



Previsione della spesa mensile per pranzi e cene fuori casa in Australia



Analizzando i risultati delle tre reti neurali appena stimate , con nnetar : la rete NN1 ($n=6$, $p=12$) mostra un errore medio (ME) di 0.02, un errore quadratico medio (RMSE) di 0.13 e un errore assoluto medio (MAE) di 0.10, con una percentuale di errore medio (MPE) di 1.96. La rete NN2 ($n=10$, $p=9$) ottiene un errore medio (ME) di 0, un RMSE di 0.16 e un MAE di 0.13, con un MPE di -1.04. Infine, la rete NN3 ($n=6$, $p=5$, $\lambda=0.01$) presenta un ME di -0.013, un RMSE di 0.14 e un MAE di 0.13, con un MPE di -1.61. Confrontando le tre reti, la NN1 mostra il miglior RMSE e MAE, indicando una migliore precisione complessiva. Tuttavia, la NN2 ha un ME perfettamente bilanciato, risultando in assenza di bias sistematico, sebbene il suo RMSE sia il più alto tra le tre reti. La NN3, nonostante abbia un RMSE e MAE intermedi, può essere considerato un compromesso tra NN1 e NN2, ma non supera NN1 in termini di accuratezza complessiva. Inoltre, mostra un ME negativo, suggerendo una leggera tendenza a sottostimare. La NN1 sembra offrire un buon equilibrio tra accuratezza e bias, rendendola la rete più performante tra le tre secondo le metriche fornite.

4.2.6 LSTM (Long Short-Term Memory)

Infine, sono state implementate quattro diverse architetture di reti neurali ricorrenti: l'LSTM (Long Short-Term Memory) , la GRU (Gated Recurrent Unit), Elman e Jordan.

I dati sono stati sottoposti ad una prima fase di pre-processing in quanto questi algoritmi richiedono che i dati di input siano centrati e scalati. Inoltre, per poter essere passati in input alla rete è stato necessario trasformare i dati in array avente 3 dimensioni: la numerosità del campione, il numero di lag (timesteps) e il numero di features. In questo caso il timestep è pari a 1, così come il numero di features.

Sono stati inoltre settati i vari parametri come il numero di epoche di allenamento del modello e la batch size, ovvero il numero di campioni che verranno propagati attraverso la rete. I modelli sono stati trainati per 50 epoche con una batch size pari a 72. Come ottimizzatore è stato utilizzato Adam con un learning rate pari a 0.001 e come funzione di perdita il mean absolute error.

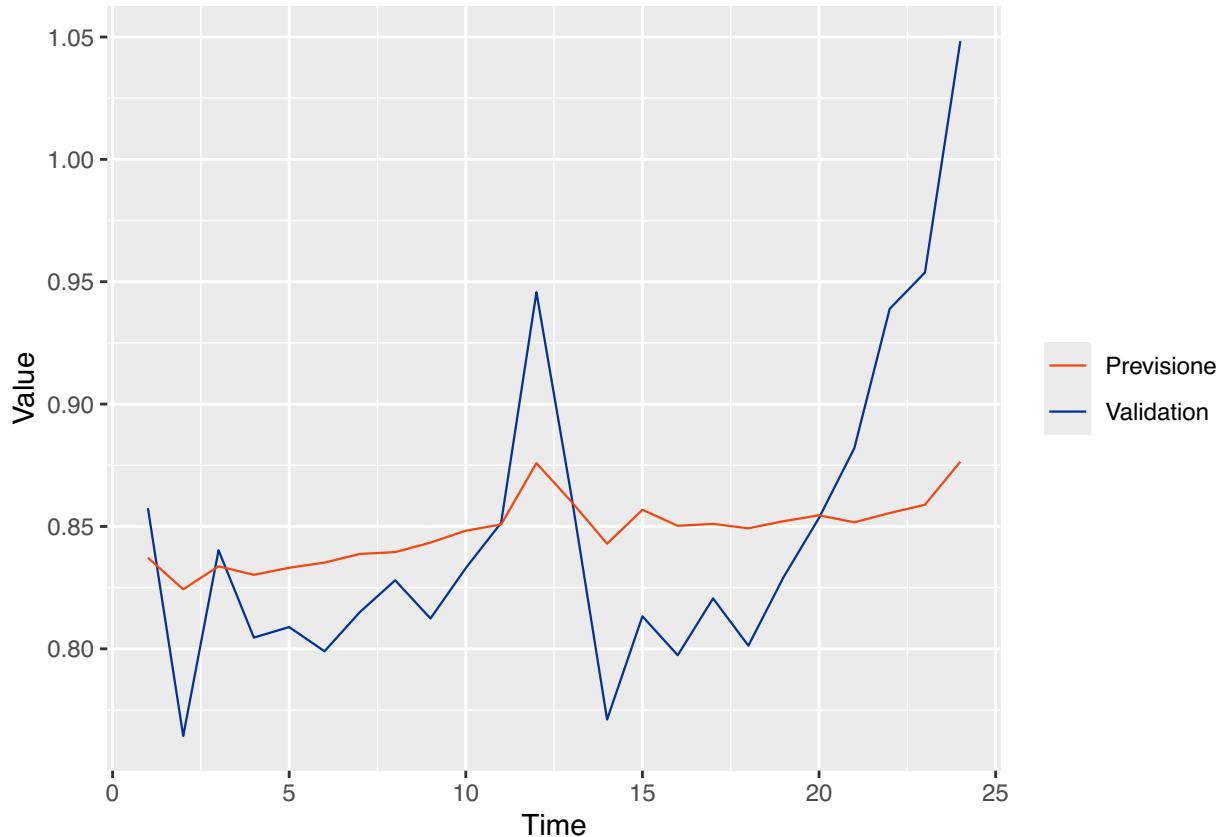
La prima rete, LSTM, è un tipo di rete neurale ricorrente (RNN) introdotta da Sepp Hochreiter e Jürgen Schmidhuber nel 1997. Risolve il problema della “scomparsa del gradiente” nelle RNN tradizionali, consentendo loro di gestire sequenze più lunghe mantenendo informazioni a lungo termine.

L'LSTM ha una struttura complessa, ma l'idea fondamentale è di avere “celle di memoria” che possono memorizzare informazioni per lungo tempo e decidere quando dimenticarle o aggiornarle. Queste celle operano attraverso tre “portali” principali:

- Porta di input (input gate): Decide quali informazioni nuove accettare nella cella di memoria.
- Porta di dimenticanza (forget gate): Decide quali informazioni nella cella di memoria mantenere o dimenticare.
- Porta di output (output gate): Decide quali informazioni nella cella di memoria utilizzare per l'output del modello.

Questi “portali” sono come filtri che regolano il flusso di informazioni nella cella di memoria. Ciò permette all'LSTM di trattenere informazioni significative per periodi di tempo più lunghi, rendendolo efficace nel trattare con sequenze lunghe e complesse.

`## 1/1 - 0s - 332ms/step`



4.2.7 RNN 2: GRU (Gated Recurrent Unit)

Si procede, quindi, con lo sviluppo della seconda rete ricorrente caratterizzata da un'architettura GRU. I dati passati in input sono gli stessi. L'addestramento per la GRU prevede l'utilizzo di 10 neuroni al livello

intermedio, il numero di epoche è fissato a 50 (poichè per combinazioni diverse il modello andava in forte overfitting). Rispetto al modello precedente questa rete neurale sembra funzionare peggio ma , in ogni modo, i valori previsti di discostano relativamente poco (0.5/0.6billioni) da quelli reali. E' chiaro che il secondo anno di previsione non sembra essere stato colto per nulla.

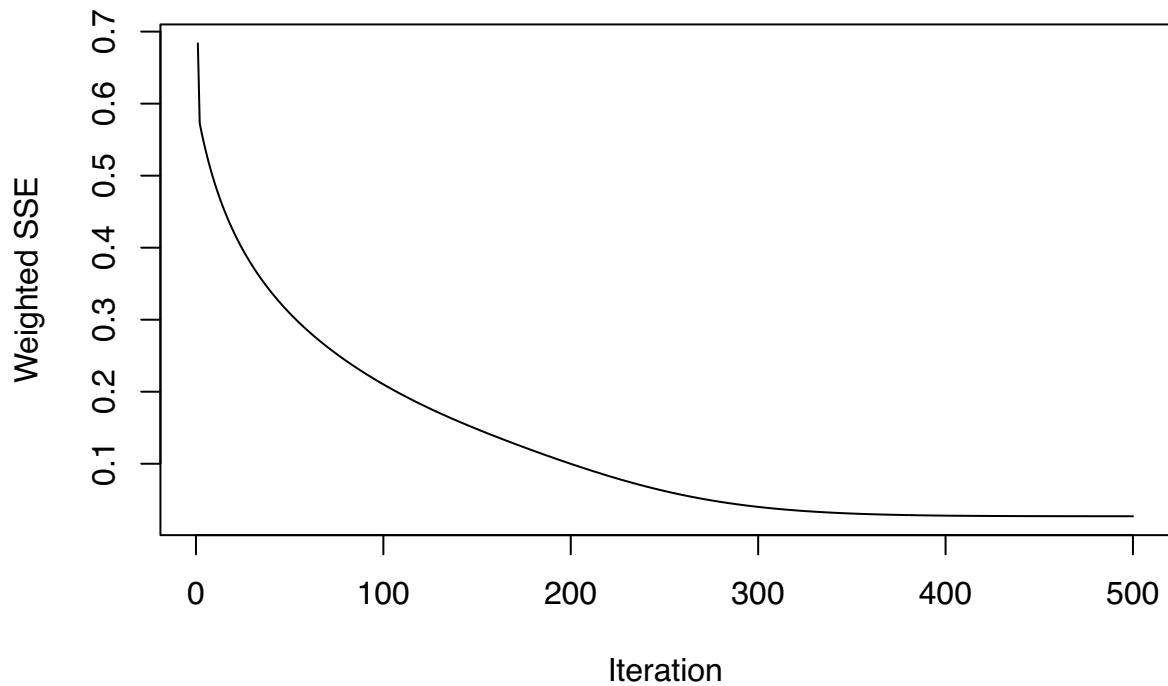
```
## 1/1 - 0s - 210ms/step
```



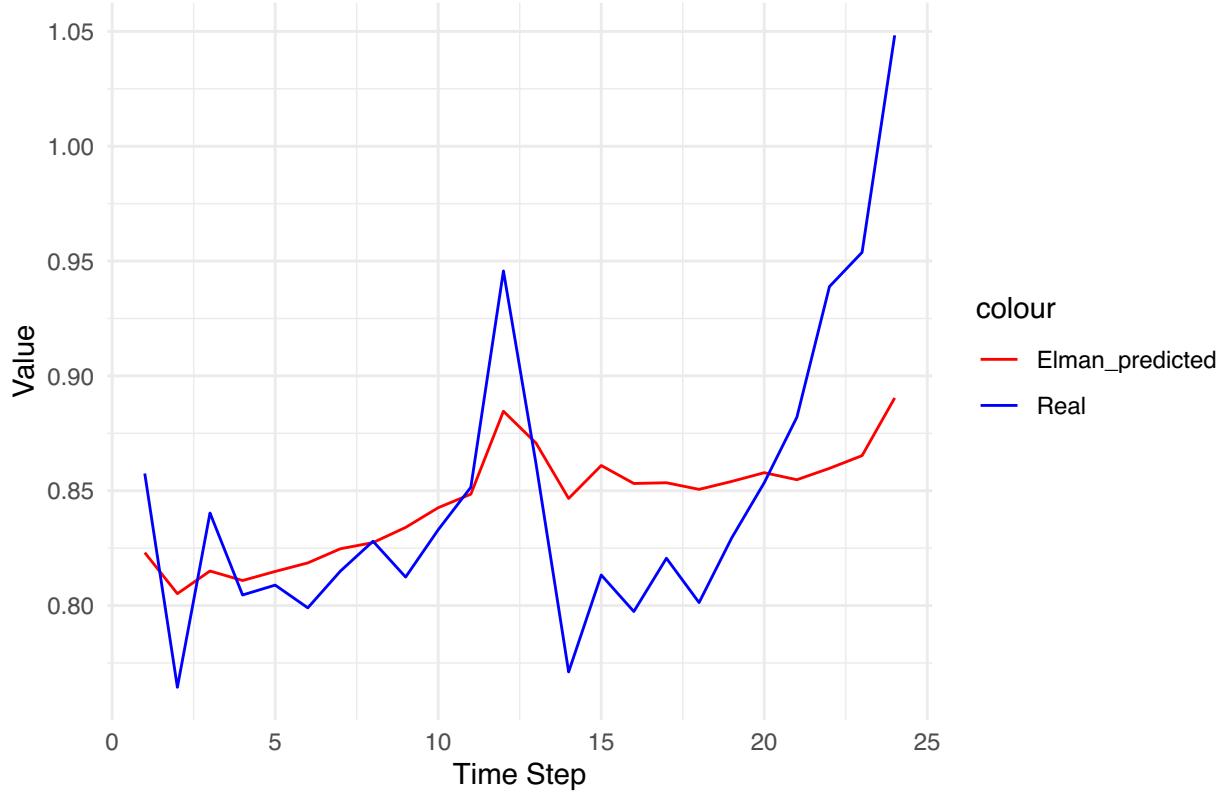
I risultati ottenuti sul test set sono abbastanza soddisfacenti. I risultati di entrambe le reti sono stati soddisfacenti, ma l'architettura LSTM ha raggiunto performance migliori. Come si puo vedere dal plot di validation, le previsioni ottenute grazie a questo modello si adattano molto bene ai dati del test set.

4.2.8 RNN 3: Elman

La rete di Elman è un tipo di rete neurale ricorrente utilizzata per elaborare dati sequenziali. La sua caratteristica distintiva è l'aggiunta di uno strato nascosto che funge da memoria a breve termine. Questo strato nascosto mantiene informazioni sulle iterazioni precedenti, consentendo alla rete di catturare dipendenze temporali nei dati di input. Durante l'addestramento, la rete di Elman utilizza algoritmi di ottimizzazione per regolare i pesi in modo che le previsioni della rete si avvicinino il più possibile ai valori desiderati. In tal caso la costruzione prevede l'utilizzo di un numero massimo di iterazioni pari a 500 e numero di neuroni al livello intermedio pari a 8. I risultati sul test set sembrano essere abbastanza accurati, per nulla inverosimili. Anche questa struttura non riesce a cogliere del tutto la struttura del secondo anno.



Elman_predicted vs Real Values

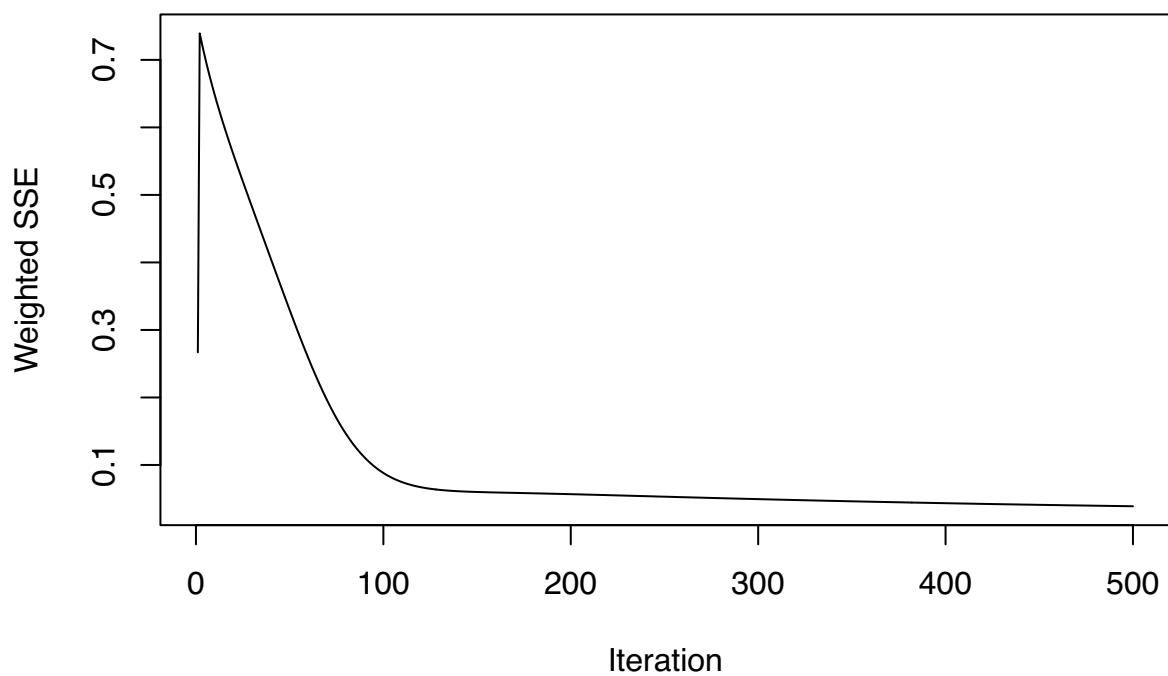


4.2.9 RNN 4: Jordan

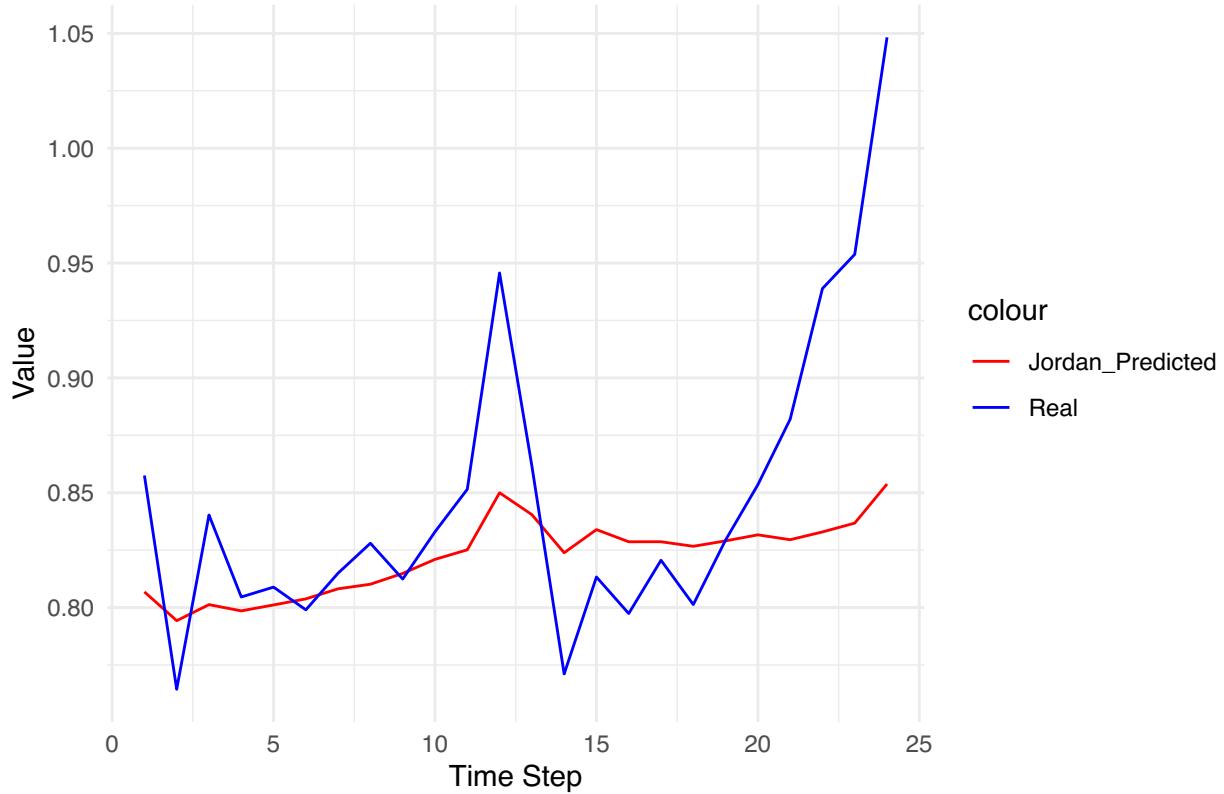
E' stata poi implementata la rete neurale di Jordan. È una variante delle reti di Elman, introdotte da Jeffrey Elman, che aggiunge un collegamento di feedback diretto dall'output della rete al suo stato nascosto. Durante il processo di addestramento, la rete di Jordan riceve in input sequenze di dati e genera delle previsioni in base al suo stato nascosto e ai collegamenti di feedback. Durante l'addestramento, l'obiettivo è regolare i pesi della rete in modo che le sue previsioni si avvicinino il più possibile ai valori desiderati. Il fatto che la loss cresca considerevolmente e poi scenda rapidamente a 0 e rimanga stabile dopo un certo numero di iterazioni può essere dovuto a diversi fattori:

- Convergenza Rapida: Il modello potrebbe avere una capacità sufficiente per adattarsi ai dati di addestramento molto rapidamente, riducendo la loss function a 0 nelle prime iterazioni.
- Stabilità della Rete: Dopo un certo numero di iterazioni, il modello potrebbe aver raggiunto una stabilità nei suoi pesi e nei suoi stati interni, il che potrebbe portare a una loss function stabile.

La rete Jordan sembra comportarsi abbastanza male sul test set di previsione. Non è in grado di cogliere le variazioni significative nel primo anno e lo stesso accade per il secondo anno. Tuttavia, tiene conto del trend crescente che viene catturato.



Jordan_Predicted vs Real Values



5. Confronto tra modelli

Nel corso dello studio, sono stati confrontati diversi modelli per la previsione di serie temporali, tra cui modelli statistici tradizionali, reti neurali artificiali e architetture di reti neurali ricorrenti, valutandone le prestazioni su un set di test utilizzando diverse metriche di errore.

Il modello Seasonal Naive, che assume che i valori futuri saranno uguali a quelli dello stesso periodo dell'anno precedente, ha mostrato una buona performance con un ME di 0.04, un RMSE di 0.10 e un MAPE di 11.20. Questo approccio semplice ha offerto un buon punto di partenza per il confronto.

L'aggiunta di componenti di trend e stagionalità ha migliorato significativamente i risultati, riducendo l'errore medio assoluto (MAE) a 0.03 e il MAPE a 4.23. Il modello ha dimostrato che incorporare le caratteristiche stagionali e di trend nella serie temporale può migliorare notevolmente le previsioni.

Un ulteriore raffinamento, con l'inclusione di quattro regressori aggiuntivi, ha portato ad un lieve miglioramento delle metriche, con un RMSE e MAE entrambi pari a 0.04 e un MAPE di 4.08, evidenziando l'importanza delle variabili esplicative nel migliorare l'accuratezza delle previsioni.

Le reti neurali artificiali, rappresentate da NN1, NN2 e NN3, hanno mostrato prestazioni variabili. NN1, con 6 neuroni nello strato nascosto e un lag di 12, ha ottenuto un ME di 0.02 e un MAPE di 1.96, ma un RMSE più alto di 0.13. NN2, con 10 neuroni e un lag di 9, ha registrato un ME di 0 ma un RMSE di 0.16 e un MAPE di -1.04, suggerendo una possibile sovrastima o sottostima sistematica. NN3, che include una regolarizzazione L2 ($\lambda=0.01$), ha mostrato un ME negativo di -0.013 e un RMSE di 0.14, indicando una leggera tendenza alla sottostima.

Le reti neurali ricorrenti hanno mostrato prestazioni più robuste. Il modello LSTM (Long Short-Term Memory) ha ottenuto un ME di 0.013, un RMSE di 0.05 e un MAE di 0.04, con un MAPE di 2.04.

Il modello GRU (Gated Recurrent Unit) ha mostrato un ME di -0.03, un RMSE di 0.06 e un MAE di 0.04, con un MAPE di -3.66. Questo risultato suggerisce che il GRU, pur essendo leggermente meno preciso dell'LSTM, è comunque un modello potente per la previsione con un numero di parametri nettamente inferiore.

Le reti neurali ricorrenti di Elman e Jordan hanno dimostrato ottime capacità di previsione, con il modello Elman che ha ottenuto un ME di 0.004, un RMSE di 0.05 e un MAE di 0.04, e un MAPE di 0.96. Il modello Jordan ha mostrato un ME di -0.02, un RMSE di 0.06 e un MAE di 0.04, con un MAPE di -2.80. Entrambi i modelli hanno beneficiato della capacità di memorizzare informazioni sulle iterazioni precedenti, ma la rete di Jordan ha leggermente sofferto di un errore di bias negativo, suggerendo una tendenza alla sottostima.

Nel complesso, i modelli di LSTM e Elman si sono dimostrati i più efficaci nel bilanciare accuratezza e capacità di generalizzazione nelle previsioni, con prestazioni superiori rispetto ai modelli statistici tradizionali e altre architetture di reti neurali.

Modello	tipo	ME	RMSE	MAE	MAPE
Seasonal Naive	Test	0.04	0.10	0.09	11.20
Modello con trend e stagionalità	Test	0.02	0.04	0.03	4.23
Modello con trend, stagionalità e 4 regressori	Test	0.02	0.04	0.03	4.08
NN1(n=6, p=12)	Test	0.02	0.13	0.10	1.96
NN2(n=10, p=9)	Test	0	0.16	0.13	-1.04
NN3(n=6, p=5, $\lambda=0.01$)	Test	-0.013	0.14	0.13	-1.61
LSTM	Test	0.013	0.05	0.04	2.04
GRU	Test	-0.03	0.06	0.04	-3.66
ELMAN	Test	0.004	0.05	0.04	0.96
JORDAN	Test	-0.02	0.06	0.04	-2.80

6. Conclusioni

L'analisi dettagliata condotta sui dati relativi alle spese totali mensili per mangiare fuori in Australia rivela una panoramica interessante. Dall'analisi esplorativa rispetto alle altre variabili di interesse è evinto che:

- visshort e resshort mostrano un andamento simile, suggerendo una possibile correlazione
- wine e expenditure mostrano un andamento opposto, indicando una relazione inversa.

Inoltre:

- wine e gas hanno una distribuzione approssimativamente normale con code destre pesanti
- permanent e resshort mostrano una distribuzione concentrata intorno alla media
- visshort, expenditure e toss presentano un andamento bimodale, probabilmente dovuto alla stagionalità.
- expenditure, la dipendente del modello, risulta fortemente correlata positivamente, relazione statisticamente significativa, con visshort, resshort, permanent e gas, e debolmente con toss e wine.

Successivamente, sono stati esplorati diversi modelli per la previsione della variabile in esame su modelli statistici tradizionali e reti neurali. L'obiettivo è stato comprendere quale approccio fosse più efficace per prevedere le spese totali mensili per mangiare fuori in Australia, utilizzando o modelli autoregressivi o variabili esplicative (nei casi di regressione lineare multipla) come il consumo di vino, la produzione di gas, e varie tipologie di partenze.

Dai risultati ottenuti, i modelli che integrano trend e stagionalità, oltre a quelli che includono regressori aggiuntivi, hanno dimostrato un miglioramento significativo nelle prestazioni rispetto ai modelli seasonal naïve. Questi modelli, incorporando informazioni aggiuntive, sono stati in grado di catturare meglio le dinamiche sottostanti alla serie temporale, riducendo significativamente l'errore medio assoluto (MAE) e l'errore quadratico medio (RMSE).

Le reti neurali artificiali (NN) hanno mostrato prestazioni variabili a seconda della configurazione utilizzata. Nonostante la loro capacità di modellare relazioni non lineari, i modelli NN hanno evidenziato difficoltà nel gestire dipendenze temporali complesse e tendenze a sovrastimare o sottostimare i valori previsti.

Le reti neurali ricorrenti, in particolare le architetture LSTM (Long Short-Term Memory) e GRU (Gated Recurrent Unit), hanno offerto risultati più robusti. Il modello LSTM ha dimostrato una capacità superiore di catturare le dipendenze a lungo termine nei dati, ottenendo un errore medio assoluto e un errore quadratico medio inferiori rispetto ai modelli tradizionali e alle NN. Anche il modello GRU ha fornito previsioni accurate, sebbene con una leggera tendenza alla sottostima.

Le reti Elman e Jordan hanno ulteriormente mostrato la loro efficacia nella previsione di serie temporali. La rete di Elman ha fornito previsioni con un errore medio molto basso, dimostrando una leggera tendenza alla sovrastima. La rete di Jordan ha avuto performance simili, ma ha evidenziato una leggera sottostima nei valori previsti.

In conclusione, i risultati dello studio suggeriscono che le reti neurali ricorrenti, in particolare i modelli LSTM e Elman, rappresentano le scelte migliori per la previsione di questo problema. Questi modelli hanno dimostrato una capacità superiore di gestire le dipendenze temporali e di catturare pattern complessi nei dati, risultando in previsioni più accurate e affidabili rispetto ai modelli tradizionali e alle reti neurali artificiali semplici.