GDP ITALY – Time Series Analysis

Davide Mascolo, 01/06/2021

1. Abstract

- In questo studio si vuole analizzare la serie storica relativa al *PIL dell'Italia.*

Per maggiori informazioni, clicca qui.

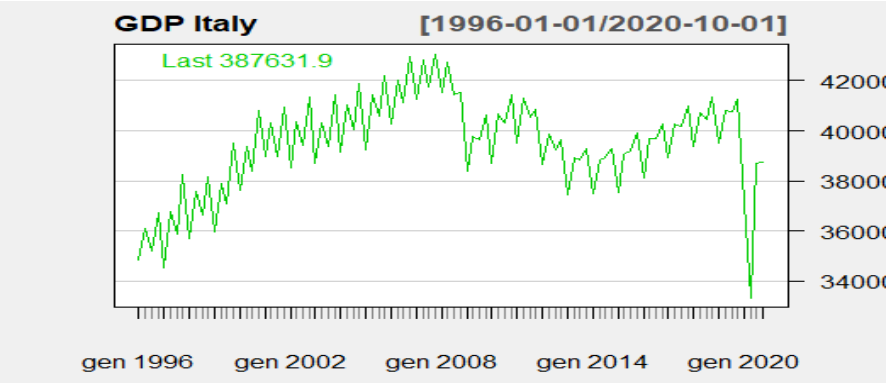
- I dati presi in considerazione vanno dal *primo trimestre dell'anno 1996 all'ultimo trimestre dell'anno 2020* e l'obiettivo di quest'analisi è *studiare le varie componenti della serie storica* utilizzando tecniche di visualizzazione e metodi di previsione.

2. Data Wrangling

- In questa fase preliminare, si trattano i dati per controllare eventuali *valori mancanti*, *valori anomali* e trasformare il set di dati in un oggetto di classe *time series*.
- Per questo studio, *non sono presenti valori mancanti e valori anomali.*

3. Analisi Descrittiva

- Partendo da una prima visualizzazione della serie, si possono evincere le principali caratteristiche.
- Dal *grafico1* si evince come la serie presenta una ***forte componente stagionale***, ***un trend crescente*** anche se a tratti e qualche improvviso *cambio di livello* nel medio-lungo periodo.



- I *picchi*, infatti, si ripetono periodicamente al *secondo ed al quarto trimestre* per ogni anno; mentre nel *primo e terzo trimestre* la serie raggiunge sempre un *calo*.
- Un evento anomalo accade nell'*ultimo trimestre del 2008,* quando la serie non va al rialzo bensì al *ribasso* e ciò è dovuto alla ***Grande Recessione***.

- Altra anomalia la si registra nel *secondo trimestre dell'anno 2020*. Infatti, per quanto detto precedentemente, il secondo trimestre di ogni anno fa registrare sempre una crescita, ma ciò non vale per il suddetto anno e quest'effetto è sicuramente legato al ***Covid***. Nell'*ultimo trimestre del 2020*, si registra un *aumento* che rispecchia l'andamento di tutti gli anni in Q4, ma questo aumento è molto *meno marcato* rispetto a tutti gli altri anni.

- Per la componente di *trend*, si nota come questa sia presente a ***tratti*** nella serie; infatti, *fino al 2008* si ha un ***trend nettamente crescente*** poi un trend decrescente, crescente e successivamente ancora decrescente fino al 2014 per poi andare ancora al rialzo fino al 2019. Proprio questo trend così altalenante e non regolare conferma che la serie sia caratterizzata da *improvvise variazioni*.
- La serie storica in esame non è ***stazionaria***, in quanto una serie si definisce tale se le sue *proprietà statistiche non dipendono dal tempo*, cioè dal momento in cui la serie viene osservata. Ciò equivale a dire che una serie storica con *trend* o *stagionalità* non sarà *stazionaria*, dato che il trend e la stagionalità influenzano il valore della serie in diversi istanti temporali.
  Se invece ci trovassimo dinanzi una seria con andamento ciclico ma senza stagionalità e trend, la serie risulterebbe sicuramente stazionaria, in quanto i cicli sarebbero aperiodici.
- Una serie ***stazionaria*** avrà un aspetto orizzontale, con variazioni più o meno costanti che ricadono nello stesso range di valori prestabilito.

- Il *grafico2* (***Seasonal Subseries Plot***), permette di capire se vi è una componente di trend e di stagionalità nella serie. Fissato il trimestre di riferimento, infatti, si osservano i valori per ogni anno. Le linee blu orizzontali indicano le medie per ogni periodo, in questo caso per ogni trimestre. La prima linea blu è il *valore medio* del PIL riferito al primo trimestre di tutti gli anni. La seconda linea blu è sempre il valore medio del PIL per tutti gli anni, ma questa volta riferito al secondo trimestre e così via.

- Si conferma quanto detto prima, ovvero che per tutti gli anni il *primo trimestre* presenta sempre un *calo* e quindi una media del PIL sempre più bassa rispetto agli altri trimestri. Il *valore medio più alto*, invece, si registra nell'*ultimo trimestre*.
  Il fatto che la media cambi e non resti costante per ogni sottoperiodo significa che c'è ***stagionalità*** nei dati. Possiamo anche capire che la serie è caratterizzata da variazioni che sono crescenti nel tempo. In particolare, in Q1 e Q3 abbiamo un forte trend crescente; anche Q4 presenta un trend crescente ma meno marcato rispetto a Q1 e Q3.
- Discorso opposto, invece, per Q2, che presenta un trend decrescente.

Grafico 2

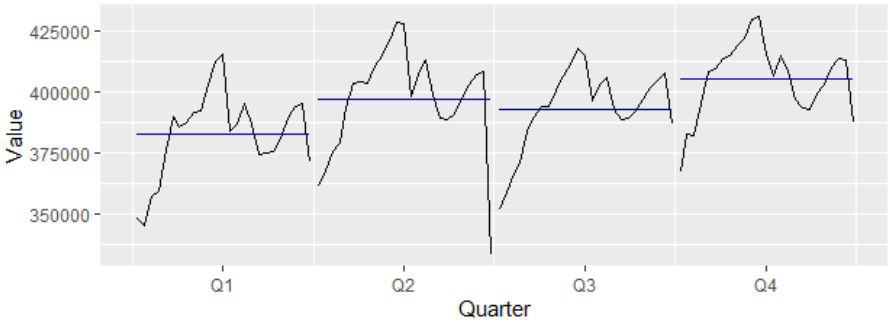
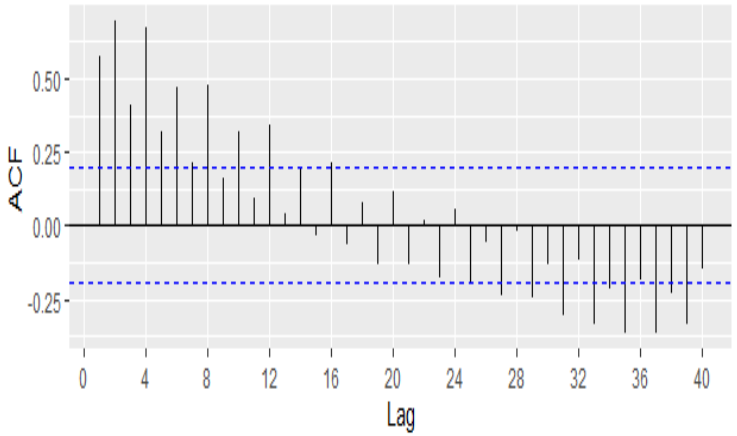
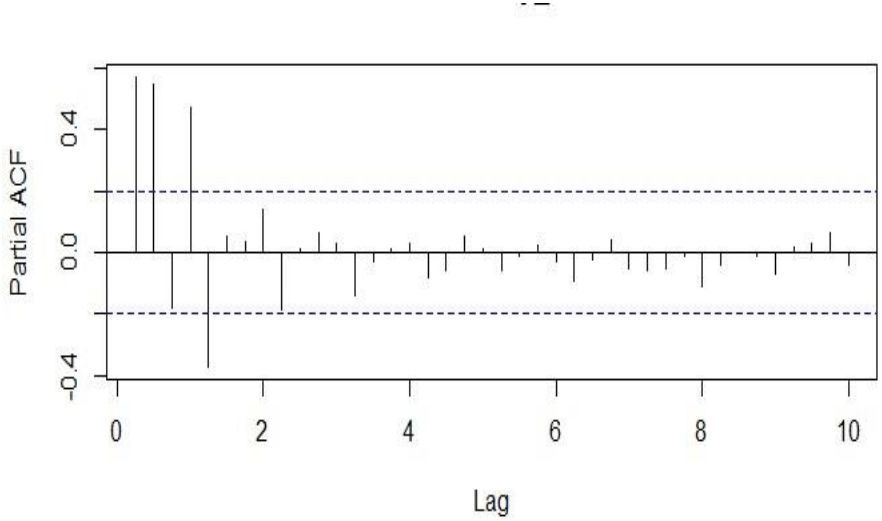


grafico3: ACF / PACF



- Il valore di ***autocorrelazione*** più alto lo osserviamo al ritardo due.
- I picchi tendono ad essere a due quarti di distanza l'uno dall'altro, quindi i punti di massimo tendono a ripetersi a due ritardi di distanza tra di loro.
- Il valore dell'autocorrelazione al ritardo zero è sempre pari a zero.
- La serie non è ***White Noise***, in quanto le autocorrelazioni sono più estreme rispetto alle bande di confidenza; infatti, per le serie white noise ci aspettiamo che *ogni autocorrelazione sia prossima allo zero*, ovviamente non esattamente uguale allo zero perchè c'è sempre una *componente d'errore* dovuta alla *variazione casuale*.



- Ancora, la serie conferma avere un trend. Infatti, le autocorrelazioni per i piccoli ritardi tendono ad assumere valori elevati e positivi, dato che le osservazioni vicine nel tempo sono vicine tra di loro anche per la dimensione.
- Quando i dati sono stagionali, invece, le autocorrelazioni sono più grandi per i multipli della frequenza stagionale.
- In questo caso, notiamo entrambi gli effetti; infatti, la diminuzione lenta dell'ACF all'aumentare del ritardo è dovuta al trend, mentre la forma ad onda è dovuta alla stagionalità.
- Infine, l'ACF mostra i picchi per il ritardo due, quattro, sei, otto e così via...questo indica una stagionalità di lunghezza due.
- Il correlogramma è utile anche per verificare se la serie storica è stazionaria o meno. In questo caso, l'ACF diminuisce lentamente e conferma quanto detto prima, ovvero la serie storica non è stazionaria.
  Il valore p^(1) è molto grande e positivo e quindi la tesi è confermata.

- Il grafico dell'***autocorrelazione parziale***, invece, propone un riepilogo della relazione tra un'osservazione in una serie temporale con osservazioni in fasi temporali precedenti con le relazioni delle osservazioni intermedie rimosse.

  <<L'autocorrelazione parziale al lag k è la correlazione che risulta dopo aver rimosso l'effetto di eventuali correlazioni dovute ai termini ai ritardi più brevi.>>
- Il grafico ***PACF*** ha dei picchi significativi solo per i primi ritardi, il che significa che tutte le autocorrelazioni di ordine superiore sono effettivamente spiegate dall'autocorrelazione lag-k.

- **Test per la Stazionarietà**

  La serie in questione non risulta essere, inizialmente, stazionaria e ciò lo si può vedere dal ***test di Dickey-Fuller***.

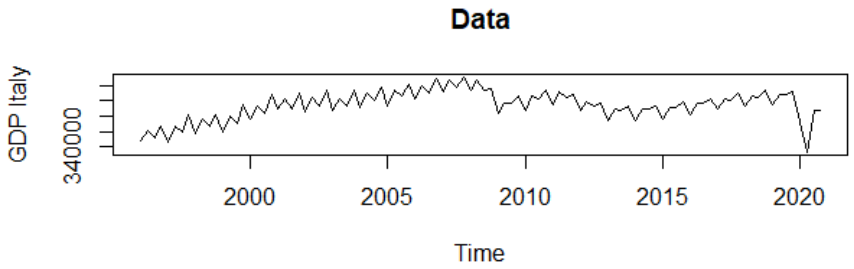
  Dal primo test, il valore associato al P-Value è maggiore del livello di alpha al 5% e quindi si rifiuta l'ipotesi nulla, concludendo che la serie non è ***stazionaria***.

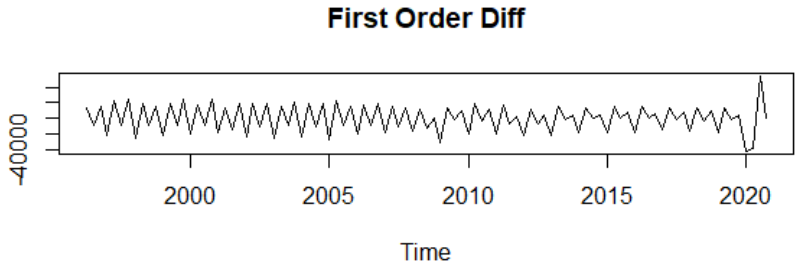
- Un modo per rendere stazionaria una serie temporale è calcolare le differenze tra osservazioni consecutive. Questo processo è noto come ***differenziazione***.

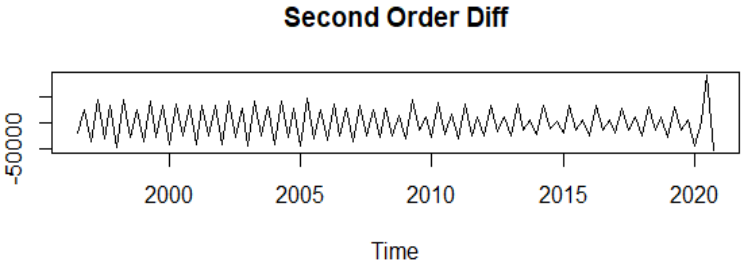
- Una ***trasformazione logaritmica*** può aiutare a stabilizzare la varianza della serie temporale. La differenza, invece, può aiutare a stabilizzare la media di una serie storica rimuovendo i cambiamenti di livello e quindi eliminando(o riducendo) la tendenza e la stagionalità.

- In questo caso, è stato necessario applicare una *differenza del primo e del secondo ordine* per rendere la serie stazionaria.

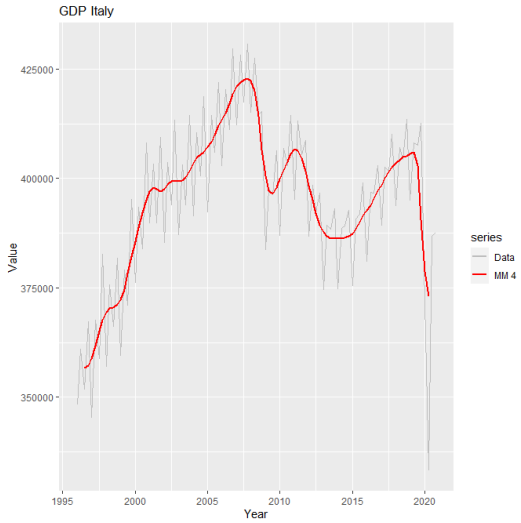
- Dal *grafico4* notiamo che la differenziazione di secondo ordine riesce a catturare l'improvviso crollo che si ha nel 2008.







- Spostando l'attenzione sulle componenti della serie, invece, si vuole catturare la componente di trend attraverso l'utilizzo di una semplice ***media mobile centrata***.
  Mettendo a confronto medie mobili di vario ordine, quella ottimale risulta essere una media centrata di ordine 4, i cui risultati sono anche simmetrici.



- La media mobile con questo ordine è in grado di riprodurre la tendenza di fondo della serie ed è anche in grado di catturare i cambiamenti di livello negli anni 2008, 2012 ed il crollo avvenuto nel 2020.
- L'ordine della media mobile determina il grado di fluidità della serie. Più l'ordine è grande e più la serie sarà liscia. Al contrario, più l'ordine è piccolo, più la serie seguirà da vicino i dati e sarà più irregolare.
- Le medie mobili sono di solito di ordine dispari, così da essere simmetriche e quindi lasciare lo stesso numero di valori mancanti sia per valori bassi che alti. Per rendere simmetrica una media mobile di ordine pari, è possibile applicare una media mobile alla media mobile. Ad esempio, si può prendere una media mobile di ordine quattro e poi applicare un'altra media mobile di ordine due ai risultati in modo da ottenere una serie simmetrica.
- In questo caso, è stato utilizzato il parametro center(in R) per centrare la media mobile, di conseguenza l'approccio sopra descritto rappresenta solo un'alternativa.

Decomposizione

- Un'altra tecnica utilizzata in questo lavoro è la ***decomposizione***, ovvero un processo che ha come obiettivo quello di decomporre la serie nelle sue componenti per capire quali di queste determinano maggiormente l'evoluzione del fenomeno nel tempo.
- ***Trend***: determina l'andamento di fondo o la tendenza di lungo periodo della serie.
- ***Stagionalità***: determina le fluttuazioni che si ripetono con regolarità nel tempo. Solitamente, questa componente ha una frequenza infra-annuale, in questo caso trimestrale.
- ***Residui***: tutto ciò che non si riesce a catturare con il modello.
- Nella decomposizione della serie, la componente ciclica e quella di trend vengono considerate insieme e per questo motivo si parla di componente di tendenza ciclica.

- La prima tecnica di decomposizione utilizzata è quella ***classica***, che può essere ***additiva*** o ***moltiplicativa***. Queste tecniche sono semplici da applicare ma hanno dei limiti, ovvero non forniscono una stima completa della componente di trend ed

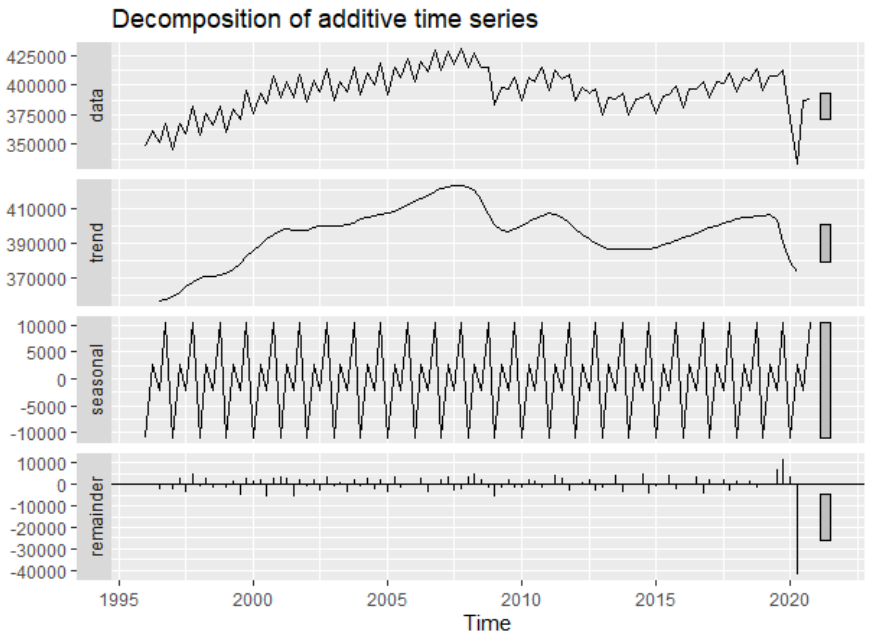
assumono che la componente stagionale sia fissa, ovviamente ciò può rappresentare un vincolo per le serie che hanno una componente stagionale variabile.

Decomposizione classica additiva

$$yt = St + Tt + Rt$$

si assume che la serie storica originale sia definita come ***somma*** delle *tre componenti*.

- Si articola nei seguenti step.
- **Step 1**: stimare la componente di tendenza ciclica con l'uso della media mobile centrata.
- **Step 2**: calcolare la serie "*detrendizzata*" e sottrarre la componente di tendenza ciclica alla serie originale.
- **Step 3**: stimare la componente stagionale sfruttando la serie "*dentrendizzata*". La componente stagionale per ogni stagione è stimata come media dei valori detrendizzati per quella stagione. In questo caso, con dati trimestrali, la componente stagionale di un trimestre è la media di tutti i valori detrendizzati di quel trimestre.
- **Step 4**: stimare la componente residua (R^t). Si può ottenere andando a sottrarre dalla serie originale la componente di trend/ciclica stimata e la componente stagionale.
- $Rt\hat{} = yt - Tt\hat{} - St\hat{}$



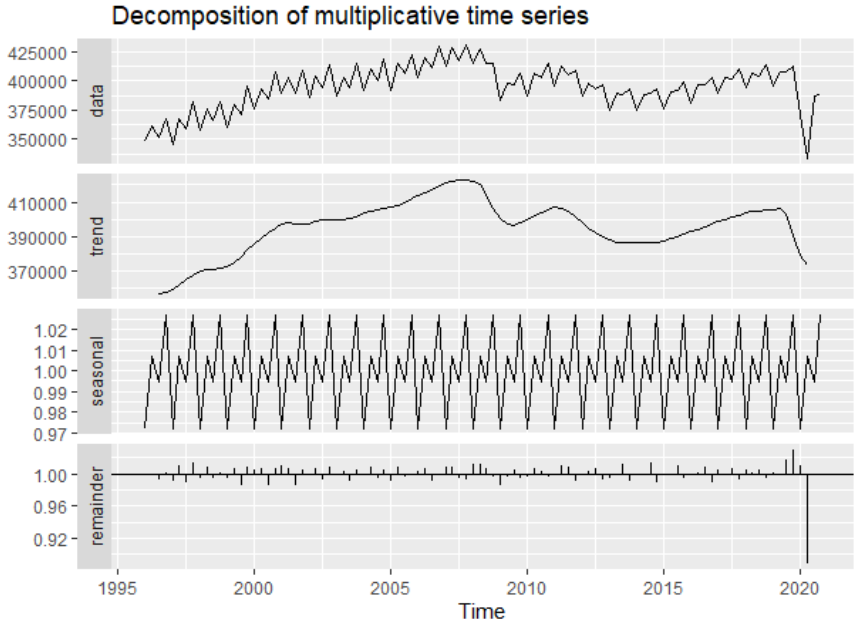
- La componente stagionale stimata presenta delle oscillazioni periodiche e regolari con ampiezza costante; ovvero la stagionalità è costante di anno in anno. Questo rispecchia l'assunzione sulla stagionalità alla base della decomposizione classica, la quale appunto ipotizza che la componente stagionale sia costante nel tempo(di anno in anno).
- La barra grigia mostra l'ampiezza relativa agli effetti. Più è grande e più la variazione della componente sarà piccola rispetto alla serie osservata. I grafici rappresentano rispettivamente la serie storica osservata, la componente di trend, la componente stagionale ed infine quella residua. La componente più influente è quella di tendenza ciclica, seguita dalla componente residua e da quella stagionale.

Decomposizione classica moltiplicativa

$$yt = St * Tt * Rt$$

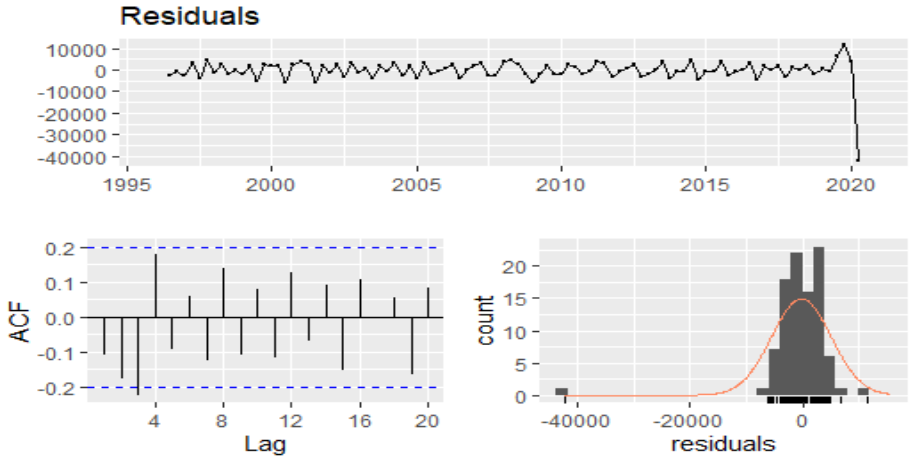
si assume che la serie storica originale sia definita come ***prodotto*** delle *tre componenti*.

- Gli step da seguire sono gli stessi, eccetto l'ultimo.
- **Step 4**: la componente residua(R^t) viene stimata dividendo la serie originale per il prodotto tra la componente di trend/ciclica e la componente stagionale.
- $Rt\hat{} = yt \ / \ (Tt\hat{} * St\hat{})$



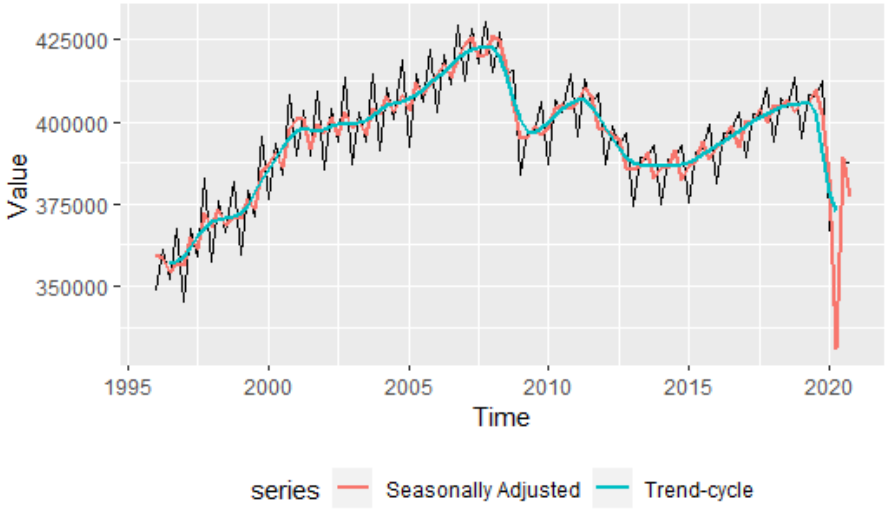
Residui

Decomposizione Additiva



- La componente residua risulta più o meno costante intorno allo zero, ad esclusione del problema legato all'anno 2020. I residui non sono correlati tra loro e si avvicinano alla distribuzione normale eccetto per delle osservazioni molto estreme nella coda sinistra.

Confronto tra la serie destagionalizzata e la componente di tendenza ciclica stimata con il metodo additivo

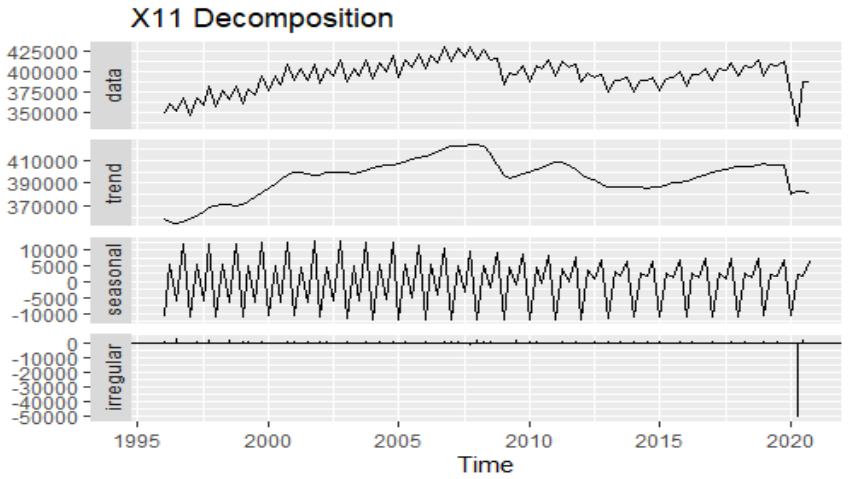


- Si vede come la componente di ***tendenza ciclica*** stimata con il metodo additivo catturi abbastanza bene le variazioni improvvise(al rialzo o al ribasso). Tuttavia, la componente sembra essere troppo ***smooth***. Proviamo a migliorare le prestazioni utilizzando un modello più complesso.

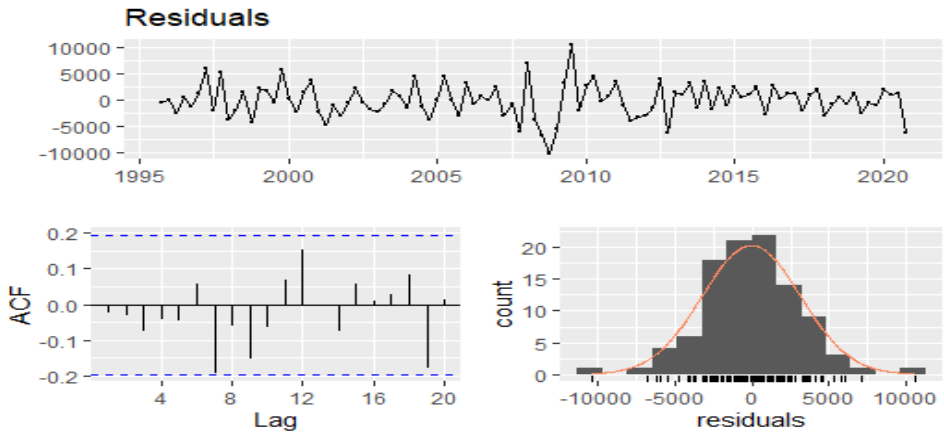
Decomposizione X11

- Metodo fortemente utilizzato per scomporre serie storiche con stagionalità mensile o trimestrale.
- Si basa sulla decomposizione classica ma richiede più passaggi e metodi più complessi per superare i limiti sopra elencati della decomposizione classica.

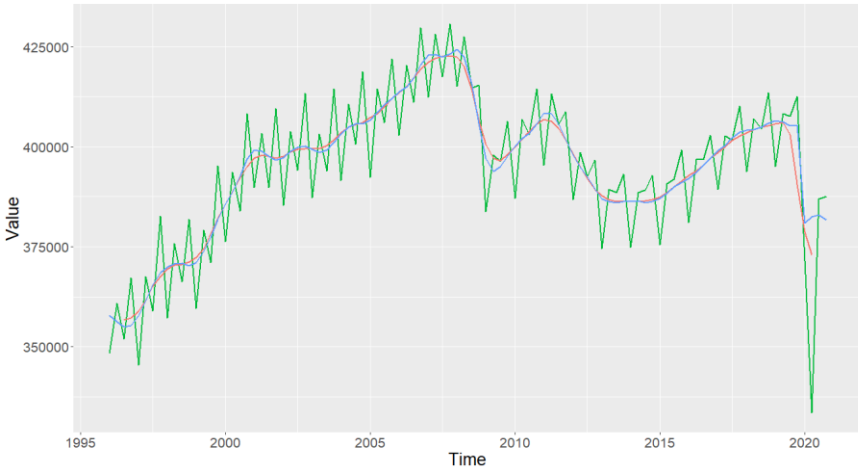
- Questa tecnica fornisce una ***stima completa*** della componente di ***tendenza ciclica*** e quindi si hanno a disposizione anche le osservazioni nella parte iniziale e finale del trend.
- Inoltre, il metodo X11 riesce a gestire la presenza di ***valori anomali*** o variazioni legate ad altri effetti(es. effetti di calendario).



Residui Decomposizione X11

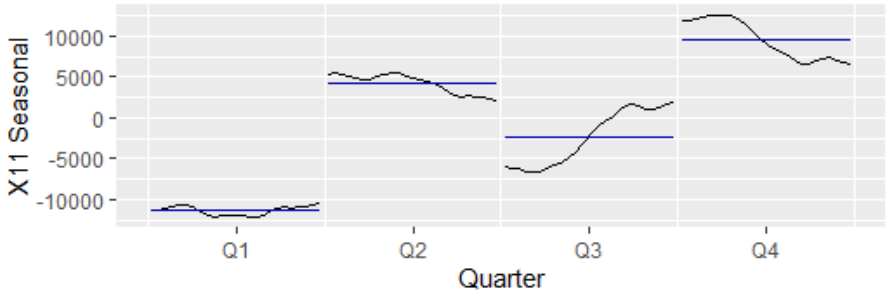


Componente di tendenza/ciclica stimata (Additiva Vs. X11)

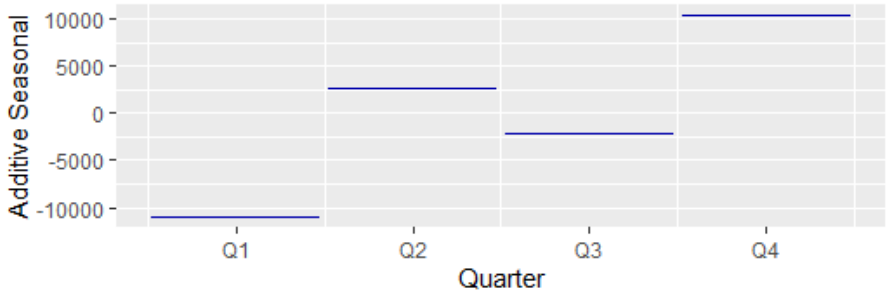


- La componente di tendenza ciclica con il metodo ***X11*** fornisce delle ***informazioni complete*** sul comportamento della serie, sia nella parte iniziale che finale, mentre questo non avviene con il metodo classico.
- Altra differenza si nota nelle fasi in cui abbiamo ***variazioni improvvise***.
  Guardando al rialzo tra il 2000 ed il 2003, infatti, la componente stimata con il metodo X11 riesce a seguire meglio questa fase al rialzo e quindi è meno smooth rispetto al metodo additivo. Stesso discorso per il picco tra il 2006 ed il 2007, ma anche per la caduta che si ha nell'anno 2008.

- Tuttavia, il metodo X11 prevede che la stagionalità cambi nel tempo, come possiamo vedere dal seguente grafico.

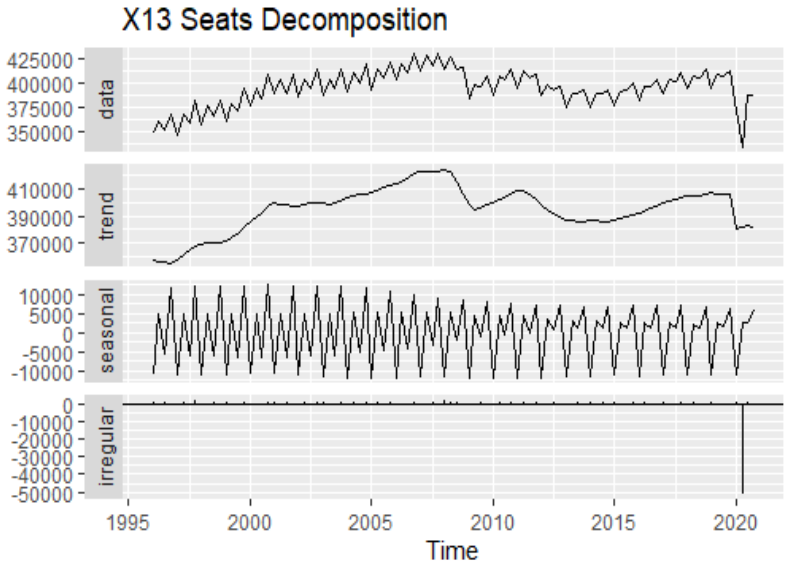


- Ciò non accade con la decomposizione additiva.



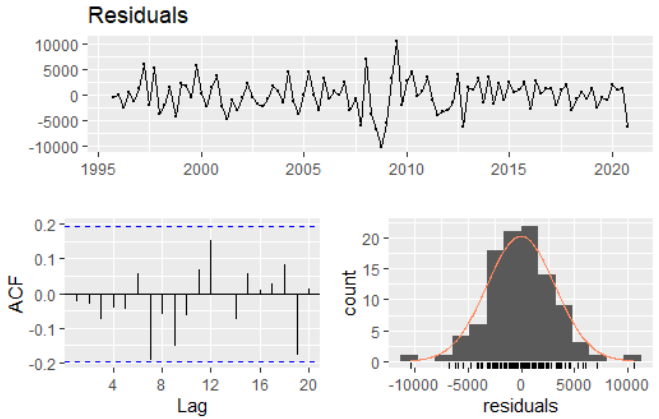
Decomposizione SEATS(Seasonal Extraction in ARIMA Time Series) - X13

- Come per il metodo X11, anche il metodo X13 SEATS funziona solo con dati trimestrali e mensili.

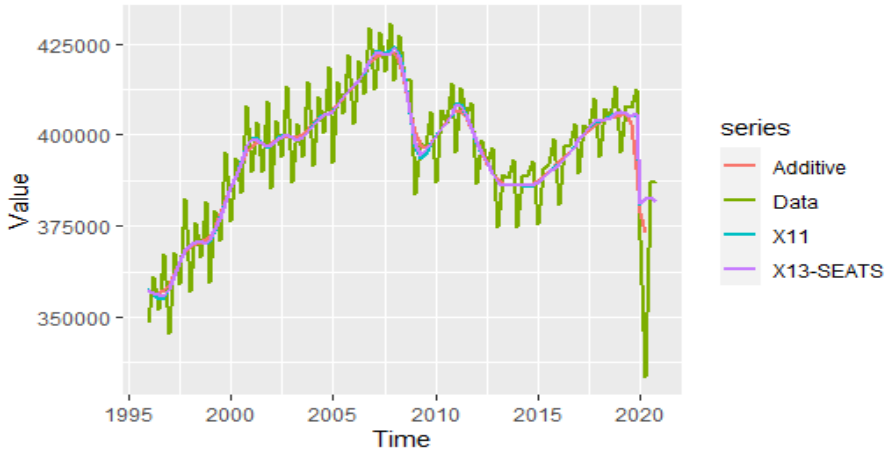


- Le *differenze* tra la decomposizione X11 e X13 sono molto *piccole*.
- Si nota come la differenza nella stima della componente di tendenza ciclica sia piccolissima. L'unica differenza che si nota, ma minima, è che la componente di tendenza ciclica ottenuta con il metodo X13 è leggermente più smooth, in alcuni punti, rispetto a quella ottenuta con il metodo X11.

Residui



Componente di tendenza ciclica (Additivo Vs. X11 Vs. X13)



Autocorrelazione

- L'autocorrelazione è un indice indipendente dall'unità di misura ed in grado di misurare la ***forza di legami lineari esistenti nella serie***, in particolare misura la forza della dipendenza lineare tra xp e xp+h, ovvero osservazioni della stessa serie ma rilevate in diversi istanti temporali. In questo studio il valore dell'autocorrelazione è pari a 0,669.

Forza del Trend e della Stagionalità

- Risulta utile, inoltre, calcolare la ***forza del trend e la forza della stagionalità***, che si presentano nella seguente forma:

$$\boldsymbol{Ft} = \max\left(0; 1 - \left(\frac{Var(Rt)}{Var(Rt + Tt)}\right); con\ 0 \leq Ft \leq 1\right.$$

- Ft vicino a 0 indica un trend debole.
- Ft vicino ad 1 indica un trend forte.

$$\boldsymbol{Fs} = \max\ \left(0; 1 - \left(\frac{Var(Rt)}{Var(Rt + St)}\right); con\ 0 \leq< Fs \leq 1\right.$$

- Fs vicino a 0 indica una stagionalità debole.
- Fs vicino ad 1 indica una stagionalità forte.

- In questo studio: **Ft = 0,901** e **Fs = 0,685**; quindi una forza della componente di trend abbastanza importante che riesce a spiegare buona parte del comportamento della serie ed una forza della stagionalità importante ma minore rispetto al trend.

Previsione

- Considerando i dati presi in esami, ovvero una serie *univariata* senza alcuna variabile da poter utilizzare come *regressore* in un modello, ci si può avvalere di alcuni *semplici metodi di previsione* che si basano sulle componenti contenute nella serie.
  Si farà, di seguito, un cenno all'approccio teorico dei diversi metodi per poi confrontare le prestazioni.

- ***Average Method***
  Con questo metodo si assume che tutti i valori futuri della serie siano *uguali alla media dei valori osservati* nella serie storica.
  Le previsioni saranno uguali a:

$$y\hat{}T + h \mid T = \frac{(y1 + y2 + \cdots + yT)}{T}$$

- ***Naive Method***
  Con questo metodo si impone semplicemente che tutte le previsioni assumano il *valore dell'ultima osservazione registrata*.

$$y\hat{}T + h \mid T = y^T$$

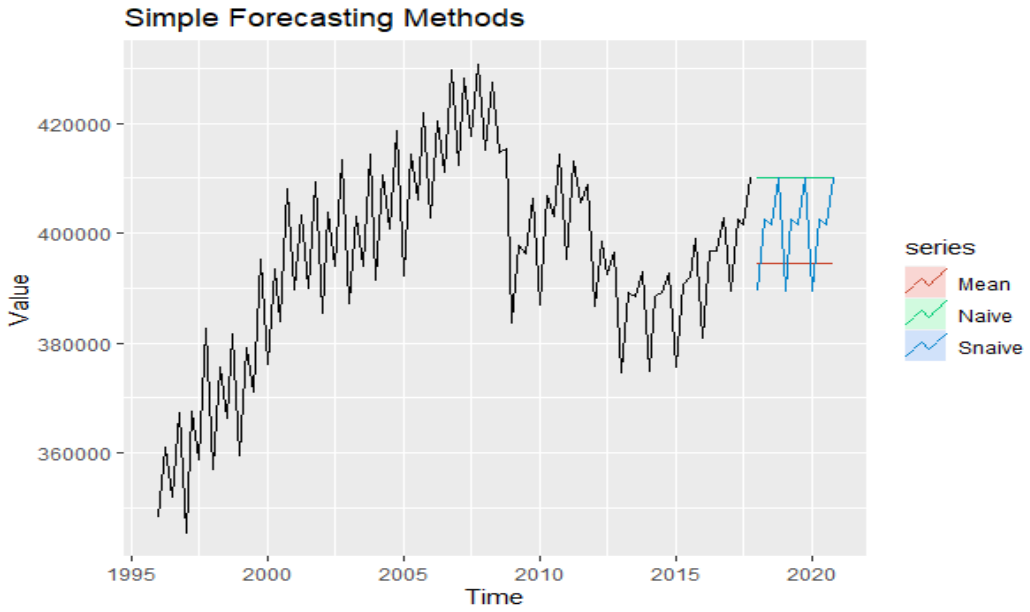
- ***Seasonal Naive Method***
  Questo metodo può essere utile per dati che presentano stagionalità. Si imposta che ogni previsione sia uguale all'*ultimo valore osservato nella stessa stagione* e ciò differenzia le previsioni ottenute da quelle Naive.
  Le previsioni saranno uguali a:

$$y\hat{}T + h \mid T = yT + h - m(k + 1)$$

- m è il periodo stagionale
- k è il numero di anni completi nel periodo di previsione.

Risultati e confronto

- Il metodo ***Average*** così come il metodo ***Naive***, si dimostrano troppo semplici e non riescono a catturare la stagionalità che è presente nei dati.
- Il metodo ***Seasonal Naive***, invece, riesce a catturare la componente stagionale presente nella serie storica.



Evaluation Metrics

	Mean	Naive	Snaive
MAE	**15505.21**	**17606.67**	**13260.31**
RMSE	**21732.47**	**27336.89**	**22346.01**
MAPE	**4.13**	**4.81**	**3.64**

- Considerando i seguenti indici, il miglior modello è il ***Seasonal Naive*** che minimizza il ***MAE*** ed il ***MAPE***.
- Discorso diverso se si considera il ***RMSE***, che viene minimizzato dal modello ***Mean***.
- Tuttavia, i residui che più si avvicinano alla distribuzione Normale sono quelli del metodo ***Seasonal***, anche se con una coda sinistra più pesante.

***** *le previsioni sono state fatte tenendo conto della divisione tra training e test set, con il test che conta le osservazioni degli ultimi tre anni, quindi non considerando l'improvviso crollo dell'anno 2020.*