UNIVERSITÀ DI PISA

PROGETTO DI STATISTICA
PARTE III

Corso di Laurea Magistrale in Artificial Intelligence and Data Engineering
Cocchella Ludovica | matricola: 532189

Sommario

1. Introduzione al caso di studio e obiettivo dell'analisi

Il Ministero del Lavoro e delle Politiche Sociali, che come noto si occupa in particolar modo dei temi che riguardano lo sviluppo occupazionale e della tutela del lavoro, ha come obiettivo lo studio dell'andamento dell'occupazione femminile in Italia e gli effetti che la pandemia di COVID-19 ha avuto su di essa.

Quindi tale Ministero ha la necessità di conoscere gli effetti che la pandemia ha avuto e che avrà nei mesi successivi sull'occupazione femminile per poter eventualmente programmare varie azioni a favore dell'occupazione stessa.

1.1. Reperimento dei dati

La serie storica che verrà analizzata è stata reperita dalla banca dati ISTAT ed è costituita dai dati riguardanti l'occupazione femminile in Italia. A partire dal link http://dati.istat.it/ è stato selezionato dall'elenco dei vari temi le seguenti sezioni nel seguente ordine "Lavoro e retribuzioni", "Offerta di lavoro", "Occupazione", "Occupati – dati mensili", "Sesso, età". Dopo aver selezionato le precedenti sezioni viene fornita una tabella dalla quale, tramite l'opzione "Personalizza" è stato selezionato come genere il sesso femminile e come periodo è stato selezionato il periodo che va da gennaio 2004 a dicembre 2020.

2. Elaborazione dei dati

La nostra analisi è stata effettuata tramite il supporto del software R e di seguito riportiamo tutti i vari step e comandi necessari per il caricamento della tabella sul software.

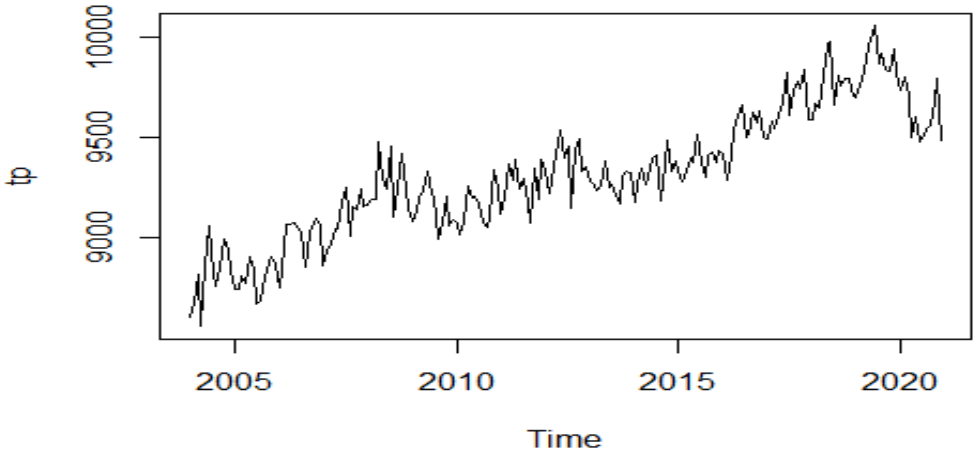
Prima di tutto è necessario impostare la directory da cui importare la tabella tramite il comando: `>setwd()`

Dopo di che carichiamo la tabella tramite il comando: `>tab=read.csv2("Tabella.csv", header=F)`

Prendiamo la colonna che ci interessa e la reinterpretiamo come serie storica indicando la data d'inizio e la frequenza tramite il seguente comando: `>tp=ts(data,frequency=12,start=c(2004,1))`

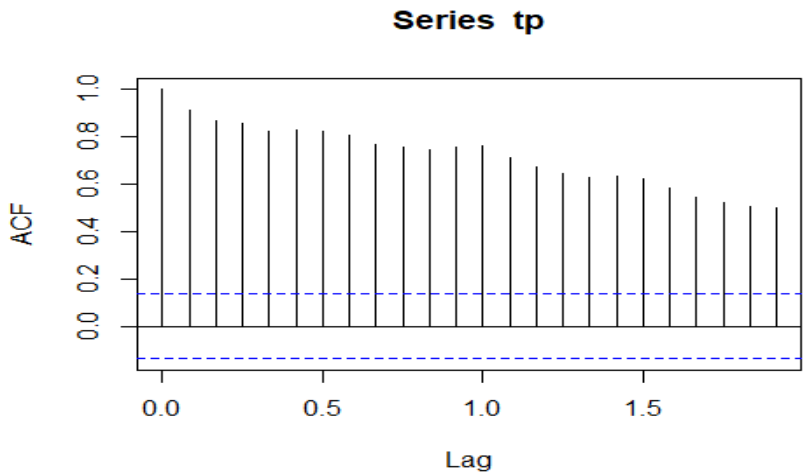
2.1 Analisi della struttura della serie storica

Come primo passo risulta fondamentale andare a comprendere la natura della serie storica al fine del suo utilizzo come strumento di previsione. Una prima analisi può essere effettuata andando a visualizzare graficamente l'andamento della serie attraverso il plot dei dati:

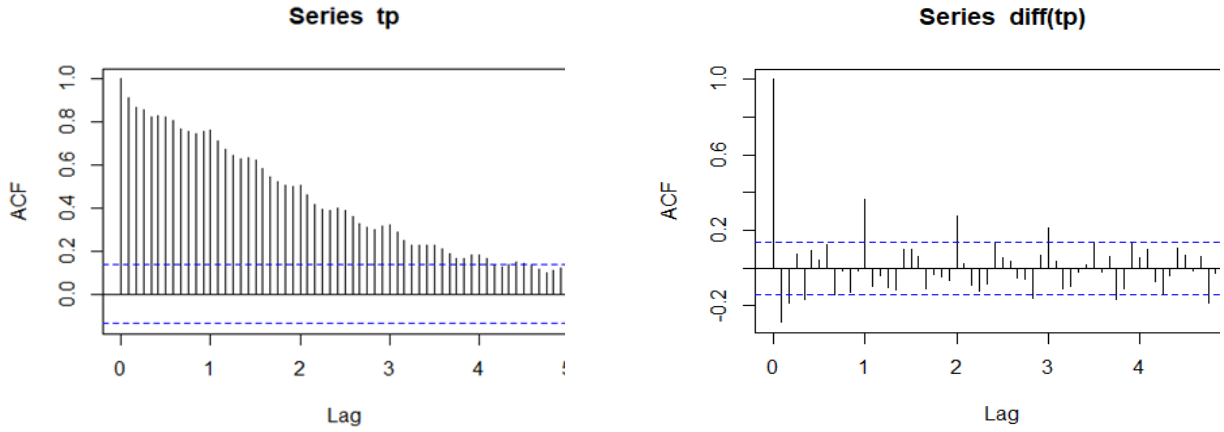


Il nostro obiettivo adesso è capire se questa serie presenta delle stagionalità. Dal grafico precedente possiamo apparentemente notare un trend ascendente che sembra prevalere su una stagionalità non molto evidente.

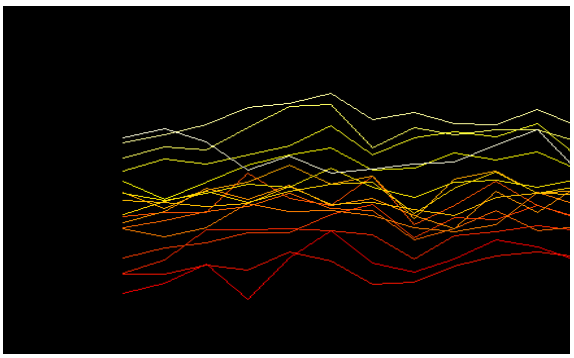
Visualizziamo e studiamo la funzione di autocorrelazione di questi dati:

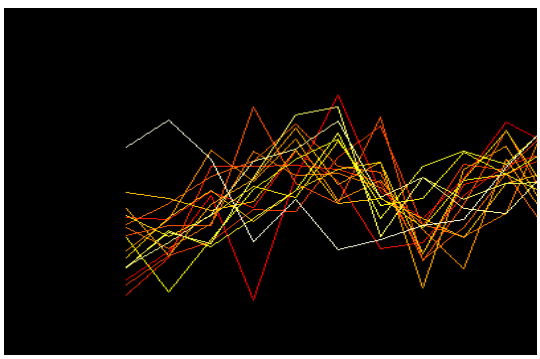


Possiamo confermare quanto affermato in precedenza infatti notiamo come sia chiaramente presente un trend e notiamo anche delle piccole variabilità che potrebbero segnalarci una leggera periodicità anche se non risultano particolarmente evidenti. Pertanto, dobbiamo cercare di capire se queste piccole variabilità sono una stagionalità nascosta oppure sono variazioni legate al fatto che la funzione di autocorrelazione non filtra perfettamente il rumore. Continuiamo ad indagare sulla presenza di stagionalità impostando lag più lunghi nella funzione di autocorrelazione e osservando la serie al netto del trend:

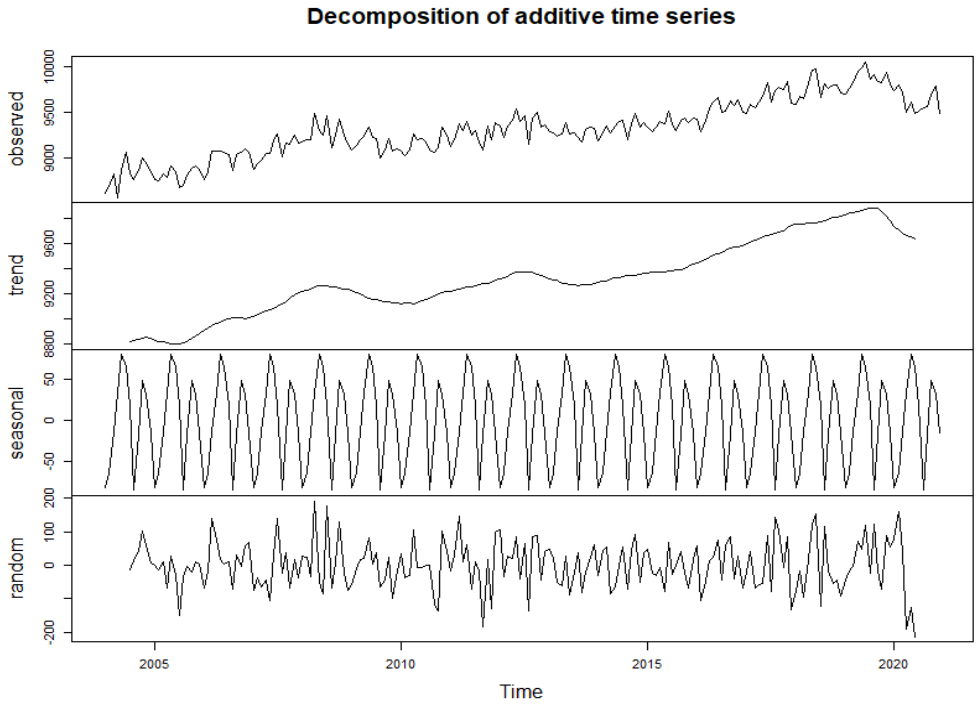


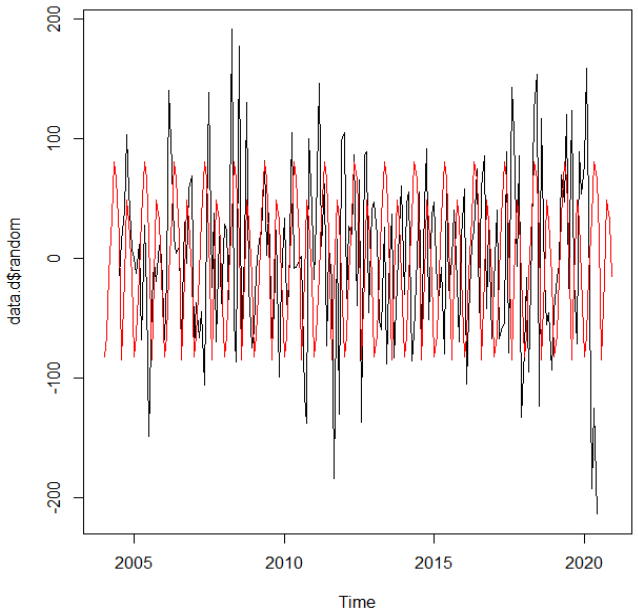
Notiamo dei leggeri picchi ma niente di troppo significativo. Ciò che possiamo dire con certezza è la presenza di un trend dominante ma non siamo ancora in grado di affermare con certezza l'assenza o la presenza di stagionalità pertanto continuiamo la nostra indagine. Rappresentiamo tutti gli anni insieme allineati sullo stesso grafico per vedere se anno dopo anno c'è una qualche sorta di ripetizione dei dati:

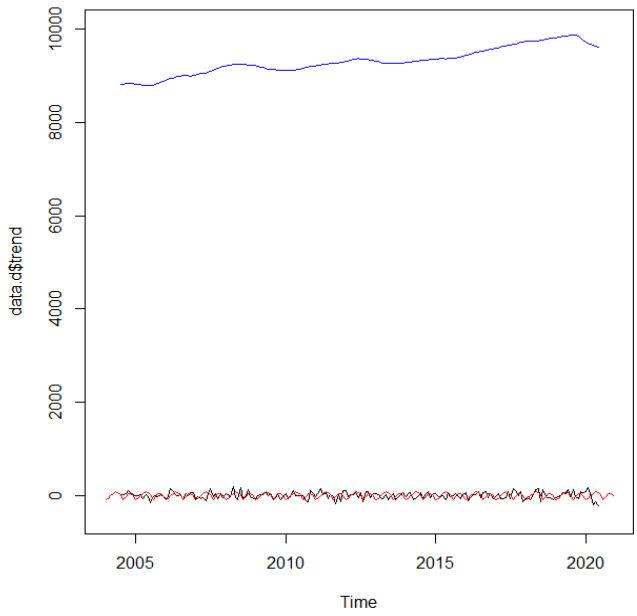




Non osserviamo una somiglianza significativa e quindi sembrerebbe non esserci stagionalità. Per poter ottenere un'ulteriore conferma vediamo cosa otteniamo dalla decomposizione:







Esaminiamo gli ordini di grandezza delle componenti soprattutto quello relativo alla componente stagionale. Notiamo che stagionalità e rumore hanno ordini di grandezza comparabili ma nettamente inferiori rispetto al trend. Quest'ultima osservazione sommata a quelle effettuate in precedenza ci portano ad affermare che ci troviamo di fronte ad una serie con trend e priva di stagionalità.

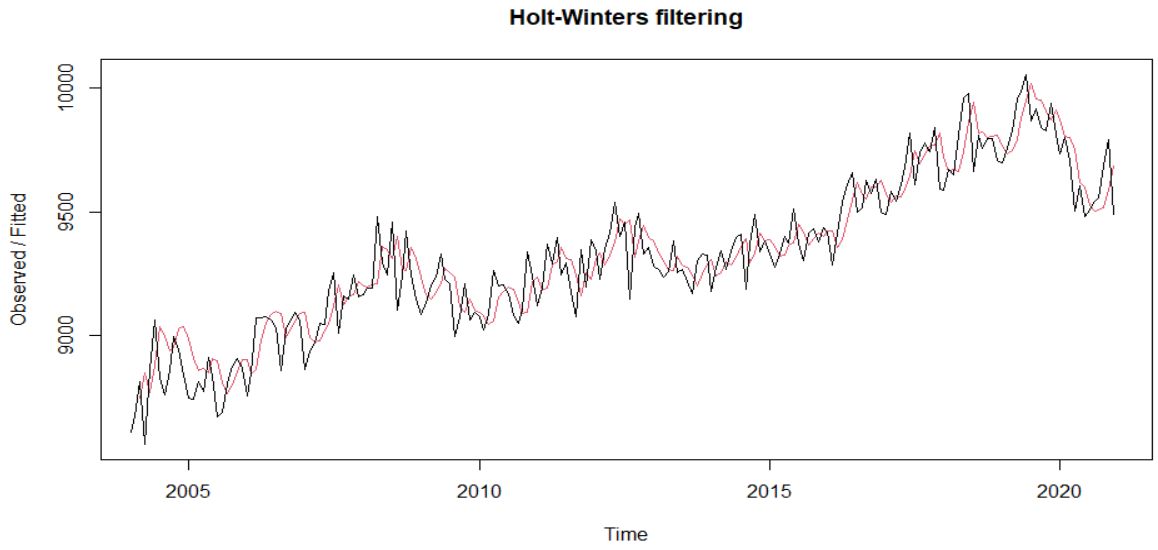
2.2 Metodi di analisi e previsione

Dopo aver compreso la natura della serie in questione possiamo passare ai vari metodi di analisi e previsione. Sceglieremo i metodi che risultano essere più adatti sulla base di quello che riteniamo essere la struttura

della serie. Dopo di che confronteremo i risultati ottenuti dai vari metodi per capire quale di questi risulti essere il metodo migliore ovvero quello che ci fornirà delle previsioni più attendibili.

2.2.1 Holt-Winters

Essendo di fronte ad una serie caratterizzata da un trend e priva di stagionalità la scelta del metodo di Holt-Winters da utilizzare ricade sul metodo di *smorzamento esponenziale con trend* . Vediamolo graficamente (in nero la serie originale e in rosso il metodo):



Vediamo i valori dei coefficienti *alpha* e *beta* assegnati automaticamente dal software:

```
> tp.set$alpha
    alpha
0.4822335
> tp.set$beta
      beta
0.06988877
```

Dal grafico possiamo notare come il metodo segue con una buona approssimazione la serie originale. In questo caso però noi sappiamo, grazie all'analisi preliminare sulla struttura della serie, che l'analisi di interesse proviene dal trend. Per questo motivo noi vogliamo un modello che riesca a catturare il trend e non segua le fluttuazioni poiché queste sono causate dal rumore. In sintesi, noi vogliamo quindi che il metodo ci restituisca un andamento che segua meno il rumore. Conseguentemente, occorre migliorare l'andamento andando a modificare i parametri. A questo punto non ci resta che esplorare diversi valori dei parametri per decidere qual è quello più adeguato:

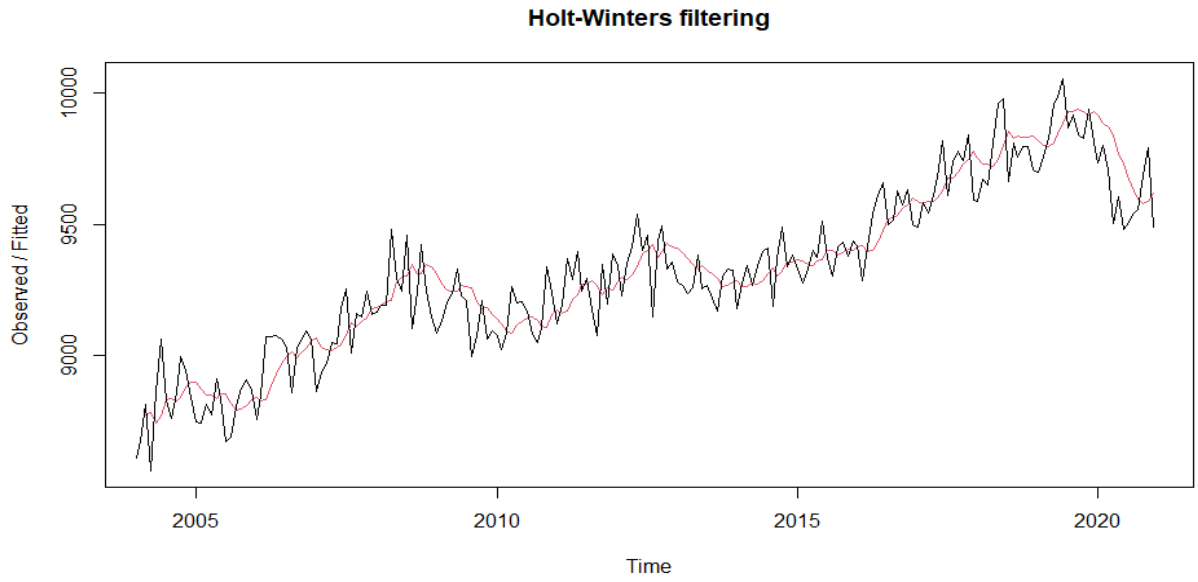
```
> for(alpha in 1:9){
+   for(beta in 1:9){
+     plot(HoltWinters(tp,alpha=alpha/10,beta=beta/10,gamma=
F),xlab=paste("alpha=",alpha/10," - beta",beta/10))
+   }
+ }
```

Quello che vogliamo ottenere è una seria che segua meno le oscillazioni quindi per ottenere ciò, essendo già il valore di *beta* molto basso, dovremmo abbassare il valore di *alpha*. In aggiunta all'esplorazione dei valori dei coefficienti abbiamo anche la possibilità di variare la condizione iniziale.

Valutando tutti i grafici ottenuti vediamo che la rappresentazione più adeguata ci viene data settando i coefficienti come segue:

```
> plot(HoltWinters(tp,alpha = 0.2,beta = tp.set$beta,gamma=F,l.start=8768.5,b.start=3.378))
```

Dal comando precedente otteniamo il seguente grafico:

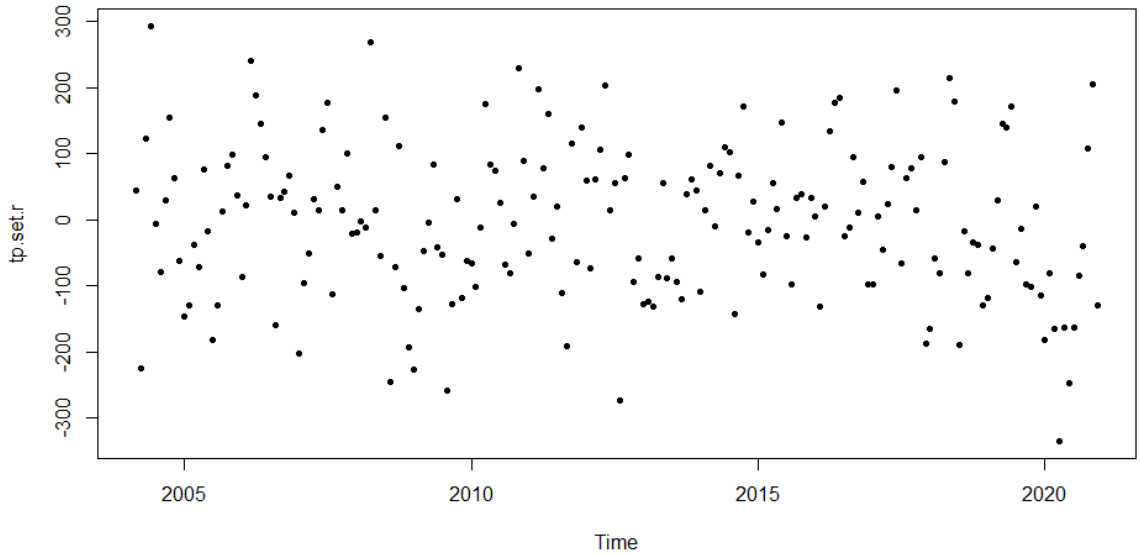


Vediamo infatti come in questo caso l'andamento tende a non seguire troppo le oscillazioni.

Non ci resta adesso che valutare la bontà del modello. Andiamo ad eseguire un'analisi dei residui al fine di stimare l'incertezza relativa alle previsioni future.

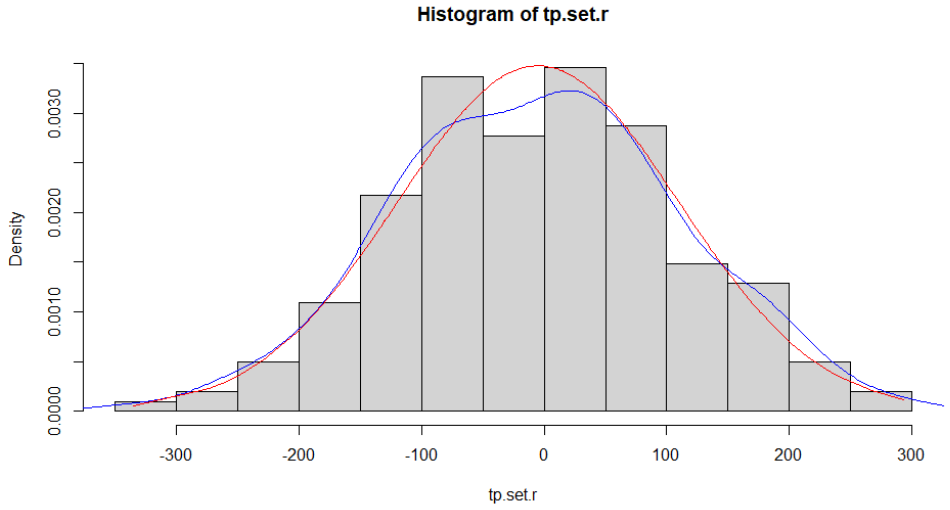
```
> tp.set.r=residuals(tp.set)
> 1-var(tp.set.r)/var(window(tp,c(2005,1)))
[1] 0.8501397
```

L'output della varianza risulta essere abbastanza alto quindi possiamo affermare che il metodo spiega bene la variabilità del problema. Andiamo a visualizzare i residui del metodo al fine di verificare la loro somiglianza a numeri casuali:

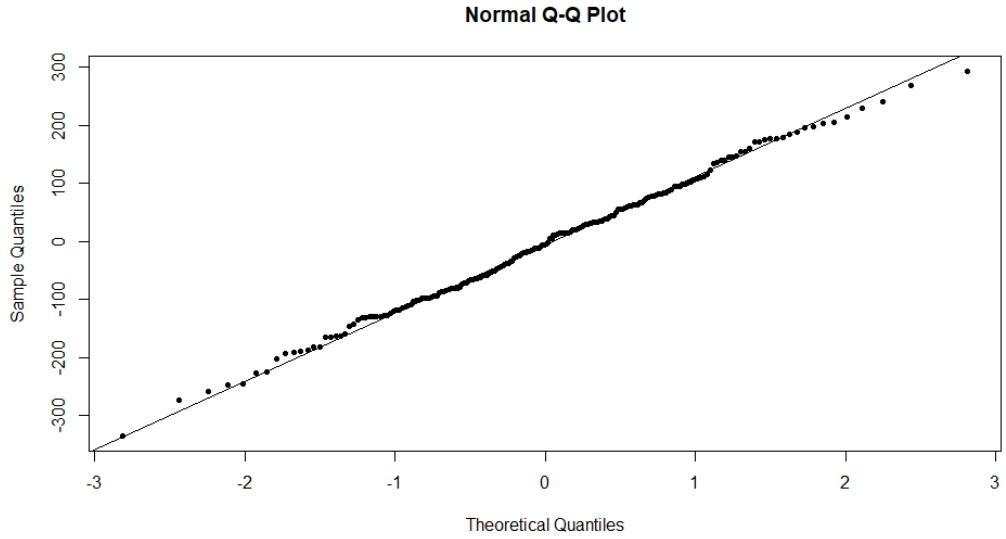


I punti sembrano essere abbastanza dispersi nel piano e ciò fa pensare a una distribuzione gaussiana.

Visualizziamo l'istogramma dei residui con la curva gaussiana:



Possiamo vedere come siano abbastanza simili. Osserviamo anche il grafico dei quantili:



Anche il grafico dei quantili sembra confermarci il fatto che i residui sono gaussiani, infatti notiamo una somiglianza elevata eccetto agli estremi. Andiamo ad effettuare un'ultima verifica con il test statistico *Shapiro-Wilk:*

```
> shapiro.test(tp.set.r)

        Shapiro-Wilk normality test

data:  tp.set.r
W = 0.99731, p-value = 0.9814
```

Otteniamo un risultato ottimo che ci permette di concludere l'analisi dei residui affermando che i residui seguono una distribuzione gaussiana.

Per valutare ulteriormente la bontà del modello andiamo a testare la capacità previsiva del metodo di smorzamento esponenziale con trend. In assenza di dati futuri già noti andiamo a prevedere valori che noi già conosciamo, tramite il test ed il training set con l'idea di fondo che se il modello è in grado di prevedere bene questi valori allora sarà in grado di prevedere bene anche quelli futuri.

```
> train=window(tp,end=c(2019,12))
> test=window(tp,2020)
> tp_t = HoltWinters(train,alpha = 0.2,beta = tp.set$beta,gamma=F,l.start=8768.5,b.start=3.37
8)
> tp_p=predict(tp_t,12)
> sqrt(mean((tp_p-test)^2))
[1] 359.7716
```

Vediamo che il metodo offre delle buone previsioni che risultano essere abbastanza vicine a quelle reali.

2.2.2 Autoregressione con il metodo dei minimi quadrati

Fra i metodi autoregressivi scegliamo di utilizzare il metodo dei minimi quadrati poichè la serie che stiamo analizzando risulta essere una serie non stazionaria. Infatti questo metodo rispetto agli altri metodi autoregressivi restituisce risultati più affidabili in caso di serie non stazionaria.

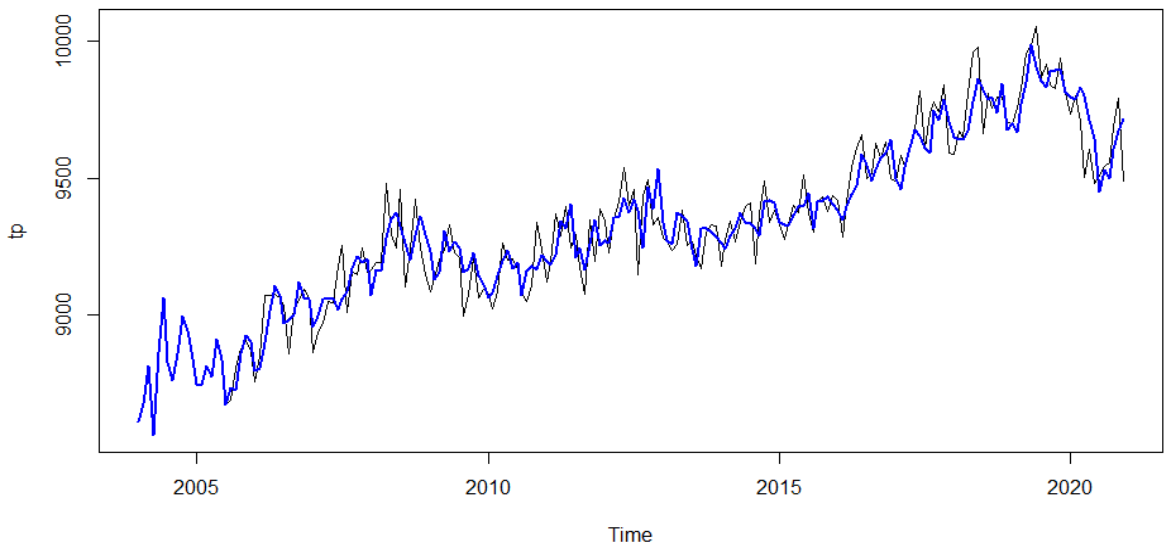
```
> tp.ls = ar(tp,method="ols")
> tp.ls$order
[1] 19
```

Il metodo prende come riferimento i 19 mesi precedenti.

```
> 1-tp.ls$var/var(tp[20:204])
[1] 0.9002073
```

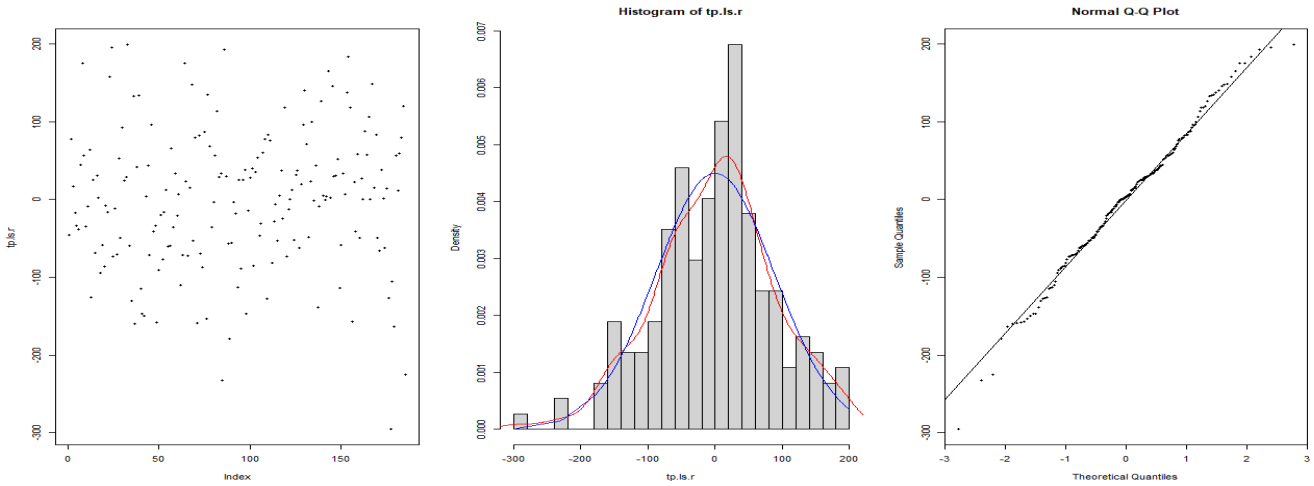
La varianza risulta avere un valore buono. Rappresentiamo graficamente quanto ottenuto con i seguenti comandi:

```
> o = tp.ls$order
> a = tp.ls$ar
> b = tp.ls$x.intercept
> tp.ls.an = tp
> for (i in (o + 1):length(tp)) {
+   tp.ls.an[i] = sum(rev(a) * tp[(i - o):(i - 1)]) + mean(tp) * (1 - sum(a)) +
+     b
+ }
> plot(tp)
> lines(tp.ls.an, col = "blue", lwd = 2)
```



In questo caso il metodo fornisce un andamento che tende a seguire le oscillazioni che non è l'ideale nel nostro caso. Infatti, nella fase di analisi della struttura della nostra serie questa risulta essere una serie con trend e senza stagionalità. Quindi noi abbiamo bisogno di un metodo che riesca a catturare il trend e a seguire il meno possibile le fluttuazioni.

Verifichiamo la bontà del modello tramite l'analisi dei residui:



Anche in questo caso è possibile affermare che i residui seguono una distribuzione gaussiana e ciò ci viene ulteriormente confermato dal test statistico *Shapiro-Wilk:*

```
> shapiro.test(tp.ls.r)

        Shapiro-Wilk normality test

data:  tp.ls.r
W = 0.99171, p-value = 0.3704
```

Per ottenere ulteriori informazioni riguardanti la bontà del modello procediamo con l'autovalidazione:

```
> train=window(tp,end=c(2019,12))
> test=window(tp,2020)
> tp_t = ar(train,method="ols")
> tp_p=predict(tp_t,n.ahead=12,se.fit=FALSE)
> sqrt(mean((tp_p-test)^2))
[1] 326.7068
```

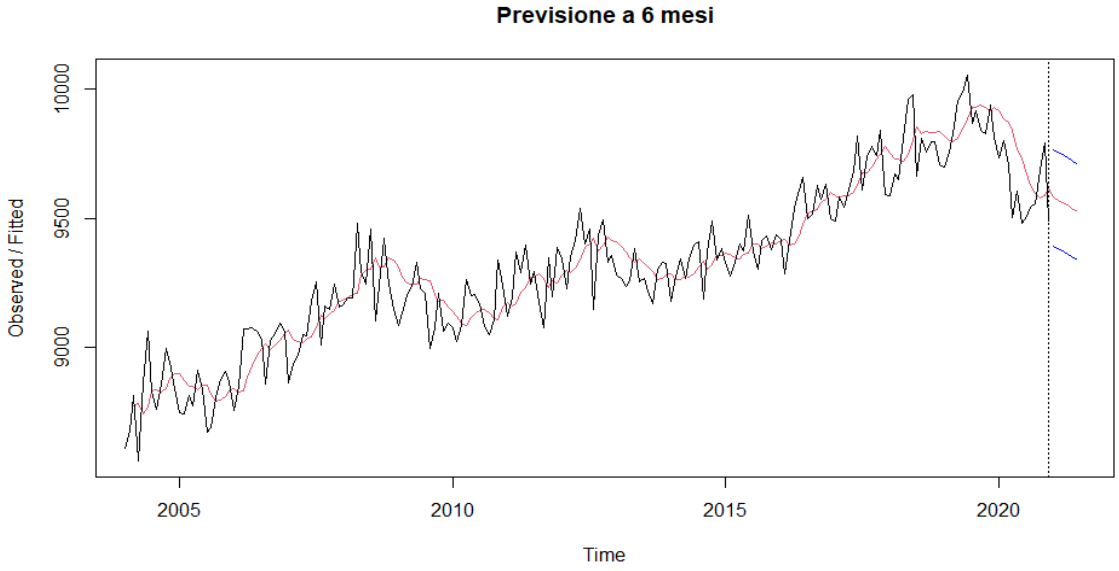
La previsione sembra essere molto vicina ai valori reali.

2.3 Scelta del metodo e previsione futura

È necessario a questo punto effettuare un confronto tra i due metodi presi in considerazione per poter decidere quale fra questi sia più giusto utilizzare per andare ad eseguire la previsione per i prossimi sei mesi. Il metodo viene scelto sulla base dei risultati dell'autovalidazione e dei residui. Possiamo affermare che nel nostro caso il compromesso migliore ci viene offerto dal metodo esponenziale con trend poiché i risultati dell'autovalidazione risultano essere paragonabili per entrambi i metodi mentre l'analisi dei residui risulta aver restituito risultati migliori rispetto al metodo dei minimi quadrati.

Avendo scelto il modello passiamo alla previsione futura:

```
> plot(tp.set,predict(tp.set,6),main="Previsione a 6 mesi")
> lines(predict(tp.set,6)+quantile(tp.set.r,0.05),col="blue")
> lines(predict(tp.set,6)+quantile(tp.set.r,0.95),col="blue")
> predict(tp.set,6)
          Jan      Feb      Mar      Apr      May      Jun
2021 9581.757 9571.155 9560.553 9549.950 9539.348 9528.745
```



3 Conclusioni

Dai risultati che abbiamo ottenuto possiamo concludere che negli anni l'ascesa in campo delle donne nel mondo del lavoro risulta essere di anno in anno sempre maggiore. L'analisi che abbiamo effettuato ci fa notare però come la pandemia abbia colpito notevolmente l'occupazione femminile, in particolare nel primo lockdown. Per quanto riguarda il 2021 possiamo vedere come la previsione che ci viene restituita dal metodo risulta avere un andamento decrescente che evidenzia che la situazione pandemica fa ancora molto preoccupare e che a causa delle misure restrittive ci vorrà ancora del tempo affinché la situazione lavorativa dell'occupazione femminile possa risollevarsi e ritornare ai numeri pre-covid.