Progetto di Statistica II

Docente: Prof. F.Flandoli

A.A. 2015-2016

Carmine Frascella

12 dicembre 2015

Indice

Capitolo 1

Vendite di automobili negli USA

1.1 Introduzione alla serie storica

Essendo un appassionato di corse automobilistiche e di automobili in generale, non potevo evitare di spulciare tra i dati relativi alle vendite in quest'ambito, per cercare di ottenere una serie storica con un *trend* accentuato.

Ho quindi raccolto i dati relativi alle vendite di automobili negli USA dal 1992 al 2014 (ho escluso i dati relativi a quest'anno perchè ovviamente incompleti), espressi in milioni di dollari statunitensi, su base mensile.

Il sito da cui ho raccolto i dati è l'equivalente americano del sito dell'ISTAT:

http://www.census.gov/retail/marts/www/adv441x0.txt

Il formato dei dati è il seguente:

```
ANNO   GEN   FEB   MAR   APR   MAG   GIU
       LUG   AGO   SET   OTT   NOV   DIC
```

Ecco i dati raccolti:

1992	30167	30457	29891	30361	30847	31260
	31729	31124	32369	32396	32336	32693
1993	34021	33087	32430	34849	35267	35171
	36479	36689	36342	37020	38267	39035
1994	38676	39648	40435	40972	40081	40475
	40339	41053	42187	42856	43014	42718
1995	42991	41749	42482	42782	43442	45312
	44520	45351	44783	44867	45343	45687
1996	45865	47731	47809	46962	47604	47172
	47359	46899	48567	48718	48006	48069

1997	49223	49794	49373	49005	47406	49176
	50346	50536	50355	49941	50318	50945
1998	50551	49882	51122	52518	52842	54306
	51326	50016	52181	54221	54802	55400
1999	55418	56116	56554	56965	58046	58287
	59314	60563	59792	59175	60835	60559
2000	62306	63801	63027	60592	60492	61345
	59995	60075	61360	61017	59479	58207
2001	60970	61168	60635	61803	61922	61566
	60916	61183	58857	75352	67325	63109
2002	62485	62979	61460	63025	59858	62104
	64660	66924	62794	62319	62718	64635
2003	64454	61336	63363	64330	64353	64984
	65680	67538	65684	64341	66239	64681
2004	63831	65586	66560	64479	67280	63515
	65708	65839	69215	68248	67804	69644
2005	66928	67503	67490	68378	67158	73880
	76746	67586	64661	62389	67189	67099
2006	70796	67779	68781	68976	67812	69213
	70612	68783	68758	69223	69983	69764
2007	69294	69642	69348	69594	70536	68923
	68964	70288	71634	70834	69166	67340
2008	67425	65834	65242	64150	62585	60861
	58034	59238	55799	50139	48916	47460
2009	49321	47756	46865	47058	48168	50003
	51909	56487	47595	50192	51343	50660
2010	51101	49572	54523	54936	55374	54959
	56262	55723	56684	58494	58318	58532
2011	59970	60672	60722	60363	58981	59723
	60379	58576	61586	63148	63267	64557
2012	64282	65867	66014	65918	66483	65691
	65988	66922	69099	67531	68683	70483
2013	70441	71596	70701	71264	72401	74513
	74184	73941	73543	74292	75358	74619
2014	72347	74594	77517	78553	78704	79262
	79679	81179	79979	81253	82259	81922

1.2 Analisi di *trend* e periodicità

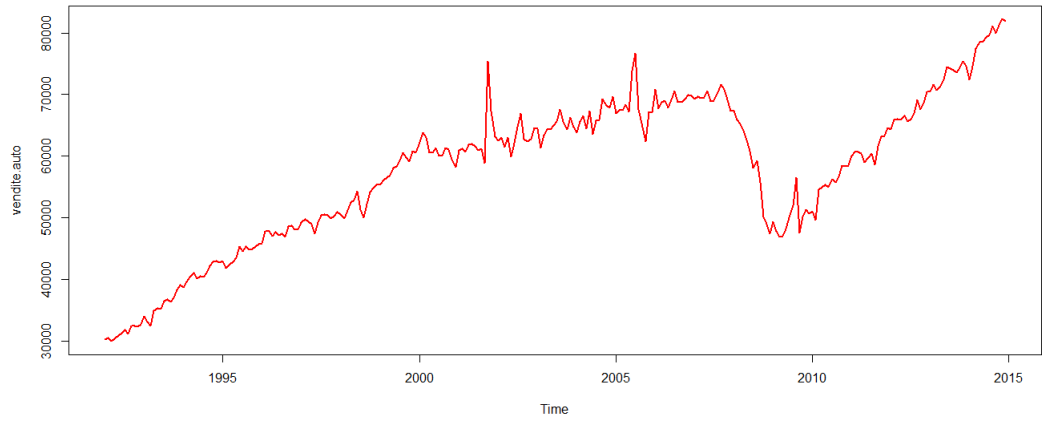
Vediamo ora di produrre un *plot* dei dati raccolti: la sensazione è che, essendo gli Stati Uniti un insieme di Stati economicamente potenti e non in crisi economica, il *trend* sia crescente, in linea con l'aumento della potenza economica della federazione.

Digitiamo allora su **R**, dopo aver copiato la tabella di dati:

```
> A.1 <- read.table("clipboard")
> A.2 = A.1[,2:13]
> X = c(t(A.2))
> vendite.auto = ts(X,frequency=12,start=c(1992,1))
> ts.plot(vendite.auto,lwd=2,col="red")
```

> **Osservazione** Il comando:
>
> ```
> > A.2 = A.1[,2:13]
> ```
>
> è stato dato per eliminare agevolmente la colonna comprendente gli anni, che non ci serve. Il comando:
>
> ```
> > X = c(t(A.2))
> ```
>
> serve invece per mettere in riga i dati nel giusto ordine.

Si ottiene quindi il grafico seguente:

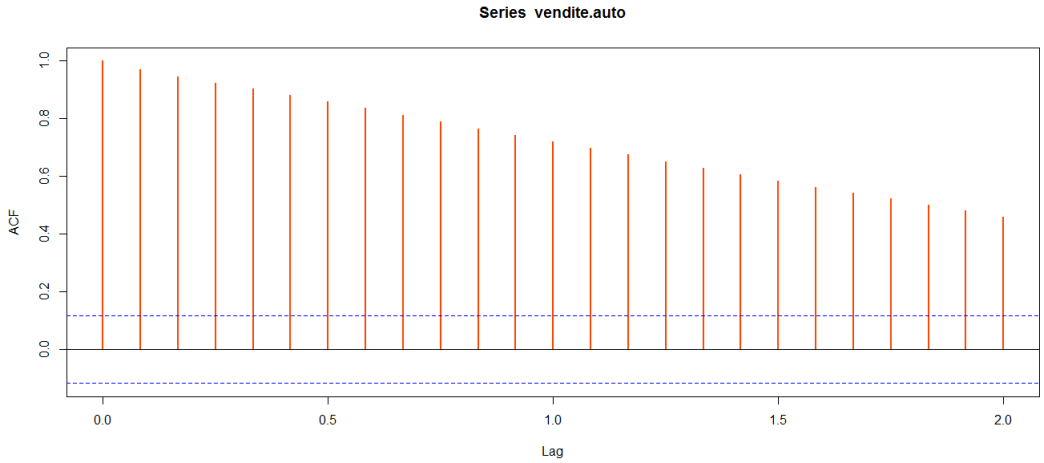


Il grafico mostra che, a meno di un calo brusco delle vendite osservato tra l'inizio del 2008 e la fine del 2009 (occorso a causa di un'effettiva recessione economica occorsa in quel periodo), il *trend* è globalmente ascendente: ciò è evidente.

Valutiamo la presenza di eventuali periodicità. Digitiamo allora:

```
> acf(vendite.auto,lwd=2,col="orangered")
```

Il grafico mostra l'assenza di periodicità rilevanti, così come ci aspettavamo dal grafico della serie storica:



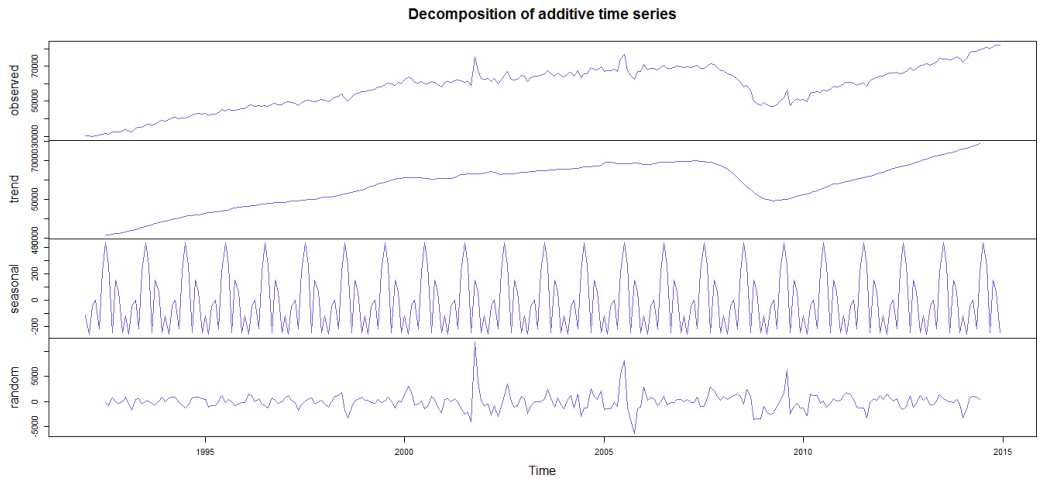
1.3 Decomposizione della serie

Operiamo ora una decomposizione sulla serie storica, per valutare il *trend*, la periodicità annuale, e l'entità del rumore.

Digitiamo allora:

```
> dec.auto = decompose(vendite.auto)
> plot(dec.auto,col="slateblue")
```

Si ottiene il seguente grafico:



Il rumore risulta abbastanza contenuto, se si escludono i tre o quattro picchi visibili a partire dal 2002.

1.4 Previsione del *trend*

Cerchiamo ora di effettuare una previsione del trend per l'intero anno del 2015.

Innanzitutto estrapoliamo il trend dalla serie. Invece di usare il comando `decompose`, però, usiamo il comando `stl`, dato che riteniamo che il trend discendente registrato negli anni intorno al 2008 sia saltuario, e dovuto a fattori esogeni, e non a un effettiva periodicità della serie.

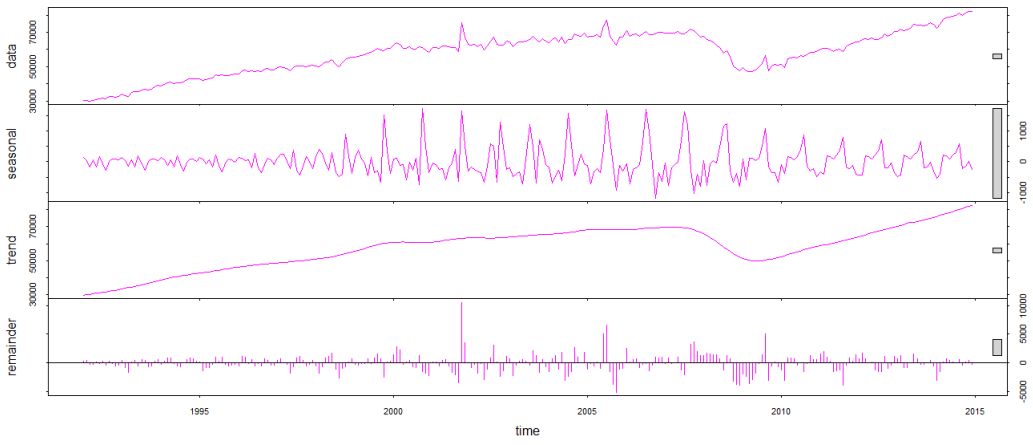
Effettuando un po' di prove, si scopre che il trend è catturato molto bene già per $k = 10$. Digitiamo quindi:

```
> k = 10
> V = vendite.auto
> S = stl(V,k)
> T = S$time.series[,2]
```

Diamo un'occhiata a quanto ottenuto, digitando:

```
> plot(S,col="magenta")
```

Si ottiene il seguente grafico:



Due osservazioni preliminari:

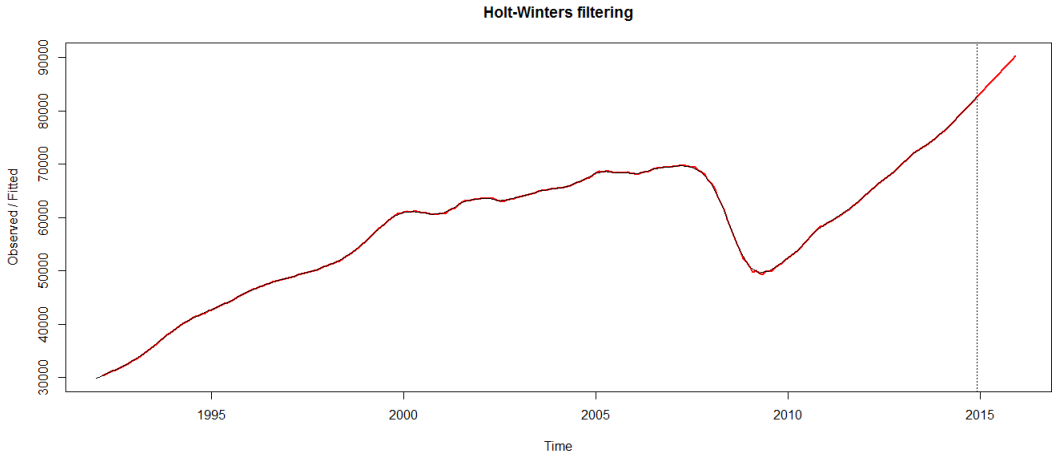
- Gli errori, soprattutto nell'ultimo periodo, sono molto contenuti;
- Il trend, nell'ultimo periodo, è ascendente con velocità pressochè costante.

Sarà dunque relativamente facile proseguire il trend, anche applicando solamente il metodo (SET) alla serie dei trend.

Digitiamo allora:

```
> HW.auto.trend = HoltWinters(T,gamma=FALSE)
> plot(HW.auto.trend, predict(HW.auto.trend,12),lwd=2)
```

Si ottiene allora il seguente grafico, molto attendibile:

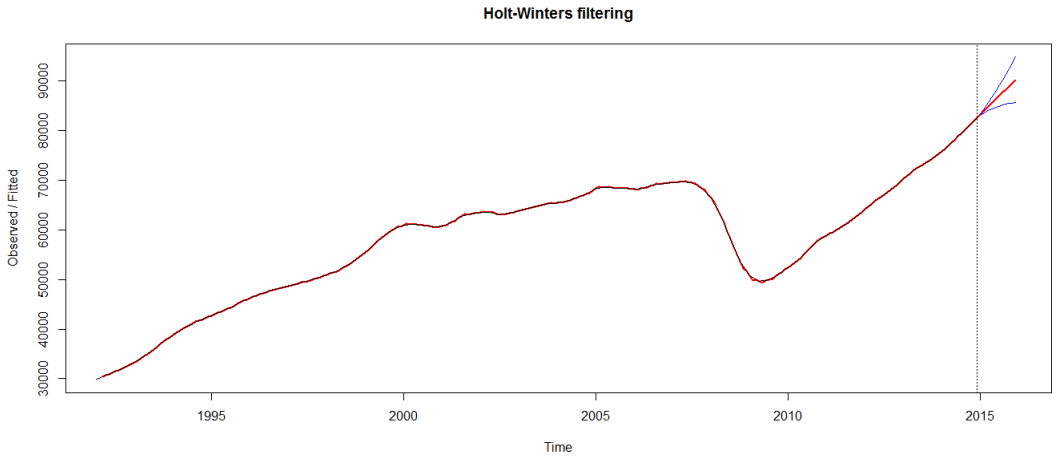


A occhio, la previsione sembra buona (nonostante i dati della previsione appartengano tutti ad una stessa retta, il che non è proprio desiderabile).

Digitando:

```
> plot(HW.auto.trend, predict(HW.auto.trend,12,prediction.interval=TRUE),lwd=2)
```

si ottiene un grafico dove è visibile una banda di confidenza per la previsione, in blu:



Infine, un dato curioso (ma non troppo). L'ottimalità del trend ottenuto con il comando `stl` è confermata dal fatto che, digitando:

```
> HW.auto.trend
```

si ottengono i coefficienti seguenti:

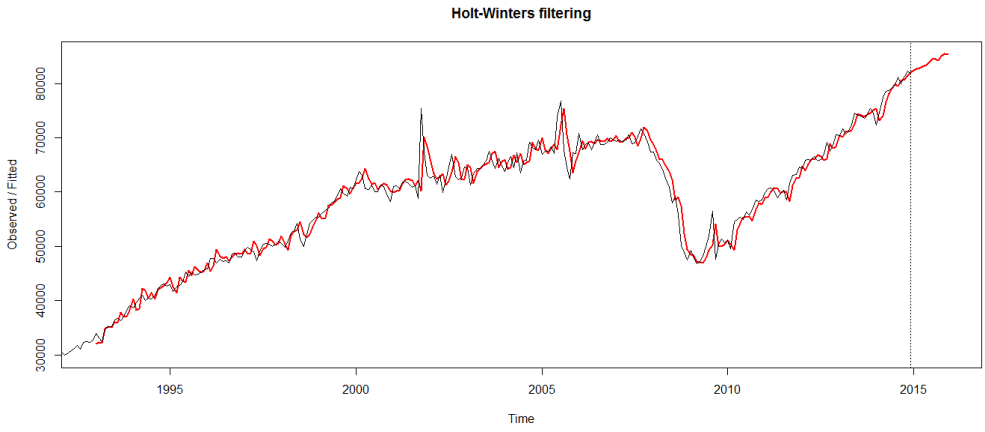
```
alpha:  1
beta:  1
```

1.5 Previsione della serie

Passiamo ora alla previsione della serie: iniziamo usando il metodo di Holt-Winters. Digitiamo quindi:

```
> HW.auto = HoltWinters(V)
> plot(HW.auto, predict(HW.auto,12),lwd=2)
```

Si ottiene il grafico seguente:



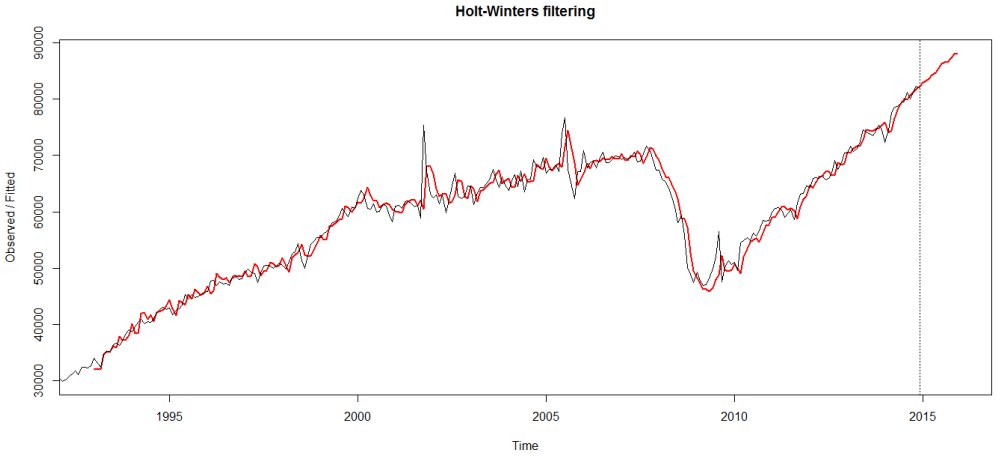
Sappiamo bene quanto il metodo di Holt-Winters sia valido, tuttavia il grafico non ci convince troppo: l'andamento sembra in calo, quando invece ce lo aspetteremo quantomeno non in calo. In questo caso, allora, conviene impostare manualmente qualche parametro. Quelli scelti dal software sono i seguenti:

```
alpha:  0.6519091
beta :  0.01723581
gamma:  0.1789215
```

Impostiamo un parametro α leggermente più conservativo, in modo da limitare un po' l'effetto delle variazioni nell'ultimo periodo. Digitando ad esempio:

```
> HW.auto.adj = HoltWinters(V,alpha=0.5)
> plot(HW.auto.adj, predict(HW.auto.adj,12),lwd=2)
```

Si ottiene una previsione un po' più corrispondente alla nostra intuizione:



Per curiosità, applichiamo unmetodo più elementare, basato sulla regressione.

Digitiamo allora:

```
> L = length(V)
> A = matrix(nrow=L-12,ncol=13)

> for (k in 1:12) {
+ A[,k] = V[(13-k):(L-k)]
+ }
> A[,13] = V[13:L]

> fit=lm(A[,13]~A[,1]+A[,2]+A[,3]+A[,4]+A[,5]+A[,6]+A[,7]+A[,8]+
+ A[,9]+A[,10]+A[,11]+A[,12])

> summary(fit)
```

Visualizzando i dati ottenuti, si scopre che solo il primo e il quarto fattore sono rilevanti. Infatti, digitando:

```
> fit = lm(A[,13]~A[,1]+A[,4])
> summary(fit)
```

I valori R^2 e *Adjusted* R^2 rimangono molto buoni.

Completiamo allora la previsione, digitando:

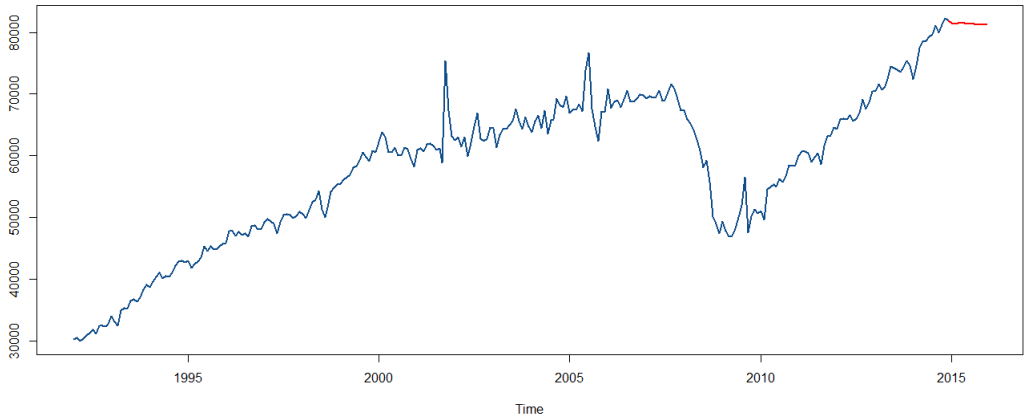
```
> P = 1:(L+12)
> P[1:L] = V
> for (k in 1:12) {
> P[L+k] = coef(fit) %*% c(1, P[L+k-1], P[L+k-4])
> }

> Pplus = ts(P,frequency=12,start=c(1992,1))
```

A questo punto, con un po' di fatica, si può visualizzare la serie storica e la previsione ottenuta:

```
> Pseries = Pprev = Pplus
> for (k in 1:(L-1)) {
> Pprev[k] = NA
> }
> for (k in (L+1):(L+12)) {
> Pseries[k] = NA
> }
> ts.plot(Pseries,Pprev,gpars=list(lwd=2,col=c("dodgerblue4","red")))
```

La previsione che si ottiene è molto diversa da quella ottenuta con il metodo di Holt-Winters (e decisamente meno credibile):



Tornando indietro, e includendo nella regressione tutti i fattori, si ottiene un risultato molto simile.

Osservazione Essendo alla fine del 2015, i dati relativi all'anno in corso sono quasi noti: analizziamoli, per vedere quale dei due metodi si avvicina di più (a occhio, diremmo il metodo di Holt-Winters).

I dati noti, che vanno da Gennaio 2015 a Ottobre 2015, sono i seguenti:

2015	82275	80403	83352	83871	85489	84037
	85391	85814	87145	86668		

Rappresentiamo allora, in un unico grafico, i dati reali relativi al 2015, e le previsioni ottenute con i due metodi.

Digitiamo allora, dopo aver copiato i dati relativi al 2015:

```
> E = scan("clipboard")
```

Digitiamo quindi:

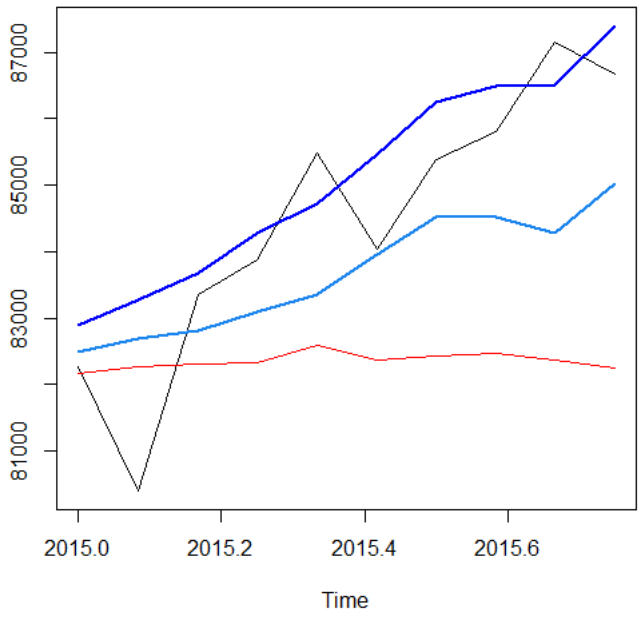
```
> E = ts(E,frequency=12,start=c(2015,1))

> HW.aux = predict(HW.auto,10)
> HW.adj.aux = predict(HW.auto.adj,10)

> l = length(Pprev)
> R.aux = as.numeric(Pprev[(l-11):(l-2)])
> R.aux = ts(R.aux,frequency=12,start=c(2015,1))

> ts.plot(E,HW.aux,HW.adj.aux,R.aux,gpars=list(lwd=c(1,2,2),
+ col=c("black","dodgerblue2","blue","red")))
```

Si ottiene allora il seguente grafico:



Come previsto, il metodo di regressione (in rosso) risulta abbastanza poco efficiente. Il metodo di Holt-Winters (in celeste) risulta decisamente più efficiente, ma la correzione del coefficiente α è risultata vincente: il grafico relativo al metodo di Holt-Winters con $\alpha = 0.5$ (in blu) costituisce la previsione migliore delle tre.

1.6 Analisi dei residui

Analizziamo ora i due metodi usati per analizzare la serie storica, tramite l'analisi dei residui. Trascuriamo d'orain avanti il modello regressivo, visto nella sezione precedente, perchè palesemente meno efficiente rispetto agli altri due.

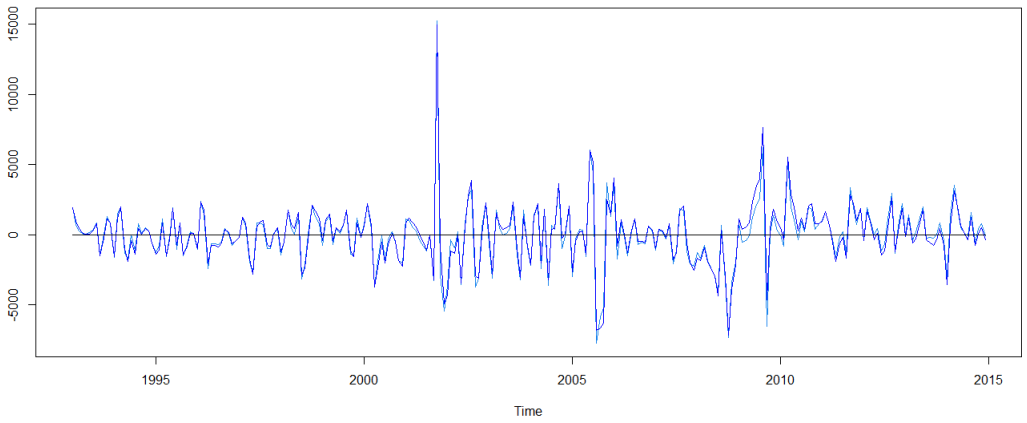
Digitiamo allora:

```
> RES.auto=residuals(HW.auto)
> RES.auto.adj=residuals(HW.auto.adj)

> Aux=X[13:276]*0
> line=ts(Aux,frequency=12,start=c(1993,1))

> ts.plot(RES.auto,RES.auto.adj,line,gpars=list(lwd=c(1,1,1),
> col=c("dodgerblue2","blue","black")))
```

Il grafico ottenuto non mostra particolari differenze tra i due metodi, in ambito di analisi:

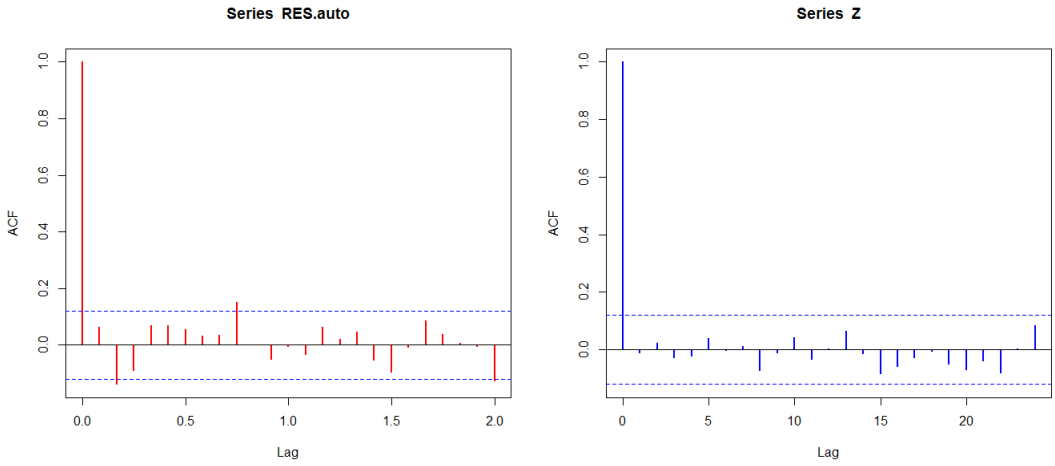


Forse il grafico celeste è leggermente migliori, ma le differenze sostanziali sono praticamente nulle: concludiamo che, anche per quanto visto in sede di previsione, il secondo metodo (parametri impostati manualmente) sembra, per ora, migliore del primo.

Effettuiamo ora un'analisi più dettagliata dei residui. Digitiamo quindi:

```
> Z=rnorm(264)
> par(mfrow=c(1,2))
> acf(RES.auto,lwd=2,col="red")
> acf(Z,lwd=2,col="blue")
```

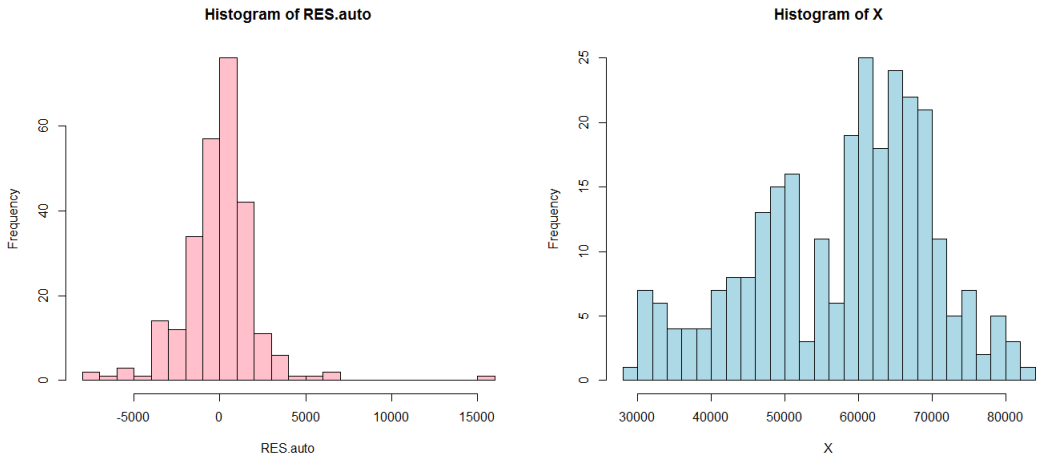
Il grafico ottenuto è abbastanza confortante:



Digitando:

```
> hist(RES.auto,20,col="pink")
> hist(X,20,col="lightblue")
```

si ottiene:



Digitando ora:

```
> 1-var(RES.auto)/var(X[13:  264])
```

si ottiene il valore:

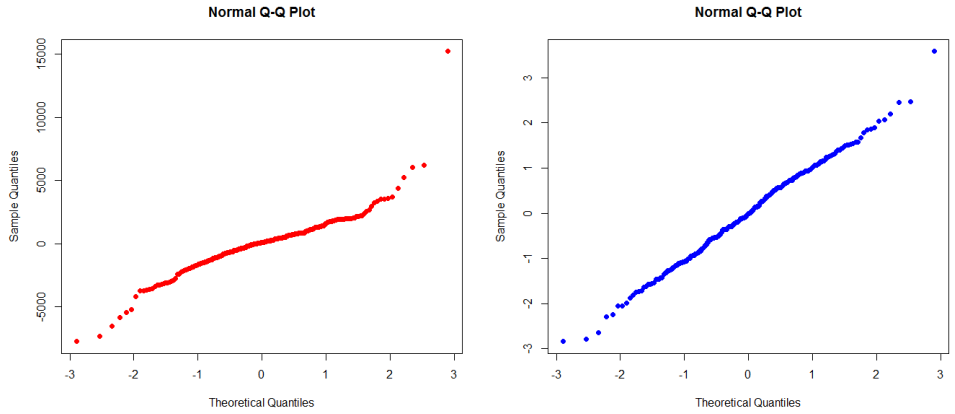
```
[1] 0.9576569
```

il quale è abbastanza buono.

Infine, digitando:

```
> qqnorm(RES.auto,pch=19,col="red")
> qqnorm(Z,pch=19,col="blue")
```

si ottiene un grafico che dimostra che i residui non sono propriamente gaussiani, e potrebbero nascondere dell'ulteriore struttura:



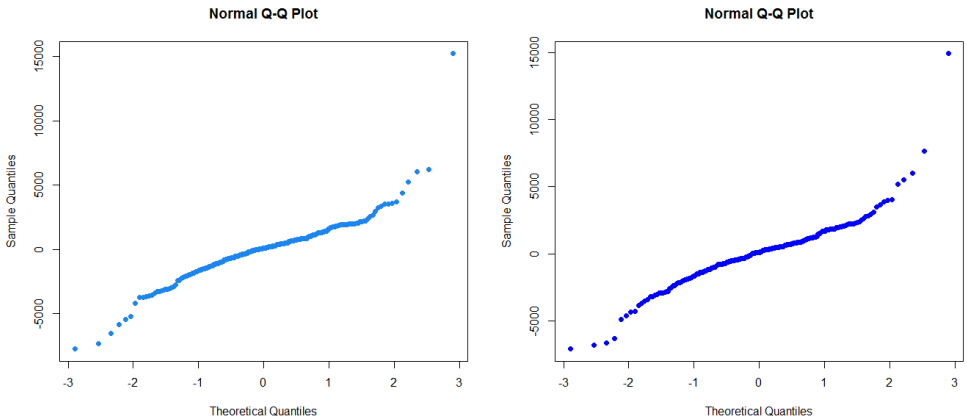
Ripetendo i comandi, in relazione al secondo metodo, si ottengono grafici simili. In particolare, la varianza spiegata risulta leggermente più bassa:

```
[1] 0.9569107
```

In particolare, mettiamo a confronto i grafici ottenuti nei due casi con il comando `qqnorm`. Digitiamo quindi:

```
> qqnorm(RES.auto,pch=19,col="dodgerblue2")
> qqnorm(RES.auto.adj,pch=19,col="blue")
```

Si ottiene il seguente grafico:



Entrambi i grafici non sono fantastici, tuttavia il secondo sembra un po' migliore del primo: un motivo in più per continuare a preferire il secondo metodo rispetto al primo.

1.7 Incertezza della previsione

Concentriamoci, d'ora in poi, sull'analisi effettuata usando il metodo di Holt-Winters con parametri scelti manualmente: abbiamo visto che, tra quelli esaminati nei paragrafi scorsi, questo sembra il migliore.

Digitiamo allora:

```
> quantile(RES.auto.adj,0.05)
> quantile(RES.auto.adj,0.95)
```

Si ottengono allora due valori, approssimativamente opposti:

```
       5%
-3174.219
```

```
      95%
 2757.572
```

Digitando allora:

```
> qmin = predict(HW.auto.adj,1)+quantile(RES.auto.adj,0.05)
> qmax = predict(HW.auto.adj,1)+quantile(RES.auto.adj,0.95)
```

si ottengono i seguenti valori, che individuano una banda di confidenza al 90% per la previsione relativa a Gennaio 2015:

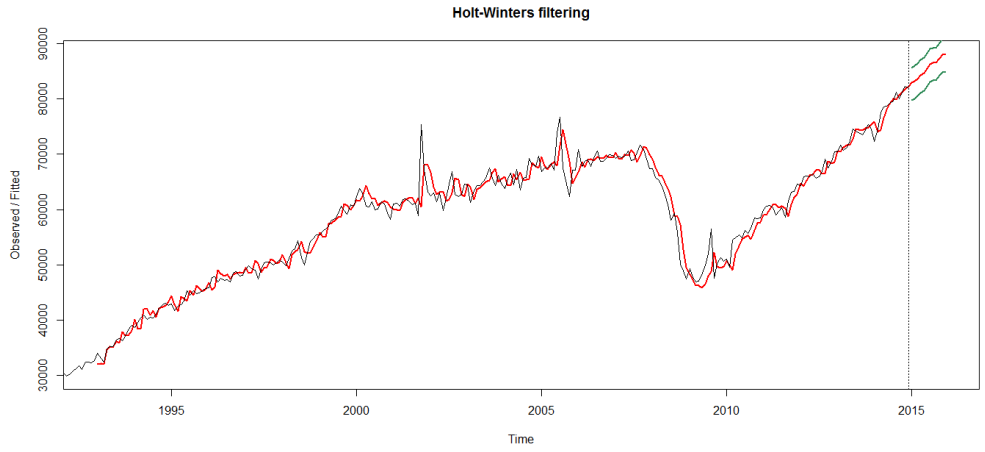
```
          Jan
2015 79723.16
```

```
          Jan
2015 85654.96
```

Passiamo quindi alla rappresentazione della banda di confidenza. Digitiamo allora:

```
> plot(HW.auto.adj,predict(HW.auto.adj,12),lwd=2)
> lines(predict(HW.auto.adj,12)+quantile(residui,0.05),col="seagreen",lwd=2)
> lines(predict(HW.auto.adj,12)+quantile(residui,0.95), col="seagreen",lwd=2)
```

Si ottiene allora il *plot* delle bande di confidenza:



Digitando invece:

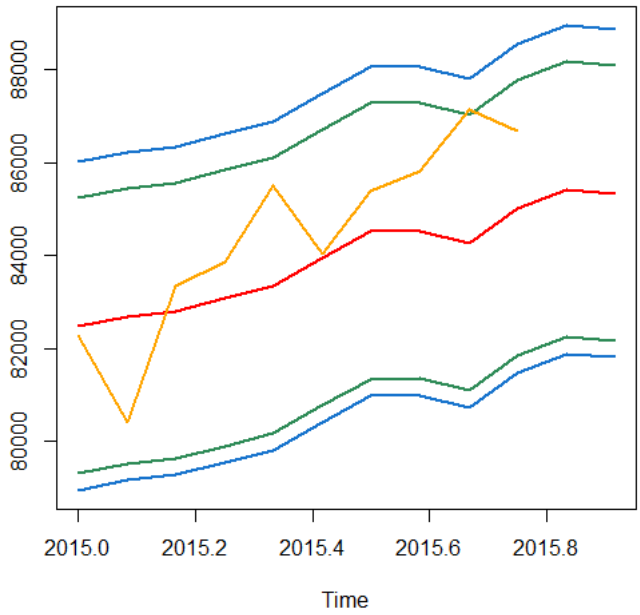
```
> ts.plot(predict(HW.auto,12),predict(HW.auto,12)+quantile(RES.auto.adj,0.05),
+ predict(HW.auto,12)+quantile(RES.auto.adj,0.95),type="n")
> lines(predict(HW.auto,12),col="red",lwd=2)
> lines(predict(HW.auto,12)+quantile(RES.auto.adj,0.05),col="seagreen",lwd=2)
> lines(predict(HW.auto,12)+quantile(RES.auto.adj,0.95), col="seagreen",lwd=2)
```

si ottiene un grafico relativo al solo anno 2015 (che non riportiamo).

Confrontiamo ora il metodo appena usato (non parametrico) con un metodo alternativo (parametrico). Supponendo che i residui abbiano legge gaussiana (e, dalla sezione precedente, possiamo affermare che questa supposizione è abbastanza forzata), possiamo mettere a confronto i due metodi con i seguenti comandi:

```
> ts.plot(predict(HW.auto,12),predict(HW.auto,12)+qnorm(0.05)*sd(RES.auto.adj),
+ predict(HW.auto,12)+qnorm(0.95)*sd(RES.auto.adj),type="n")
> lines(predict(HW.auto,12),col="red",lwd=2)
> lines(predict(HW.auto,12)+quantile(RES.auto.adj,0.05),col="seagreen",lwd=2)
> lines(predict(HW.auto,12)+quantile(RES.auto.adj,0.95),col="seagreen",lwd=2)
> lines(predict(HW.auto,12)+qnorm(0.05)*sd(RES.auto.adj),col="dodgerblue3",
+ lwd=2)
> lines(predict(HW.auto,12)+qnorm(0.95)*sd(RES.auto.adj),col="dodgerblue3",
+ lwd=2)
```

Si ottiene il seguente grafico, nel quale appare un grafico arancione in più, che tra poco spiegheremo:



La banda individuata con il metodo parametrico (in azzurro) è più larga dell'altra (in verde); inoltre essa è, per motivi teorici, simmetrica rispetto alla previsione (e questa può essere una caratteristica non utile, in molti casi). Deduciamo allora che, almeno in prima istanza, il metodo non parametrico appare più efficiente di quello parametrico.

Osservazione Riprendiamo i dati relativi al 2015, e rappresentiamoli sul grafico con le bande di confidenza, aggiungendo i comandi:

```
> E = scan("clipboard")
> E[11] = E[12] = NA
> E = ts(E,frequency=12,start=c(2015,1))
> lines(E,col="orange",lwd=2)
```

Il grafico visto in precedenza, dove in arancione compaiono proprio i dati relativi al 2015, conferma a pieno le nostre aspettative.

Inoltre esso evidenzia, in qualche senso, l'ottimalità (o comunque la bontà) dell'uso del metodo non parametrico per creare la banda di confidenza.

1.8 Fattori esogeni

Cerchiamo ora di generare un modello regressivo più accurato, tramite l'uso di un fattore esogeno.

Dal sito seguente:

https://ycharts.com/indicators/us_monthly_gdp

ho scaricato i dati, su base mensile, da Aprile 1992 a Dicembre 2014, relativi al prodotto interno lordo (in inglese, GDP) degli Stati Uniti. I dati sono espressi in migliaia di miliardi di dollari.

> **Osservazione** Se si prova a trovare questi dati sul sito, questi saranno inaccessibili: per ottenerli, infatti, bisogna iscriversi al sito.

Riportiamo di seguito i dati raccolti (in ogni coppia di righe adiacenti ci sono i dati di un anno, dal 1992 al 2014):

1992	NA	NA	NA	6.270	6.252	6.364
	6.383	6.380	6.406	6.452	6.484	6.544
1993	6.534	6.548	6.552	6.585	6.631	6.653
	6.644	6.679	6.742	6.750	6.831	6.860
1994	6.875	6.928	6.946	6.986	7.089	7.058
	7.103	7.155	7.137	7.225	7.230	7.289
1995	7.329	7.273	7.321	7.317	7.337	7.413
	7.393	7.445	7.519	7.517	7.521	7.589
1996	7.611	7.620	7.683	7.792	7.780	7.828
	7.860	7.886	7.932	8.012	8.044	8.013
1997	8.089	8.161	8.162	8.272	8.235	8.324
	8.376	8.400	8.454	8.490	8.477	8.549
1998	8.506	8.634	8.662	8.654	8.681	8.760
	8.771	8.827	8.943	8.998	9.042	9.043
1999	9.088	9.153	9.205	9.222	9.265	9.271
	9.380	9.381	9.454	9.530	9.606	9.687
2000	9.636	9.700	9.793	9.922	9.937	9.988
	9.937	10.040	10.070	10.130	10.140	10.120
2001	10.160	10.150	10.180	10.260	10.340	10.300
	10.280	10.400	10.230	10.350	10.300	10.470
2002	10.500	10.450	10.540	10.580	10.580	10.650
	10.720	10.670	10.710	10.710	10.760	10.820
2003	10.830	10.930	10.900	10.980	10.970	11.090
	11.170	11.250	11.350	11.370	11.420	11.450
2004	11.480	11.610	11.680	11.700	11.820	11.770
	11.920	11.930	11.960	12.090	12.120	12.170
2005	12.350	12.330	12.410	12.460	12.440	12.600
	12.660	12.740	12.780	12.830	12.870	13.000
2006	13.090	13.100	13.290	13.270	13.340	13.370
	13.380	13.430	13.490	13.520	13.620	13.620

```
2007    13.680 13.830 13.760 13.940 14.000 13.990
        14.000 14.120 14.260 14.190 14.240 14.330
2008    14.360 14.180 14.270 14.330 14.340 14.580
        14.490 14.370 14.330 14.200 14.160 13.880
2009    13.930 13.950 13.890 13.870 13.880 13.910
        13.890 13.990 13.980 14.160 14.160 14.080
2010    14.210 14.250 14.350 14.400 14.400 14.450
        14.540 14.550 14.640 14.710 14.690 14.810
2011    14.720 14.740 14.990 15.040 15.040 14.930
        15.130 15.690 15.580 15.870 15.790 15.790
2012    15.920 16.140 16.060 16.120 16.150 16.130
        16.280 16.200 16.330 16.280 16.300 16.420
2013    16.520 16.460 16.530 16.580 16.550 16.640
        16.650 16.730 16.810 16.860 17.010 17.010
2014    16.910 16.980 17.070 17.180 17.280 17.350
        17.440 17.560 17.570 17.600 17.660 17.590
```

Carichiamo i dati su **R**:

```
> PIL = scan("clipboard")
> Ec = ts(PIL,frequency=12,start=c(1992,1))
```

Supporremo, d'ora in avanti, che le variazioni del prodotto interno lordo abbiano effetto immediato sulle variazioni delle vendite di automoboli: tale supposizione è plausibile.

Rivedendo la regressione applicata nella sezione (1.5), digitiamo allora:

```
> L = length(V)
> A = matrix(nrow=L-12,ncol=4)

> A[,1] = V[12:(L-1)]
> A[,2] = V[19:(L-4)]
> A[,3] = Ec[12:(L-1)]
> A[,4] = V[13:L]

> fit = lm(A[,4]~A[,1]+A[,2]+A[,3])
```

Digitando:

```
> summary(fit)
```

si ottengono in particolare valori R^2 e *Adjusted* R^2 praticamente identici (R^2 è un po' più alto, ma ciò era scontato; *Adjusted* R^2 è identico).

Digitando:

```
> ccf(PIL[4:L],X[4:L])
```

si ha che la supposizione effetuata è valida: concludiamo allora che, perlomeno con questo fattore esogeno, non si sono avuti miglioramenti netti (d'altronde era abbastanza difficile aumentare un valore della varianza spiegata già in partenza elevatissimo).

Capitolo 2

Quantità di pioggia nella Scozia Settentrionale

2.1 Introduzione alla serie storica

Volendo ora analizzare una serie storica con periodicità moderata, ho raccolto i dati sui millimetri di pioggia caduti nella Scozia Settentrionale, dall'inizio del 2000 alla fine del 2014 (al solito, i dati del 2015 sono stati evitati perchè incompleti).

I dati sono stati raccolti dal sito:

http://www.metoffice.gov.uk/hadobs/hadukp/data/monthly/HadNSP_monthly_qc.txt

Il formato dei dati è il seguente:

```
ANNO   GEN   FEB   MAR   APR   MAG   GIU   LUG   AGO   SET
       OTT   NOV   DIC
```

I dati raccolti sono i seguenti:

```
2000 203.7 206.9 165.4 103.3  63.2  95.9  41.5 102.9 124.2
     216.7 175.7 193.6
2001  89.1 113.3  77.2  65.9  54.2  89.7 116.5 135.5 135.0
     262.3 172.4 157.5
2002 201.2 214.8 124.2  76.6  76.2 120.1 128.4  72.2  58.9
     162.3 167.1 106.9
2003 218.9  70.3  91.1  55.8 121.0  98.4  67.2  67.6 128.3
     142.7 154.1 168.4
2004 197.1 115.7 101.3  93.4  58.2 127.4  78.4 143.5 156.8
     192.1 144.3 175.4
2005 223.4 124.6 143.5  98.0 109.4 122.3  80.2 158.3 180.5
     173.0 162.0 136.6
```

```
2006 107.7  83.5 115.5 130.5 105.5  81.0  83.5  95.3 144.6
     281.5 232.4 255.6
2007 267.3 142.1 148.4  64.6 134.4  58.7 148.6 157.6 130.6
     114.8 170.3 156.8
2008 271.8 166.6 175.8  94.4  20.5 116.9  91.4 123.1 104.2
     269.4 160.5 164.4
2009 156.2 110.1 178.1  80.4 114.3  67.3 140.6 194.2 146.6
     219.1 238.9 105.3
2010  98.2 103.4 100.4  90.1  63.4  38.0 171.3 116.3 162.6
     151.2 118.8  68.3
2011 139.3 144.9 101.0  84.9 194.0  87.2  80.4 183.0 197.7
     190.8 167.4 310.6
2012 228.8 123.7  70.5 107.2  87.4  83.7  90.9 132.6 170.4
     157.8 159.4 236.5
2013 142.9  88.7  49.6 129.4 114.2  62.9  70.4 115.3 124.4
     186.0 178.8 294.2
2014 200.9 191.8 133.2  94.6  98.8  61.9  84.9 220.8  66.6
     269.0 125.0 252.9
```

2.2 Analisi di *trend* e periodicità

Vediamo ora di produrre un *plot* dei dati raccolti: la sensazione è che il *trend* sia sostanzialmente costante, menre vi sia una periodicità moderatamente accentuata.

Digitiamo allora su **R**, dopo aver copiato la tabella di dati:

```
> B.1 <- read.table("clipboard")
> B.2 = B.1[,2:13]
> Y = c(t(B.2))
> mm.pioggia = ts(Y,frequency=12,start=c(2000,1))
> ts.plot(mm.pioggia,lwd=2,col="dodgerblue4")
```

Osservazione Il comando:

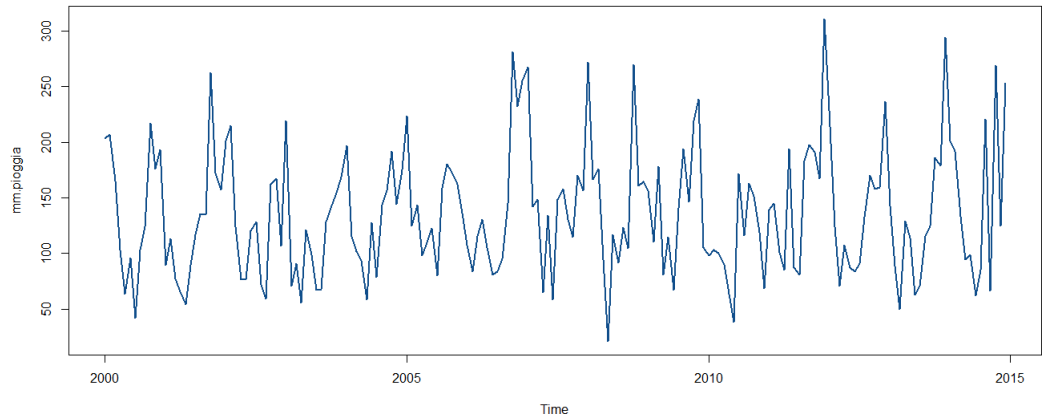
```
> B.2 = B.1[,2:13]
```

è stato dato per eliminare agevolmente la colonna comprendente gli anni, che non ci serve. Il comando:

```
> Y = c(t(B.2))
```

serve invece per mettere in riga i dati nel giusto ordine.

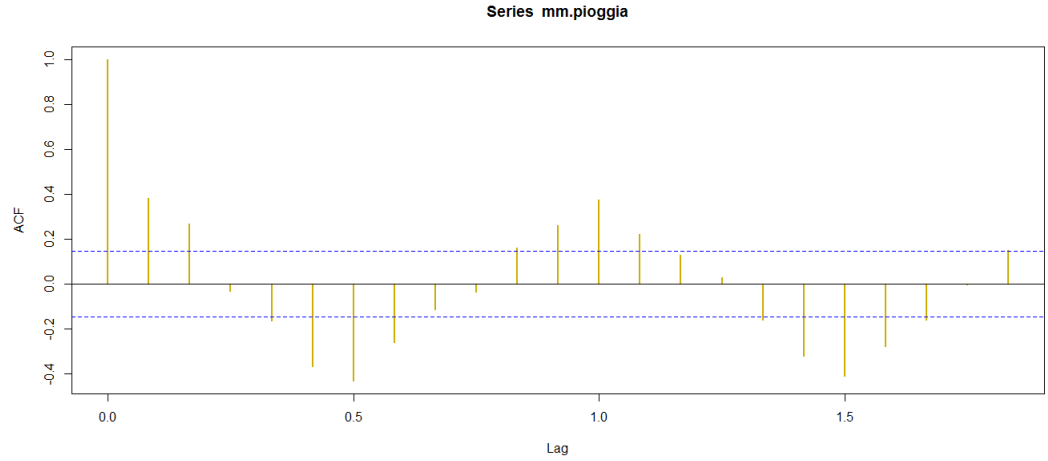
Si ottiene quindi il grafico seguente:



Come previsto, il *trend* è all'incirca stazionario, mentre la periodicità annuale è abbastanza accentuata. Ciò è confermato digitando:

```
> acf(mm.pioggia,lwd=2,col="gold3")
```

e osservando il seguente grafico:



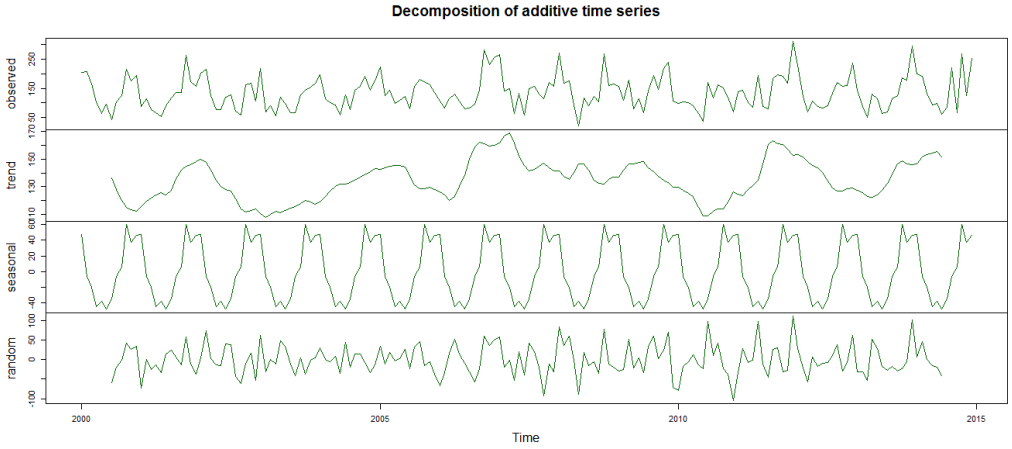
2.3 Decomposizione della serie

Operiamo quindi una decomposizione sulla serie storica, per valutare il *trend*, la periodicità annuale, e l'entità del rumore.

Digitiamo allora:

```
> dec.pioggia = decompose(mm.pioggia)
> plot(dec.pioggia,col="darkgreen")
```

Si ottiene il seguente grafico:



Come previsto, il *trend* è stazionario, e la periodicità è accentuata. A differenza della serie storica precedente, però, in questo caso il rumore è abbastanza elevato.

2.4 Previsione del trend

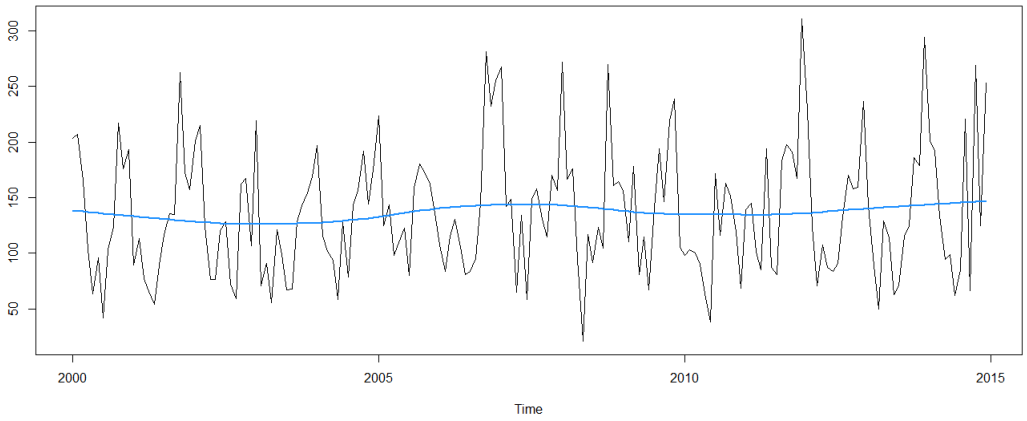
Cerchiamo ora di effettuare una previsione del trend per l'intero anno del 2015.

In questo caso useremo sempre il comando `stl`, ma con k molto basso (e non `decompose`, in questo caso particolarmente poco efficiente).

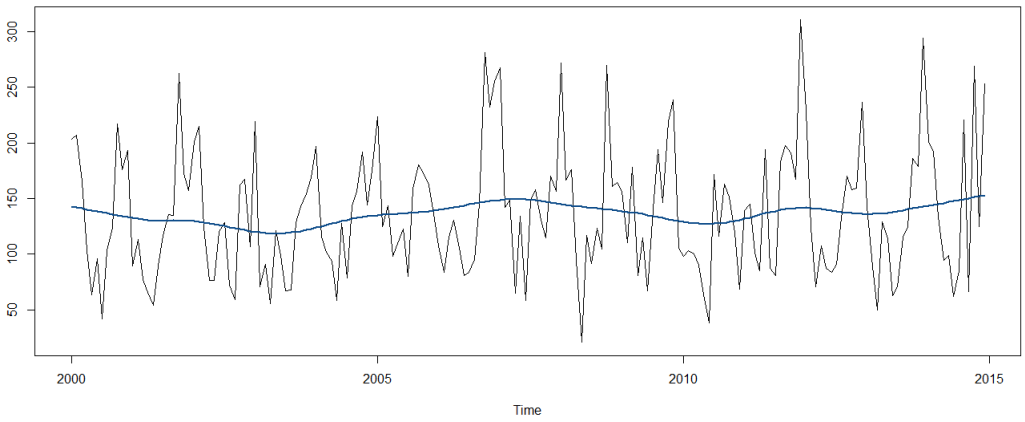
Scegliamo $k = 2$, e digitiamo:

```
> k = 2
> Stl.rain = stl(mm.pioggia,k)
> Stl.trend = Stl.rain$time.series[,2]
> ts.plot(mm.pioggia,Stl.trend,gpars=list(lwd=c(1,2),
+ col=c("black","dodgerblue")))
```

Si ottiene quindi il seguente grafico:



Per $k = 3$, invece, con comandi analoghi si ottiene il seguente grafico:



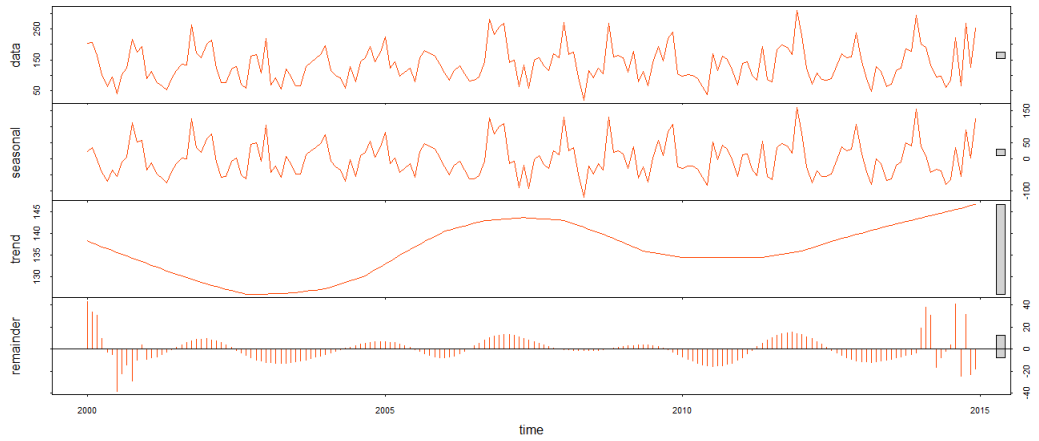
Forse è meglio optare per la scelta $k = 2$. Estrapoliamo la serie del trend:

```
> k = 2
> V = mm.pioggia
> S = stl(V,k)
> T = S$time.series[,2]
```

Diamo un'occhiata ai residui:

```
> plot(S,col="orangered")
```

Si ottiene:



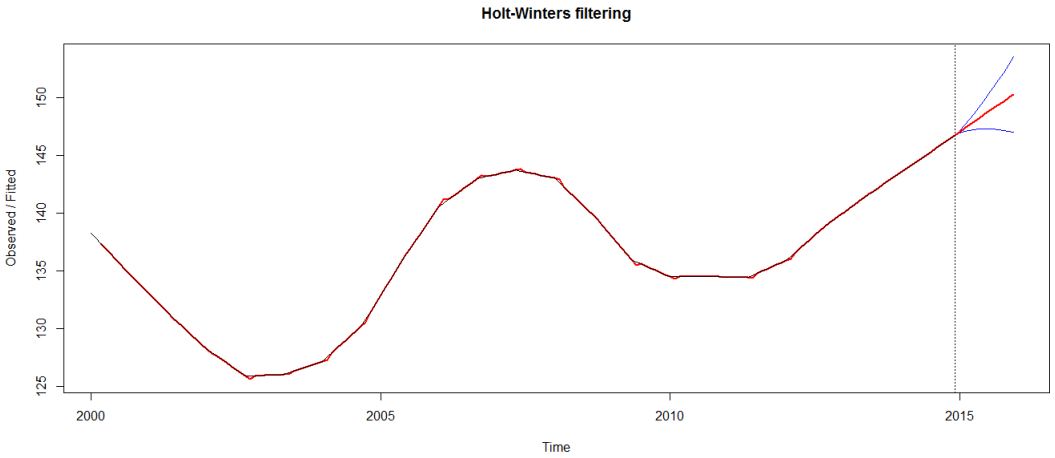
Osservazione Osserviamo la strana (neanche troppo) forma dei residui, che comunque risultano abbastanza contenuti.

Vediamo ora di proseguire il trend in maniera accettabile.

Facciamo un primo tentativo:

```
> HW.rain.trend = HoltWinters(T,gamma=FALSE)
> plot(HW.rain.trend, predict(HW.rain.trend,12,prediction.interval=TRUE),lwd=2)
```

Si ottiene allora il seguente grafico:



Il risultato è assolutamente deludente: non resta che tentare di proseguire il trend *a mano*.

Osservazione Notiamo che una delle due bande di confidenza, quella inferiore, sembra decisamente più attendibile come previsione del trend. Proviamo allora a effettuare una media tra i valori ottenuti e i valori appartenenti a questa curva.

Iniziamo digitando:

```
> T = S$time.series[,2]
> q = predict(HW.rain.trend,12)
> q.num = as.numeric(q)
> q.num[2]-q.num[1]
```

Il software restituisce il seguente valore:

```
[1] 0.2936564
```

Sulla base di tale valore, modifichiamo i valori previsti, creando una parabola discendente. Digitiamo allora:

```
> for (k in 3:12) {
+ q.num[k] = q.num[k-1]+0.025*(11-k*1.05)
+ }
```

A questo punto digitiamo:

```
> L = length(T)
> T.aux = as.numeric(T)
> TT = qq=1:(L+12)
```

```
> for (j in 1:(L-1)) {
> qq[j] = NA
+ TT[j] = T.aux[j]
+ }
```

```
> qq[L] = TT[L] = T.aux[L]
```

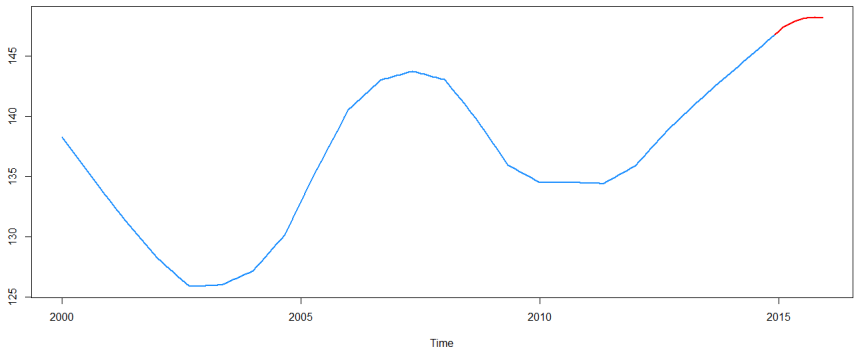
```
> for (j in (L+1):(L+12)) {
> qq[j] = q.num[j-180]
+ TT[j] = NA
+ }
```

Per concludere, digitiamo:

```
> T.plot = ts(TT,frequency=12,start=c(2000,1))
> q.plot = ts(qq,frequency=12,start=c(2000,1))
```

```
> ts.plot(T.plot,q.plot,gpars=list(lwd=c(2,2),col=c("dodgerblue","red")))
```

Il grafico ottenuto è decisamente migliore di quello precedente:



Conserveremo nel seguito questa previsione del trend, dunque.

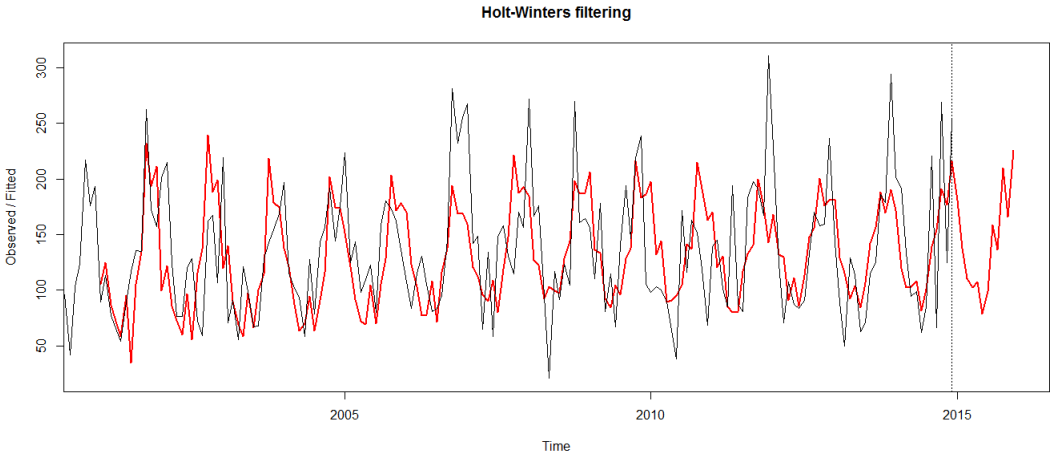
2.5 Previsione della serie

Passiamo ora alla previsione della serie.

Iniziamo con il metodo di Holt-Winters, digitando:

```
> HW.rain = HoltWinters(V)
> plot(HW.rain, predict(HW.rain,12),lwd=2)
```

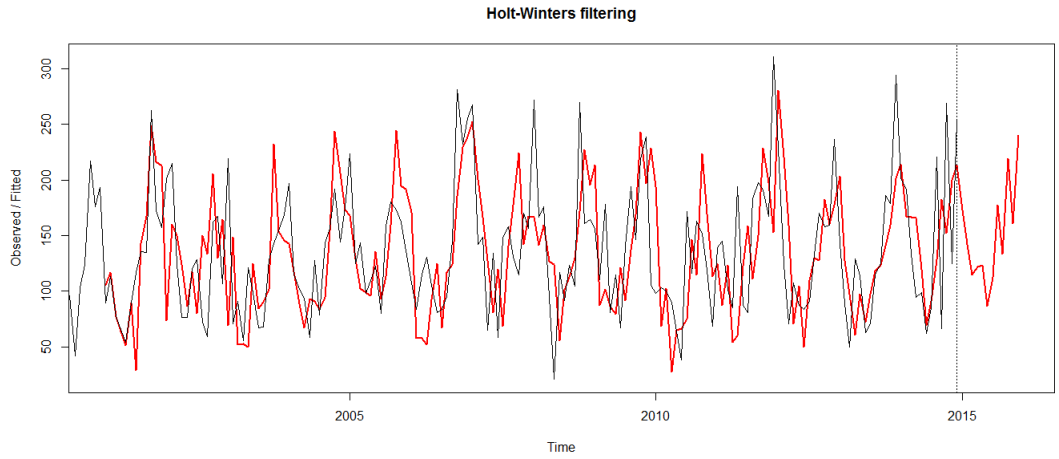
Il grafico ottenuto non è eccezionale:



Forse modificando i parametri si riesce a ottenere qualcosa di meglio. Digitiamo ad esempio:

```
> HW.rain.adj = HoltWinters(V,alpha=0.5,gamma=0.5)
> plot(HW.rain.adj, predict(HW.rain.adj,12),lwd=2)
```

Si ottiene allora un grafico un po' migliore:



Applichiamo ora la regressione manuale. Digitiamo quindi:

```
> L = length(V)
> A = matrix(nrow=L-12,ncol=13)

> for (k in 1:12) {
+ A[,k] = V[(13-k):(L-k)]
+ }

> A[,13] = V[13:L]
> fit = lm(A[,13]~A[,1]+A[,2]+A[,3]+A[,4]+A[,5]+A[,6]+A[,7]+A[,8]+
+ A[,9]+A[,10]+A[,11]+A[,12])
> summary(fit)
```

Guardando il *summary*, ci si può rendere conto che il risultato finale non sarà eccellente: i valori R^2 e *Adjusted* R^2 sono infatti bassissimi. La situazione non migliora nemmeno eliminando dei fattori. Giusto per curiosità, portiamo a termine in ogni caso il metodo, non prima di aver eliminato qualche fattore, però:

```
> fit = lm(A[,13]~A[,1]+A[,2]+A[,5]+A[,6]+A[,12])
> P = 1:(L+12)
> P[1:L] = V

> for (k in 1:12) {
+ P[L+k] = coef(fit) %*% c(1,P[L+k-1],P[L+k-2],P[L+k-5],P[L+k-6],P[L+k-12])
+ }

> Pplus = ts(P,frequency=12,start=c(1992,1))
> plot(Pplus,lwd=2,col="dodgerblue4")
```

Digitiamo quindi, per concludere:

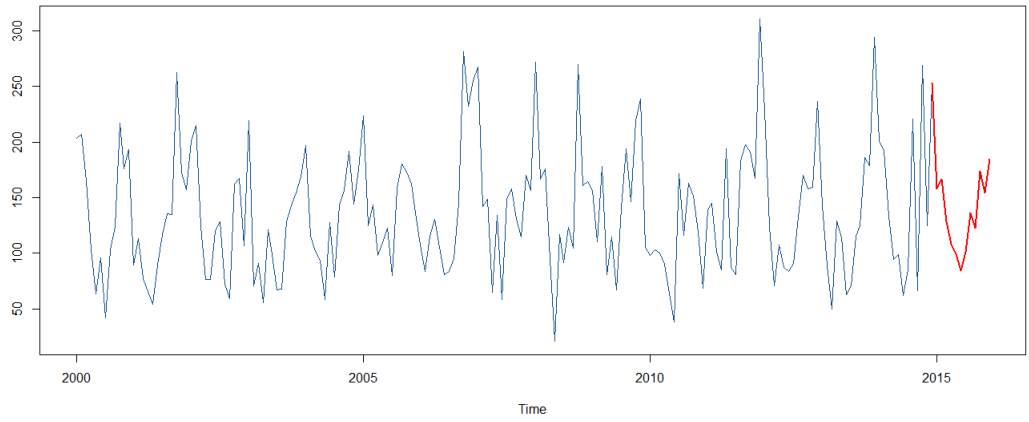
```
> Pseries = Pprev = Pplus

> for (k in 1:(L-1)) {
+ Pprev[k] = NA
+ }

> for (k in (L+1):(L+12)) {
+ Pseries[k] = NA
+ }

> ts.plot(Pseries,Pprev,gpars=list(lwd=c(1,2),col=c("dodgerblue4","red")))
```

Il grafico ottenuto non è poi così brutto:



Osservazione I grafici ottenuti con il metodo di Holt-Winters (soprattutto il secondo), per ora, sembrano i migliori. Essi però si basavano su un *trend* che, nella scorsa sezione, abbiamo pensato non fosse attendibile.

Vedremo ora, quindi, come usare il *trend* creato manualmente, a nostro parere più attendibile, per creare una altrettanto attendibile previsione della serie.

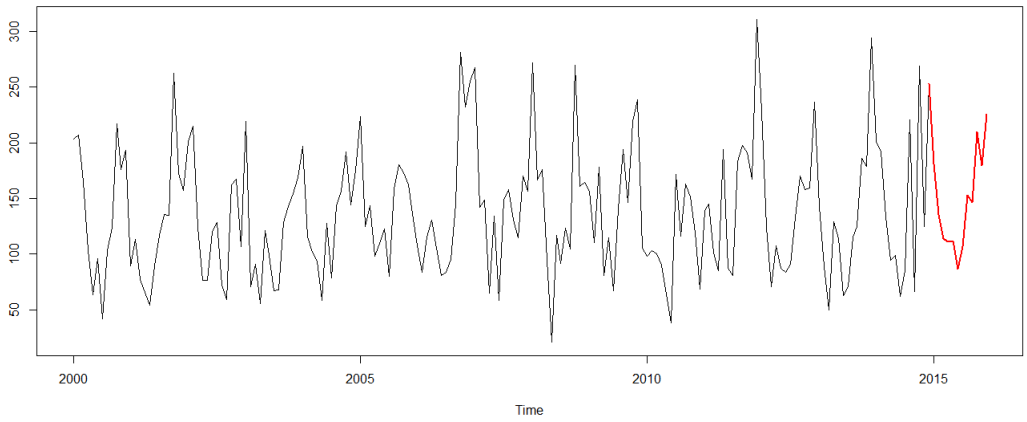
Iniziamo effettuando la previsione della stagionalità, che andremo poi a sommare al *trend* da noi creato: usiamo parametri più conservativi, in modo che risalti la differenza di trend rispetto alle altre serie:

```
> Seasonal = S$time.series[,1]
> HW.S = HoltWinters(Seasonal,alpha=0.2,gamma=0.2)
> plot(HW.S,predict(HW.S,12),col=c("black","red"))
```

A questo punto digitiamo:

```
> w = as.numeric(predict(HW.S,12))
> q.num = as.numeric(q.plot)
> q.num[181:192] = q.num[181:192]+w[1:12]
> q.num[180] = V[180]
> qq.plot = ts(q.num,frequency=12,start=c(2000,1))
> ts.plot(V,qq.plot,gpars=list(lwd=c(1,2),col=c("black","red")))
```

Si ottiene il grafico seguente, abbastanza attendibile:



Osservazione Anche in questo caso, i dati del 2015 sono ormai noti, dato che l'anno è quasi terminato.

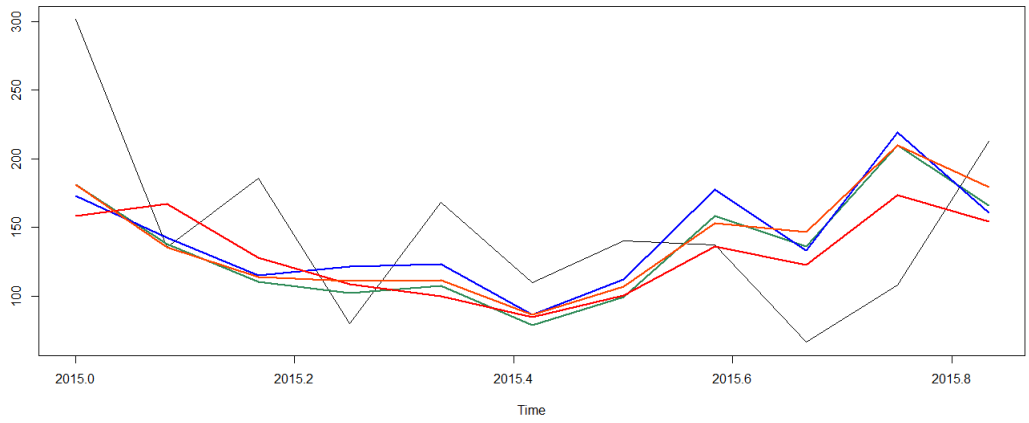
Vediamo ora quale previsione si rivela la migliore. Carichiamo su **R** il vettore dei dati relativi al 2015 (da Gennaio a Novembre compreso):

```
2015  301.4  135.4  185.7   80.0  168.5  109.8
      140.5  137.4   66.3  108.3  212.5
```

A questo punto digitiamo:

```
> E = ts(E,frequency=12,start=c(2015,1))
> HW.aux = predict(HW.rain,11)
> HW.adj.aux = predict(HW.rain.adj,11)
> l = length(Pprev)
> R.aux = as.numeric(Pprev[(l-11):(l-1)])
> R.aux = ts(R.aux,frequency=12,start=c(2015,1))
> L = length(qq.plot)
> R.handmade = as.numeric(qq.plot)[(L-11):(L-1)]
> R.handmade = ts(R.handmade,frequency=12,start=c(2015,1))
> ts.plot(E,HW.aux,HW.adj.aux,R.aux,R.handmade,gpars=list(lwd=c(1,2,2,2,2),
> col=c("black","seagreen","blue","red","orangered")))
```

Si ottiene allora il seguente grafico:



È davvero difficile dire quale delle 4 previsioni è migliore, ma certamente quella, fatta a mano usando il *trend* costruito a mano, non sfigura di fronte alle altre (nel disegno, è la previsione in arancione).

> **Osservazione** Una piccola nota di colore: ho pensato di chiedere a undici miei amici, scelti a caso, di scegliere quale fosse per loro la previsione migliore (senza ovviamente condizionarli dicendo loro quale fosse quella fatta a mano da me). I voti sono stati i seguenti:
>
> - Arancione: 4 voti;
> - Rossa: 3 voti;
> - Blu: 2 voti;
> - Verde: 2 voti.
>
> Certamente i voti per la previsione blu e quella verde non potevano essere tanti, vista la semplicità del loro algoritmo, e l'assenza quasi totale di aggiustamenti. In ogni caso, è un dato decisamente appagante!

Vediamo ora qual è la previsione migliore delle quattro, calcolando gli scarti quadratici. Digitiamo allora:

```
> Orange = as.numeric(R.handmade)-as.numeric(E)
> Red = as.numeric(R.aux)-as.numeric(E)
> Blue = as.numeric(HW.adj.aux)-as.numeric(E)
> Green = as.numeric(HW.aux)-as.numeric(E)

> Orange %*% Orange
> Red %*% Red
> Blue %*% Blue
> Green %*% Green
```

Si ottengono i seguenti valori:

```
[1,] 43437.20
[1,] 43447.98
[1,] 47697.45
[1,] 44808.54
```

Il metodo fatto a mano, dunque, risulta il migliore di quelli visti finora, anche col criterio dei minimi quadrati. Un'ulteriore conferma!

2.6 Analisi dei residui

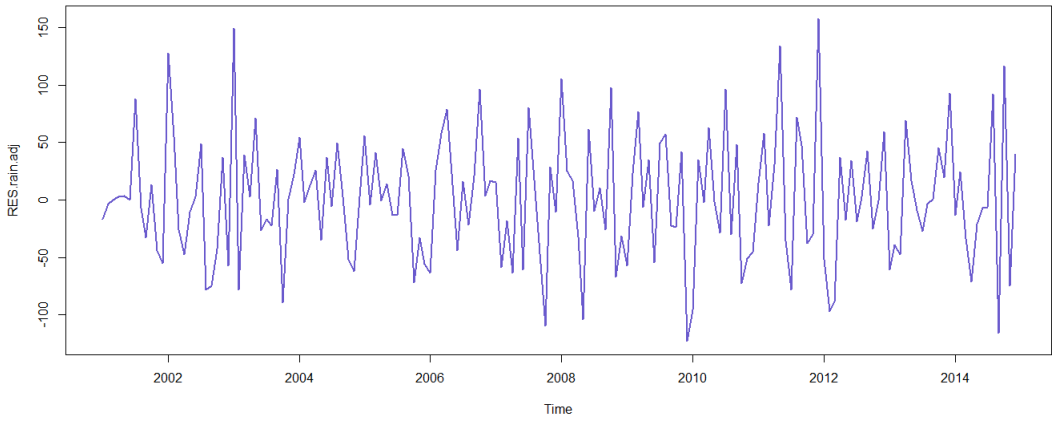
Effettuiamo ora l'analisi dei residui, così come abbiamo fatto per la serie delle vendite di automobili. In questa sezione, allora, analizzeremo due metodi:

- Il metodo di Holt-Winters con parametri modificati;
- Il metodo regressivo.

Per il primo metodo, digitiamo:

```
> RES.rain.adj = residuals(HW.rain.adj)
> plot(RES.rain.adj,lwd=2,col="slateblue")
```

Si ottiene allora il grafico seguente:



È immediato notare che gli errori non sono affatto contenuti.

L'analisi può proseguire digitando i seguenti comandi:

```
> par(mfrow=c(1,2))
> Z = rnorm(length(Y)-12)
> acf(RES.rain.adj,lwd=2,col="red")
> acf(Z,lwd=2,col="blue")
```

```
> hist(RES.rain.adj,20,col="lightpink")
> hist(Y,20,col="lightblue")
```

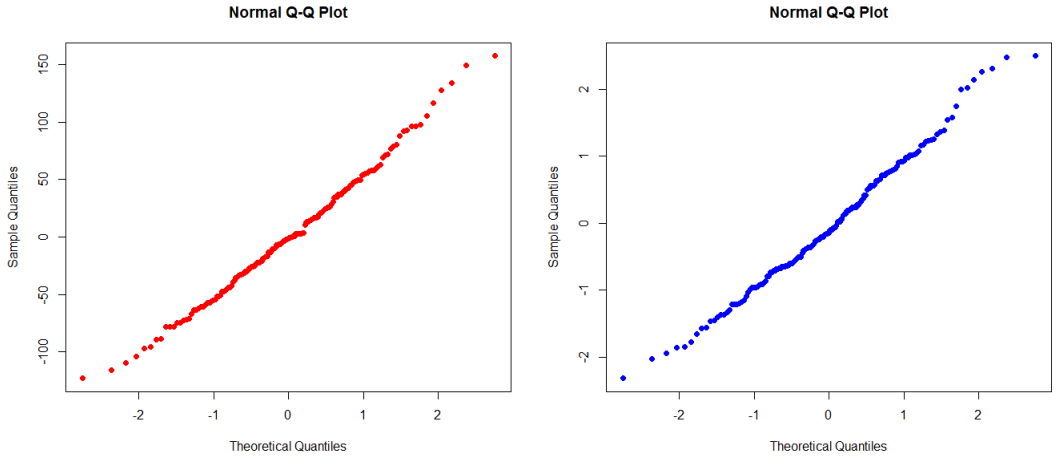
```
> 1-var(RES.rain.adj)/var(Y[13:168])
```

```
> qqnorm(RES.rain.adj,pch=19,col="red")
> qqnorm(Z,pch=19,col="blue")
```

I risultati ottenuti rivelano lati positivi e lati negativi dell'analisi. Da una parte, la varianza spiegata è incredibilmente bassa:

```
[1] 0.1128825
```

Dall'altra, però, i residui sembrano proprio seguire una distribuzione gaussiana, quindi non si poteva fare molto meglio:



Osservazione Ciò è dovuto, in parte, al fatto che l'intensità della pioggia è una variabile abbastanza casuale, a differenza ad esempio della temperatura minima, che in genere segue un andamento sinusoidale molto più accentuato.

Vediamo ora di analizzare i esidui derivanti dall'uso del metodo regressivo: in questo caso, i comandi per estrapolarli sono un po' diversi.

Innanzitutto digitiamo:

```
> par(mfrow=c(1,1))
```

```
> L = length(V)
> A = matrix(nrow=L-12,ncol=13)
```

```
> for (k in 1:12) {
+ A[,k] = V[(13-k):(L-k)]
+ }
```

```
> A[,13] = V[13:L]
```

```
> fit = lm(A[,13]~A[,1]+A[,2]+A[,5]+A[,6]+A[,12])
```

```
> P = 1:(L+12)
> P[1:L] = V
> for (k in 1:12) {
+ P[L+k] = coef(fit) %*% c(1,P[L+k-1],P[L+k-2],P[L+k-5],P[L+k-6],P[L+k-12])
+ }
> Pplus = ts(P,frequency=12,start=c(2000,1))
```

A questo punto digitiamo:

```
> RES.rain.reg = residuals(fit)
```

per ottenere i residui.

Per concludere, possiamo digitare i vari comandi già visti per i residui precedenti: non riportiamo i grafici ottenuti. In particolare, i residui appaiono ancora sostanzialmente gaussiani (e ciò era abbastanza prevedibile), ma il fatto più rilevante (in senso positivo, per fortuna) è che la varianza spiegata diventa:

```
[1] 0.3340997
```

Il valore è ancora molto basso, ma molto più alto di quello determinato col metodo di Holt-Winters modificato: questo metodo, allora, appare migliore (e ciò si era già visto durante l'analisi degli scarti quadratici dei valori predetti).

Usando invece il metodo di Holt-Winters classico, si ottiene:

```
[1] 0.2766235
```

Nella scorsa sezione abbiamo costruito a mano un modello previsivo che è risultato più efficiente di tutti gli altri, basandoci su delle condiserazioni sul *trend*. Se si riguarda quanto fatto, si noterà che le scelte effettuate erano molto spinte verso la previsione, e poco verso l'analisi.

Per completezza, la varianza spiegata con quel metodo risulta uguale a:

```
[1] 0.251404
```

2.7 Incertezza della previsione

Vediamo ora di generare una banda di confidenza al 90% per la previsione, di tipo non parametrico, basandoci sul metodo che è risultato, in media, migliore degli altri visti nelle scorse sezioni: quello regressivo.

Digitiamo allora:

```
> quantile(RES.reg,0.05)
> quantile(RES.reg,0.95)
```

Si ottengono i seguenti valori, decisamente molto alti:

```
       5%
-68.86791
```

```
      95%
 78.76734
```

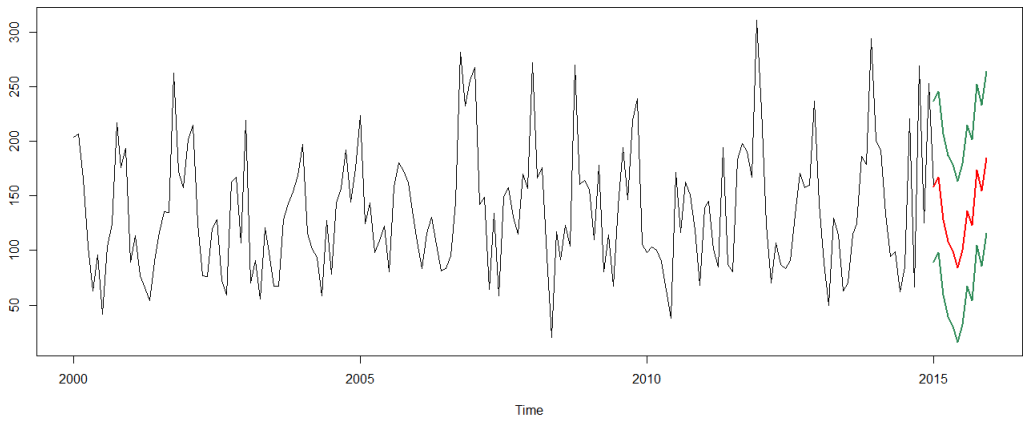
Per concludere, digitiamo:

```
> Y1 = Y
> Y1[181:192] = NA
> V1 = ts(Y1,frequency=12,start=c(2000,1))
> P1 = Pprev
> V1[181] = Pprev[181]
> P1[180] = NA
```

```
> Pmin = P1+quantile(residui,0.05)
> Pmax = P1+quantile(residui,0.95)
> Pmin[1:180] = Pmax[1:180]=NA
> Pmin = ts(Pmin,frequency=12,start=c(2000,1))
> Pmax = ts(Pmax,frequency=12,start=c(2000,1))
```

```
> ts.plot(V1,P1,Pmin,Pmax,gpars=list(lwd=c(1,2,2,2),
+ col=c("black","red","seagreen","seagreen")))
```

Si ottiene allora il grafico seguente, che come temevamo mostra delle bande di confidenza molto ampie (e di sicuro la situazione non migliora di molto usando un metodo di tipo parametrico):



Osservazione Concludiamo qui l'analisi della serie, visto che questa, a differenza di quella precedente, ci ha regalato ben poche soddisfazioni.