Using the R-package to forecast time series: ARIMA models and Application

2 authors:

Eralda Dhamo Gjika
University of Tirana

Llukan Puka
University of Tirana

Using the R-package to forecast time series: ARIMA models and Application

E.DHAMO, LL.PUKA

University of Tirana, Faculty of Natural Science, Department of Mathematics
E-mail: eralda.dhamo@unitir.edu.al, eralda_dhamo@yahoo.com
E-mail: lpuka2001@yahoo.co.uk

Abstract

Forecasting time series is a need in the financial sector or other fields, economic or not. We present here the software *R* as an important tool for forecasting and especially for studying the time series models. *R* has many features in common with both functional and object orientated programming languages. It is a widely used environment for statistical analysis. The striking difference between *R* and other statistical software is that it is a free software and that is maintained by scientists for the scientists. In particular, functions in *R* are treated as objects that can be manipulated or used recursively. The packages can be installed free of charge from the internet site www.r-project.org. "An Introduction to *R*", is also available via the *R* help system.

The paper describes some tools of *R* related to the time series modeling by ARIMA processes, providing graphical and numerical results for some real data. Criterions like AIC or other are used to choose the best forecasting method. We show that the best way to learn to do a time series analysis in *R* is through practice and 'hands-on' experience.

Keywords: Time series, forecast, *R*, ARIMA, AIC criterion, modeling real data.

1. Introduction and motivation

Knowledge of demographic trends, (total population, annual number of births, mortality, migration etc), is essential for drawing up appropriate government policies in the social, economic, education, housing, migration and regional planning fields.

Time series (univariate or not) automatic forecast is very helpful for this purpose. It is often quite hard to track one suitably trained to produce time series models, so an automatic forecast algorithm is an essential tool. Automatic forecast algorithm must determine an appropriate time series model, estimate the parameters of the model and compute the forecast. The most used automatic forecast algorithms, are based on exponential smoothing or ARIMA models.

R is a software and programming language that enables one to study time series effectively.

This paper is organized as follows: Section 2 gives details of the data used and introduces the methods of exponential smoothing, ETS and ARIMA. Section 3 introduces the forecast package in *R* language and fits the time series (the number of births in Albania from 1990 to 2008). We compare the performance of the already aforementioned models in Section 2. These models may be used then for life tables, mortality tables etc. In Section 4 we present some comments and findings.

2. The data, Exponential smoothing, ETS and ARIMA models

The data used to demonstrate the *forecast* package in *R*, is taken from the Albanian Institute of Statistics (INSTAT) (www.instat.gov.al). The data consist of the number of births per month from 1990 to 2008. The entire dataset contains 228 observations.

Rather than using the entire dataset, we first consider only the observation from January 1990 to December 2005, leaving aside the three years (from 2006 to 2008). Trying to build an appropriate model for the data, we compare three models for this subset of observations. The assessment of the best model is made through the predictive ability of each model.

The three models we consider, are: exponential smoothing, ETS and ARIMA.

Exponential smoothing methods have been around since the 1950s and were originally classified by Pegel's taxonomy (1969), extended later by other researchers, giving a total of fifteen methods. The table below shows the fifteen combinations.

Table 1. The fifteen exponential smoothing methods

		Seasonal Component		
		N	A	M
Trend Component		(No seasonality)	(Additive)	(Multiplicative)
N	(No trend)	N, N	N, A	N, M
A	(Additive)	A, N	A, A	A, M
A_d	(Additive damped)	A_d, N	A_d, A	A_d, M
M	(Multiplicative)	M, N	M, A	M, M
M_d	(Multiplicative damped)	M_d, N	M_d, A	M_d, M

Forecast methods are numerous and they improve continuously. Some of them are: *moving average, exponential smoothing, ARIMA, GARCH, Croston, Theta, cubic spline, Random Walk etc*. The forecast methods are classified in three main groups:

- *Univariate* - used of past models ex: moving average, trend.
- *Multivariate* - used of past relation between multivariate variables ex: regression analysis.
- *Qualitative* - used of subjective judgement and other information.

In Table 1, the commonly used methods are: cell (N,N) which describes the simple exponential smoothing method (or SES), cell (A,N) describes Holt's linear method, cell (A,A) describes the additive Holt-Winter's method and cell (A,M) gives the multiplicative Holt-Winter's method.

The first model, an *exponential smoothing model*, is an algorithm producing point forecast only. The *second model* we consider, is proposed by Hyndman (2008) and it is noted (E,T,S). The triplet (E,T,S) refers to the three components: error, trend and seasonality. The notation helps in remembering the order in which the components are specified.

The third model we consider in this paper, is the *ARIMA model*. Many people working with forecast, have difficulty using Autoregresive Integrated Moving Average (ARIMA) because of the order selection process. Actually many researchers have proposed methods to identify the order of an ARIMA model (Makridakis and Hibbon (2000); Liu (1989); Goodrich (2000); Reilly (2000)).
For non-seasonal data we can consider ARIMA (p, d, q) models and for seasonal data we can consider ARIMA (p, d, q) $(P, D, Q)_m$ where *m* is the seasonal frequency. Based on the model of Box and Jenkins (1970) the seasonal autoregressive integrated moving average model is given by:

$$\Phi_p(B^s)\phi(B)\nabla_s^D\nabla^d X_t = \alpha + \Theta_Q(B^s)\theta(B)w_t$$

Where,
s = seasonal lag,
ϕ = coefficient for AR process,
Φ = coefficient for seasonal AR process,
θ = coefficient for MA process,
Θ = coefficient for seasonal MA process.

B is the backward shift operator, $\nabla_s^D = (1 - B^s)^D$ and $\nabla^d = (1- B)^d$, w_t is an uncorrelated random variable with mean zero and constant variance.

3. Forecast package in R

All methods are acceptable in certain circumtances, but the quality of forecast is related with the selection of the model. The calculations need time and often are doubvious for the fitness of the model. Generally, there is not one method that performs better in all time series. For time series with different specification, there are different methods that perform in a more efficient way.

Use of forecasting techniques in *R* language, needs installation of some statistics package. Some of the main support package for forecast in *R* are: **expsmooth, Mcomp, fma, pastec, psych, Hmisc, nls2, nlme, dynlm, dynamicGraph, lmtest, psplin;** see for details in http://CRAN.R-project.org/package=forecasting .

3.1 Reading and organisation of data in R

In most cases in our practice, the data is saved in an Excel format. The data from INSTAT-Albania is saved in Excel 2007 format. Working with this data in *R* enviroment, means that they must be saved in ***csv*** format (*comma separated delimited*). The comand `read.csv (file.choose())` is one of the reading commands to transfer an Excel sheet in *R*.

When working with time series in *R*, first, the data must be converted in a time series format so that *R* may recognise. So, we **convert in a time series** the number of births per month from January 1990 to December 2005 using the `ts()` command.

```
> T=read.csv(file.choose())#chooses the excel file where are the
data. T is the vector of values from January 1990 to December
2005.
> SS=ts(T)  #converts the data in a time series format
> SS

Time Series:
Start = 1
End = 192
Frequency = 1
             X0
      [1,] 6077
      [2,] 6488
            ...
> plot(SS, main="Number of Births form 1990 to 2005",ylab="Number
of births",col="blue")
```

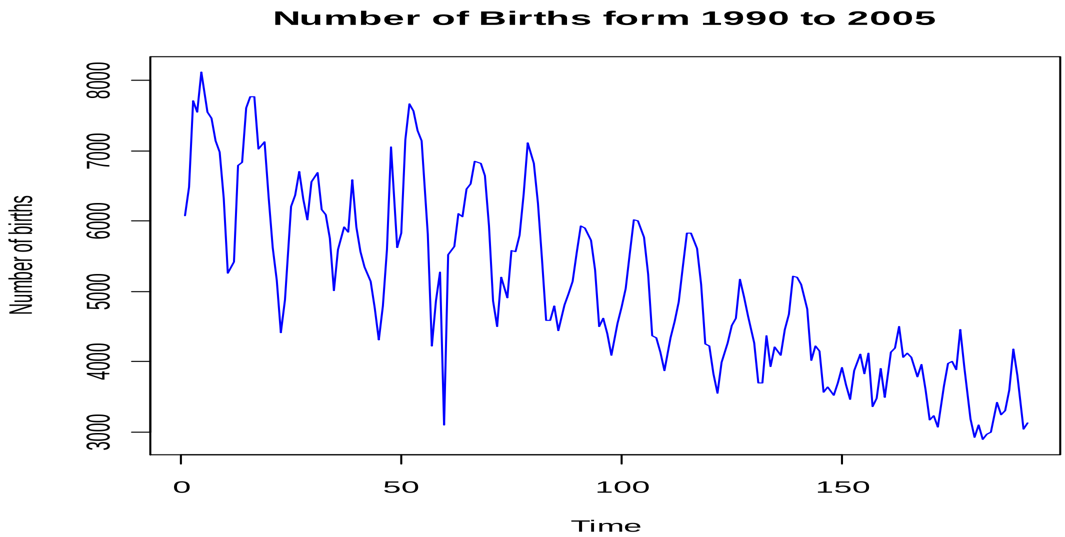

Figure 1. The number of births per month form January 1990 to December 2005

To build the model we need to arrange the data with frequency *m=12*. Data as follows:

```
> SS5=ts(SS,start=1990,frequency=12)
> SS5
       Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov
1990   6077  6488  7720  7555  8130  7555  7473  7145  6981  6324  5257
...          ...   ...   ...   ...   ...   ...   ...   ...   ...
2005   3092  2887  2957  2991  3419  3245  3307  3586  4177  3795  3034
       Dec
1990   3122
...    ...
2005   5420
```

3.2 Exponential smoothing, Holt Winter's method

Using the `HoltWinters()`, we smooth the time series and find the smoothing parametres.

```
> HW=HoltWinters(SS5)
```

In our data the model is: Holt-Winter's exponential smoothing with trend and additive seasonal component and the smoothing parameters are: *alpha: 0.7294214 ; beta : 0; gamma: 1* (see the smoothing in Figure 2).

```
> plot(HW,col="blue")
```

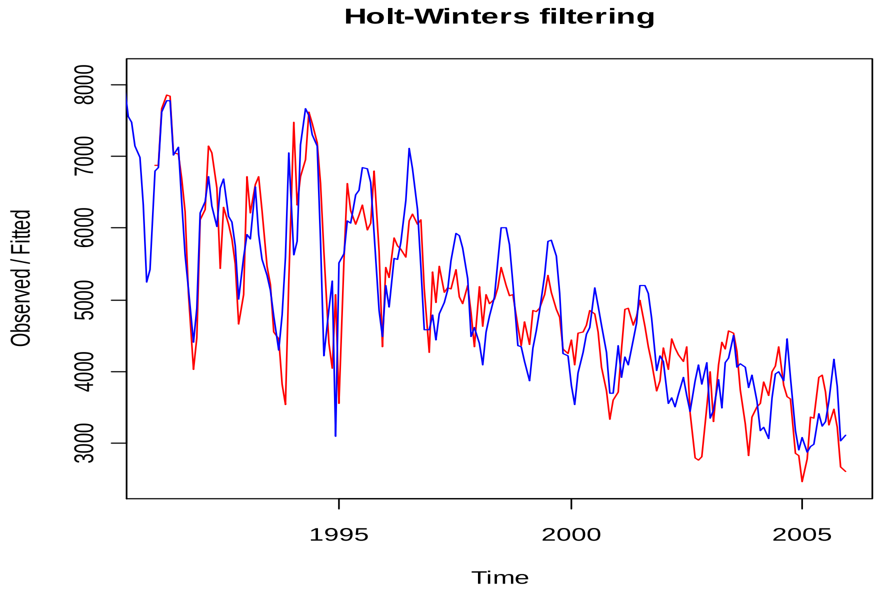

Figure 2. The Holt-Winter's smoothing

(***blu line***= *the original data*, ***red line***= *the smoothed data*)

In order to judge the accuracy of the model, we consider the data from January 1990 to December 2008. The accuracy of the forecast is evaluated by one of the following criteria according to circumstance: *ME*, *RMSE*, *MAE*, *MPE*, *MAPE*, *MASE*, *AIC*, AIC_C, *BIC*.

```
>HoltWintersForecast<-predict(HW,36,prediction.interval= TRUE)
# the smoothing time series and the forecast with the prediction
interval
> HoltWintersForecast
            Fit         upr         lwr
Jan 2006    2707.779    3899.939    1515.61869
Feb 2006    2430.338    3905.949    954.72754
            ...         ...         ...
```

A plot of the real data, forecasted values and intervals is shown in the following:

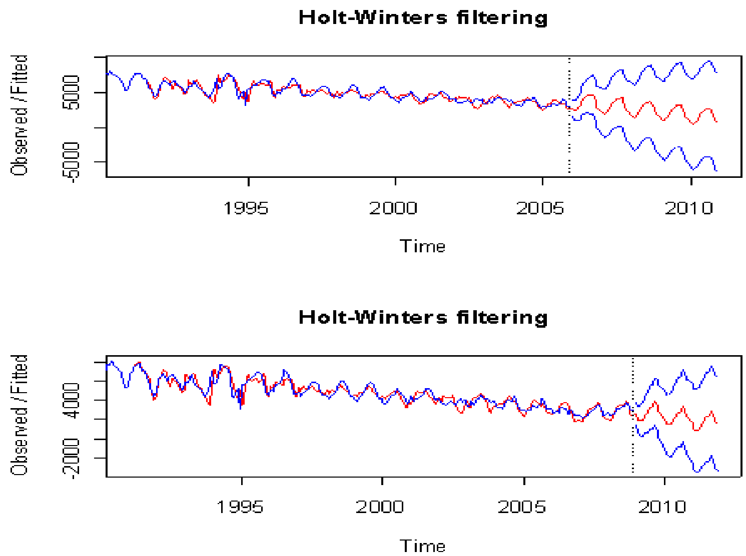

Figure 3. The Holt-Winter's smoothing (***blu line***= *the original data and the upper and lower border of the forecast interval,* ***red line***= *the smoothed data*)

A detailed graph of the observed data from January 1990 to December 2008, the smoothed values using the time series with observations 1990-2005 and the smoothed values using the time series with observation 1990 – 2008 is shown below.

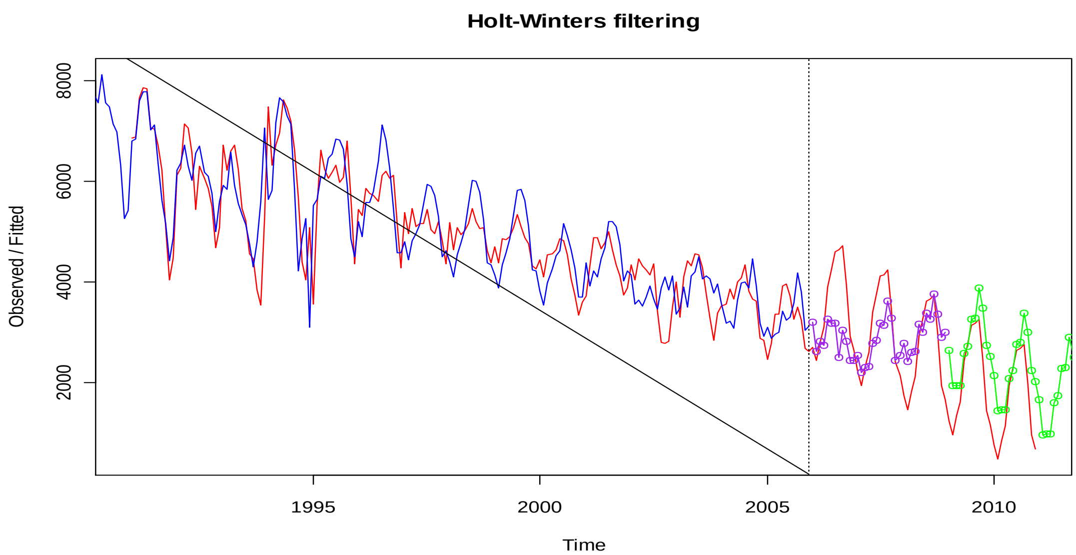

Figure 4. Holt-Winter's forecast (***blue line***=*the original data '90-'05,* ***red line***= *smoothed data based on original data '90-'05,* ***purple line***= *original data '06-'08,* ***green line***= *smoothed data based on the original data '90-'08*)

3.3 Second model : ETS model

ets() function gives only the model for the observed values and the smoothing parameters.

The *summary ()* command gives a summary of the model, the smoothing parameters and the forecasted values. So, the model for our data is: ETS(A,N,A) which has additive errors, no trend and additive seasonality.

```
> summary(forecast(SS))
Forecast method: ETS(A,N,A)
Model Information:
ETS(A,N,A)
Call:
 ets(y = object)
  Smoothing parameters:
    alpha = 0.9999
    gamma = 1e-04
  Initial states:
    l = 7499.4998
    s = -560.9959 -627.2871 -90.4 205.6223 397.7078 696.0149
         416.9466 299.9661 39.389 43.8771 -435.9757 -384.8651
  sigma:  480.7723
     AIC     AICc      BIC
3408.790 3411.163 3454.395
In-sample error measures:
      ME      RMSE              MAE         MPE        MAPE
-19.8798397 480.7723060 329.4195487  -0.9382922   6.7947608
  MASE
0.7959913
Forecasts:
Point        Forecast  Lo 80     Hi 80      Lo 95       Hi 95
Jan 2006     3298.174 2682.0391 3914.308  2355.87724   4240.470
  …          ….          ….         ….          …
Dec 2007     3122.000  103.8945 6140.105  -1493.79378 7737.793
```

We compare the values fitted from the ETS model (for the first 192 observations of the time series) and the whole time series (228 observations) in the graph derived by the commands:

```
>par(mfrow=c(2,1))
>plot(forecast(SS),xlab="Time",ylab="Number of births")
>lines(SS8,type="l",ylim=R,col="black")
>plot(forecast(SS8),xlab="Time",ylab="Number of births")
```

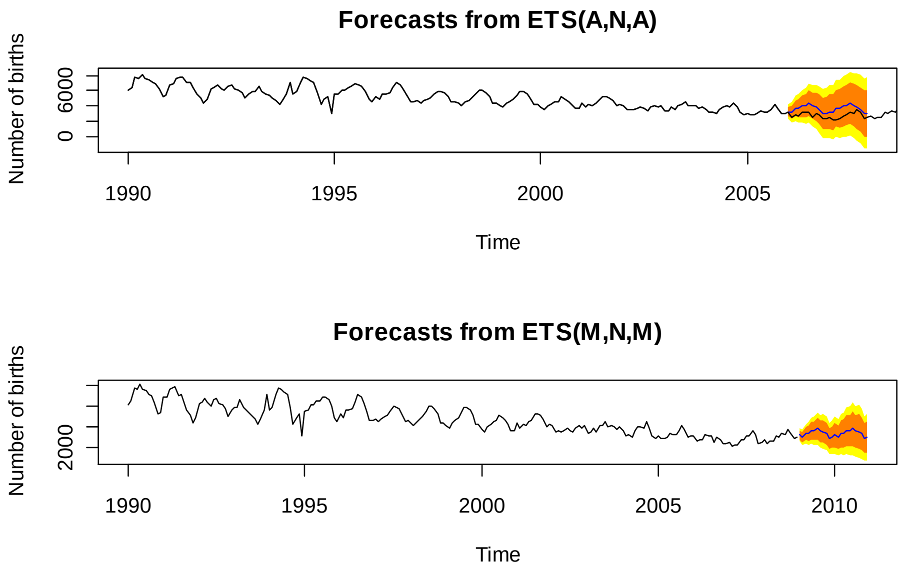

Figure 5. The first graph shows the forecasted value and intervals of confidence from ETS(A,N A) model of number of births (***black line***= *original data '90-'08,* ***blue line***=*forecast from ETS model,* ***orange zone***=*interval of confidence 80%,* ***yellow zone***=*interval of confidence 95%).*The second graph shows the time series (1990-2008) the forecasted values and intervals of confidence from the ETS(M,N,M) model.

As seen from Figure 5 the models are different. First model is ETS(A,N,A) and then when we add the data from January 2006 to December 2008 the model becomes an ETS(M,N,M).

Figure 6 shows the time series with all the observation value from January 1990 to December 2008.

```
> plot(SS8,main="Number of Births form 1990 to 2008",ylab="Number
of births",col="blue")# SS8 is the time series with 228 values
                       (1990 to 2008)
```

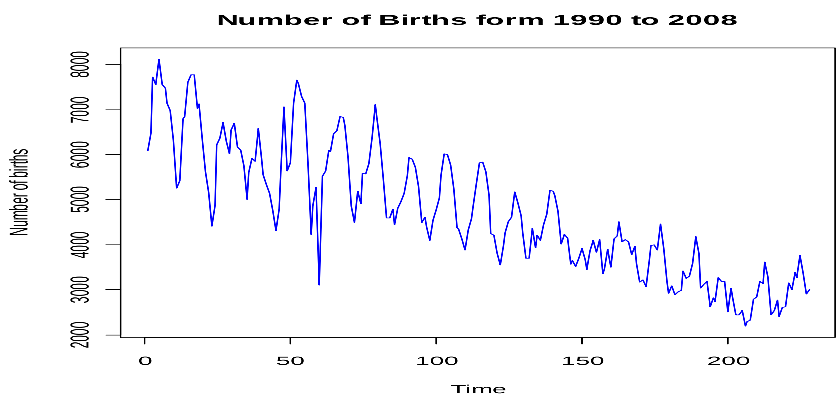

Figure 6. Original data taken from INSTAT Albania of number of births per month form 1990 to 2008

3.4 Third model: ARIMA model

To fit the data by ARIMA model we use the command:

```
 > auto.arima(SS5) # SS5 time series data from 1990-2005
Series: SS5
ARIMA(1,1,1)(1,0,1)[12]
Call: auto.arima(x = SS5)
Coefficients:
         ar1      ma1    sar1     sma1
      0.7004  -0.9657  0.9307  -0.7090
s.e.  0.0582   0.0155  0.0425   0.0871

sigma^2 estimated as 218215:  log likelihood = -1448.58
AIC = 2907.16   AICc = 2907.48   BIC = 2923.42

> auto.arima(SS8) # SS8 time series data from 1990-2008
Series: SS8
ARIMA(1,1,1)(1,0,1)[12]
Call: auto.arima(x = SS8)
Coefficients:
         ar1      ma1    sar1     sma1
      0.7075  -0.9674  0.9428  -0.7175
s.e.  0.0528   0.0142  0.0323   0.0722

sigma^2 estimated as 188954:  log likelihood = -1705.24
AIC = 3420.48   AICc = 3420.76   BIC = 3437.61



> forecast(auto.arima(SS8))
        Point Forecast   Lo 80      Hi 80      Lo 95      Hi 95
Jan 2009    2932.194   2375.117   3489.270   2080.2184   3784.169
…           …              … …        …       …        …       …
Dec 2010    2447.406   1497.329   3397.482    994.3883   3900.423
```

The model given by the command `auto.arima()` is a SARIMA (Seasonal Autoregressive Integrated Moving Average) model and is denoted as an ARIMA(p,d,q)(P,D,Q)[m] where *m* is the seasonal component.

So the process is:

`ARIMA(1,1,1)(1,0,1)[12]` (for the data from 1990 to 2005)

$s = 12$, $\phi = 0.7004$, $\Phi = 0.9307$, $\theta = -0.9657$, $\Theta = -0.7090$

And,

`ARIMA(1,1,1)(1,0,1)[12]` (for the data from 1990 to 2008)

$s = 12$, $\phi = 0.7075$, $\Phi = 0.9428$, $\theta = -0.9674$, $\Theta = -0.7175$

Graphical views of the models are shown below:

```
> par(mfrow=c(2,1))
> plot(forecast(auto.arima(SS5)),xlab="Time",ylab="Number of
births")#SS5 time series from ‘90-‘05
> lines(SS8,type="l",ylim=R,col="black")#SS8 time series from
’90-‘08
> plot(forecast(auto.arima(SS8)),xlab="Time",ylab="Number of
births")
```

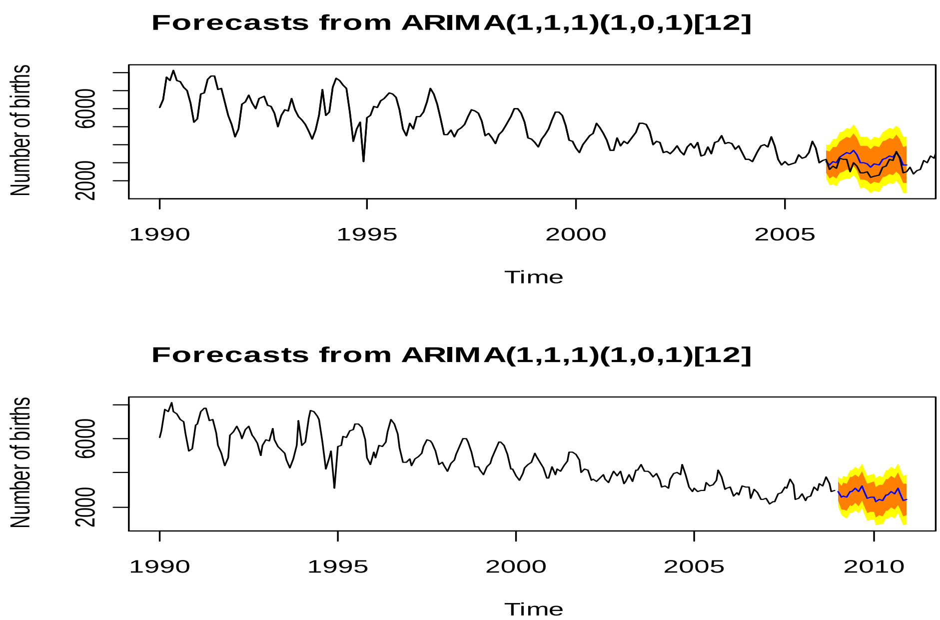

Figure 7. The first graph in Figure 7 shows the forecasted data and the confidence intervals from ARIMA(1,1,1)(1,0,1)[12] (1990 to 2005) model (***black line***= *original data ’90-’08,* ***blue line*** = *forecast,* ***orange zone****=interval of confidence 80%,* ***yellow zone****=interval of confidence 95%).*
The second graph shows the time series (1990-2008) and the forecast of ARIMA(1,1,1)(1,0,1)[12] model for the period January 2009 to December 2010.

As seen from the above, the model does not alter when we add the three years that we had previously excluded (2006, 2007 and 2008) but the coefficients does.

Table 2 and **Table 3** show some of the results and criteria used for the evaluation of the models.

Table 2. The Exponential smoothing and the ETS models results

Coefficients	Exp.Smoothing ’90-‘05	Exp.Smoothing ’90-‘08	ETS (A,N,A)	ETS (M,N,M)
alpha	0.7294214	0.6811137	0.9999	0.7654
beta	0	0		
gamma	1	1	1 e-04	1 e-04
a	3693.27807	2826.18221		
b	-40.81367	-40.81367		
s1	-944.68574	-156.94856	-560.9959	0.8879
s2	-1181.31251	-815.97413	-627.2871	0.8668
s3	-770.64387	-757.11361	-90.4	0.9898
s4	-433.74647	-715.89701	205.6223	1.0572
s5	403.08912	-53.79534	397.7078	1.063
s6	793.09847	145.47920	696.0149	1.1367
s7	1194.72672	716.06274	416.9466	1.0807
s8	1270.97556	778.20432	299.9661	1.0722
s9	1400.52101	1409.93459	39.389	1.0012
s10	648.44785	1065.26665	43.8771	0.99
s11	-330.24878	353.50813	-435.9757	0.8972
s12	-571.27807	179.81779	-384.8651	0.9573
ME			-19.8798397	-24.5326955
RMSE			480.7723060	465.4780546
MAE			329.4195487	323.1767185
MPE			-0.9382922	-1.0516275
MAPE			6.7947608	6.9702644
MASE			0.7959913	0.8231815
AIC			3408.790	4036.809
AIC_C			3411.163	4038.781
BIC			3454.395	4084.820

Table 3. The SARIMA models

Model	ar1	ma1	sar1	sma1	AIC	AIC_C	BIC
ARIMA(1,1,1)(1,0,1)[12] 1990-2005	0.7004	-0.9657	0.9307	-0.7090	2907.16	2907.48	2923.42
s.e	0.0582	0.0155	0.0425	0.0871			
ARIMA(1,1,1)(1,0,1)[12] 1990-2008	0.7075	-0.9674	0.9428	-0.7175	3420.48	3420.76	3437.61
s.e	0.0528	0.0142	0.0323	0.0722			

4. Results and comments

We used the forecast package of *R* language to show how to work on time series, when the main aim is forecast. The forecast package in *R* language works as an automatic forecast algorithm which can determine an appropriate time series model, estimate the parameters of the model and compute the forecast. *R* achieves the evaluation and forecast in a few seconds on modern computers. As *R* is created and maintained by scientist for the scientist, it is a language which can help not only in the applied field but in the research field as well.

We want to emphasize that, finding the model with the lowest error criterion used, does not imply that it is the best model. The best way of choosing the appropriate model between those offered by the forecast package in *R* is to know how to manage your data and understand them.

References

Hyndman, R. J., Koehler, A. B., Snyder, R. D. & Grose, S. (2002): A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting* **18**, 439–454.

Hyndman R.J. ,King M.L., Pitrun I., Billah B. (2005): Local linear forecast using cubic smoothing splines. Australian & New Zealand Journal of Statistics, Volume 47, Issue 1 87-99

Hyndman R. J., Athanasopoulos G., Song H., Wu D.C., (2008): The tourism forecasting Competition

Hyndman R.J., Khandakar Y. (2008)**:** Automatic Time Series Forecasting: The forecast Package for R, Monash University, *Journal of Statistical Software*, Volume 27, Issue 3. (http://www.jstatsoft.org)

Hydman R.J., Kostenko A.V. (2007): Minimum sample size requirements for seasonal forecasting models

Pegels, C. C. (1969) Exponential smoothing: some new variations, *Management Science*, **12**, 311–315.

http://cran.r-project.org/web/views/TimeSeries.html,
http://CRAN.R-project.org/package=forecasting.
http://www.stat.pitt.edu/stoffer/tsa2/Examples.html

Some useful commands in forecast package (R)

abline( )	Graph command
acf( )	Autocorrelation function
arima( )	Fit an ARIMA model to the data
arima.sim( )	Simulate an ARIMA model
c( )	Command for a vector
diff( )	Vector of difference
filter( )	Filter a time series
hist( )	Histogram
HoltWinters( )	Holt Winter's procedure
length( )	Length of a vector
lines( )	Graph command
lm( )	Linear model
log( )	Logarithm of values
lsfit( )	Least square estimation
mean( )	Arithmetic mean
pacf( )	Partial autocorrelation function.
plot( )	Graph command
predict( )	Prediction
read.csv( )	Import data in *csv* format
rep( )	Vectorial command
sd( )	Standard deviation
seq( )	Vectorial command
shapiro.test( )	Shapiro–Wilk test
stl( )	Seasonal decomposition of time series
summary( )	Summary function
ts( )	Converts the data to a time series
tsdiag( )	Diagnostic of time series
qqnorm( )	QQ plot
par()	Activate a new window
lag.plot()	Lag- plot
auto.arima()	Fit an ARIMA model to the data
kpss.test()	Kwiatkowski, Phillips, Schmidt and Shin (1992) test
pp.test()	Phillips-Perron Unit Root Tests
adf.test()	ADF test