

# Development of a Modelling Script of Time Series Suitable for Data Mining

Víctor Sanz-Fernández<sup>1</sup>, Remedios Cabrera<sup>1\*</sup>, Rubén Muñoz-Lechuga<sup>1</sup>,  
Antonio Sánchez-Navas<sup>2</sup>, Ivone A. Czerwinski<sup>1</sup>

<sup>1</sup>Departamento de Biología, Facultad de Ciencias del Mar y Ambientales, Universidad de Cádiz, Campus de Excelencia Internacional del Mar (CEIMAR), Puerto Real, Cádiz, Spain

<sup>2</sup>Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Universidad de Cádiz, Puerto Real, Cádiz, Spain

Email: \*reme.cabrera@uca.es

Received 30 May 2016; accepted 19 July 2016; published 22 July 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

Data Mining has become an important technique for the exploration and extraction of data in numerous and various research projects in different fields (technology, information technology, business, the environment, economics, etc.). In the context of the analysis and visualisation of large amounts of data extracted using Data Mining on a temporary basis (time-series), free software such as R has appeared in the international context as a perfect inexpensive and efficient tool of exploitation and visualisation of time series. This has allowed the development of models, which help to extract the most relevant information from large volumes of data. In this regard, a script has been developed with the goal of implementing ARIMA models, showing these as useful and quick mechanisms for the extraction, analysis and visualisation of large data volumes, in addition to presenting the great advantage of being applied in multiple branches of knowledge from economy, demography, physics, mathematics and fisheries among others. Therefore, ARIMA models appear as a Data Mining technique, offering reliable, robust and high-quality results, to help validate and sustain the research carried out.

## Keywords

Data Mining, ARIMA Models, Time Series, Script, R

---

## 1. Literature review

During the last few years and due to the numerous advances in information systems, the use of means designed for

\*Corresponding author.

**How to cite this paper:** Sanz-Fernández, V., Cabrera, R., Muñoz-Lechuga, R., Sánchez-Navas, A. and Czerwinski, I.A. (2016) Development of a Modelling Script of Time Series Suitable for Data Mining. *Open Journal of Statistics*, 6, 555-564. <http://dx.doi.org/10.4236/ojs.2016.64047>

the acquisition of data such as web and mobile applications, social networks, etc. has massively increased. As a result of this “information revolution”, the world of science has been saturated with data of varied origin. It is estimated that 90% of all data have been created in the last two years (2013-2015) [1]. At the IOD (Information On Demand) Conference held in 2011, IBM presented the explosion of data in today’s society as a problem, and put forward how companies are facing the challenge of obtaining relevant and valuable information from this vast amount of data. The amount of data in the world is expected to double every two years, according to the data scientist Mark van Rijmenam, founder of *Datafloq*, in addition to increase 2.5 exabytes per day [1].

This enormous amount of information is known as Big Data. The vast majority of these data, which come from astronomy, genomics, telephony, credit card transactions, Internet traffic and web information processing, primarily, are acquired systematically with a certain frequency, being therefore time series [2]-[4]. The tendency to manipulate large quantities of data is due to the need, in many cases, to include the data obtained from the analysis of large databases in new databases, such as business analyses [5]. Besides data manageability, other factors to consider are the speed of analysis/scanning speed, access, search and return of any element. It is important to understand that conventional databases are a significant and relevant part of an analytical solution [6] [7].

Today, the explosion of data poses a problem given the amount of these increases overwhelmingly; in fact this situation reaches occasionally the point when it is not possible to gain an useful insight from them. Therefore, it is necessary to organise, classify, quantify and of course exploit this information to obtain maximum performance for the benefit of scientific research. In response to this difficulty the concept of Data Mining arises that refers to the non-trivial automated process which identifies valid, previously unknown, potentially useful and fundamentally understandable patterns in the data.

The literature shows that Data Mining techniques are used to extract information from very diverse backgrounds as the power consumption of a region [8], modelling and optimisation of wastewater pumping systems [9] and the establishment of the position of wind turbines to obtain the maximum possible wind currents [10].

A common pattern of all previous studies is the use of time series for the analysis and visualisation of information. A way to perform the processing of time series is through the creation of mathematical models that identify and predict their behaviour. One of these are the ARIMA models [11], that extract the most relevant data from the dataset identifying the patterns of the series at different levels of the timescale and simplify a large amount of data in a simple equation, hence their utility and application in Data Mining. ARIMA models are within the Data Mining techniques, as these are used in time series, therefore being a very useful tool to extract relevant information from Big Data.

In the field of the analysis and visualisation of data, the development of free software is a good tool for both analytical and visual integration of information. In this section of the processing of data, software for the analysis and visualisation of data allow to work with large volumes of data completed over a period of time [12]. The development of statistical software that allows to work on the analysis of time series further facilitates the implementation of ARIMA models.

The use of free access software as Rstudio, which is an integrated development environment for R, has the advantage of enabling programming statistical packages as required, as well as of applying all kinds of time series analysis, in addition to reducing economic costs in any research project. In the present work a script has been developed in the environment of programming R language that allows the implementation, processing and visualisation of ARIMA models, in order to make it easier for scientists to know about the exploration, exploitation and manipulation of large volumes of univariate data carrying associated timescales. The script development and implementation structure is shown in **Figure 1**.

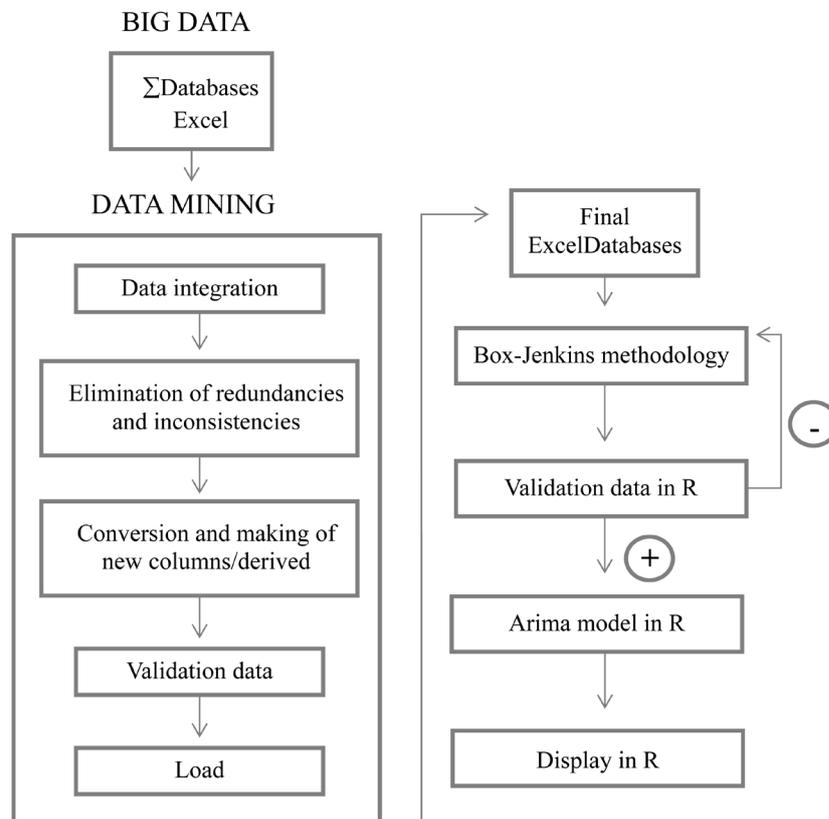
Therefore, the use of this script achieves the implementation of the Box-Jenkins methodology [11] for the development of ARIMA-models; in this way, the researcher is able to decompose the time series and to obtain the most relevant information of the characteristics of the temporal series, showing the extent to which this script helps in the exploration, exploitation and manipulation of data.

## 2. Information about the ST.R File

This document provides information about what is and how to use the ST.R script.

### 2.1. What Is ST.R?

ST.R is a code in R language developed for the treatment of time series and the realisation of ARIMA models following the Box-Jenkins methodology [11]. The script is split into two blocks. The first one is a collection of



**Figure 1.** Structure of development and implementation of the script in R. The different actions to be followed for the implementation of the script are shown. It is a conceptual model of implementation where Excel is used as a possible tool for data management. Source: own elaboration.

commands for the numerical and graphic description of the time series, and the development of the ARIMA models. In the second block the commands of different precision measurements are set up, which allow to compare the forecasts made by the models with the actual data with the aim of selecting the model with the most optimal fit to actual observations [13] [14].

## 2.2. How to Use ST.R

In order to successfully run the ST.R script, the necessary libraries are `lmtest` and `tseries`. These libraries are available from the repository Comprehensive R Archive Network (CRAN) at <http://CRAN.R-project.org/package=OptGS>. In this work, the R “stat” package version 3.3.0 was used, using “ARIMA” argument. The fitting methods are described in the R manual [15].

## 2.3. ST.R Structure

1) Graphical representation: graphical representation of the time series to visualise its components (trend, cycle, stationarity and random or irregular component (Figure 2 and Figure 3).

2) Trend analysis: the existence or non-existence of the trend is studied from the graphical results. A linear trend will be removed with first differences. However, for a nonlinear trend two differences are used. The Dickey-Fuller [16] and KPSS [17] tests are used for the analysis (Figure 4).

3) Homocedasticity analysis: This is done from both a visual and a mathematical perspective. From a visual point of view, it is carried out through the study of the thickness of the series. If this thickness remains constant, with no major irregularities observed, the series will be homocedastic; otherwise, the series will be considered heterocedastic. From a mathematical, it is carried out with the application of the homoscedasticity Breusch-pagan test [18] (Figure 5).

```

1 #Library load
2 library(tseries)
3
4 #Graphical representation
5 #Definition of the time series with our own data
6
7 #X = Dataset Name, X1 =Dataset Main Variable
8 X<-X$X1
9 X<-ts(X,start=#series start year#
10      ,frequency= #series frequency#)
11
12 #Verification that the series has been defined correctly
13 class(X)
14 start(X)
15 end(X)
16 frequency(X)
17 cycle(X)
18 time(X)
19
20 #Graphical representation of the series
21 plot(X, type="o",col ="red")
22 plot(aggregate((X), FUN=mean), type="o",col ="red")
23
24 X.STL <- stl(X,s.window='periodic')
25 lines(X.STL$time.series[,2],col='red') # Trend
26 lines(X.STL$time.series[,1]+X.STL$time.series[,2],col='blue') # Trend + stationarity
27 moving.mean_x<-decompose(X)$trend
28 lines(moving.mean_x, col="red", lty="dashed")
29
30 #Moving mean filters|
31 mva3_X <- filter(X,filter=1/3*c(1,1,1))
32 mva5_X <- filter(X,filter=1/5*c(1,1,1,1,1))
33 mva7_X <- filter(X,filter=1/7*c(1,1,1,1,1,1,1))

```

Figure 2. An R Graphical User Interface (GUI) for step 1. Graphical representation.

```

34
35 #Correlation graphics
36 acf(X,lag.max=7)
37 pacf(X,lag.max=7)
38
39 #Analysis of dependence of correlation coefficients
40 Box.test(X, lag = 7, type = c("Box-Pierce", "Ljung-Box"))
41
42 #Durbin-watson test, to verify the hypothesis of no autocorrelation
43 #against the alternative of first-order autocorrelation
44 #under a scheme autoregressive -AR ( 1)
45 library(lmtest)
46 n<-length(X)
47 ti<-1:n
48 dwtest(X~ti)
49
50 #Model choice, additive or multiplicative
51 #Additive model
52 dif<-diff(X)
53 sd(dif)/abs(mean(dif))
54 #Multiplicative model
55 inc<-X[-1]/X[-length(X)]
56 sd(inc)/abs(mean(inc))
57 #The chosen model has the lowest value
58
59 #Decomposition of the series in components
60 plot(decompose(X,type=#"additive or multiplicative"#))
61 desco_multi<-decompose(X,type=#"additive or multiplicative"#)
62 str(desco_multi)
63 X.STL <- stl(X,s.window='periodic', decompose(X,type=#"additive or multiplicative"#))
64 plot(X.STL)

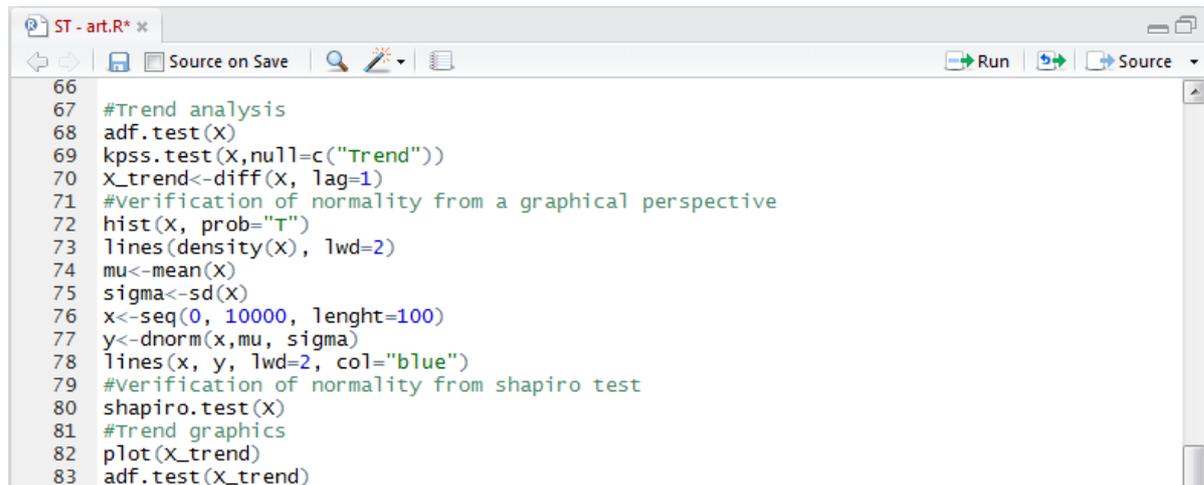
```

Figure 3. An R Graphical User Interface (GUI) for step 1. Graphical representation.

4) Stationarity analysis: As a result of the steps above, when neither seasonal cycle, nor trend, nor a significant thickness alteration of the series are to be perceived, the series is regarded as stationary (**Figure 5**).

5) Model identification: the most optimal model type is determined from the order of the Autoregressive procedure and moving averages of the constituents, both uniform and seasonal. This choice is made from autocorrelation (FAC) and partial autocorrelation (partial FAC) functions (**Figure 5**).

6) Estimation of the coefficients of the model: the order of the model having been established, the estimation of its parameters is made. Given it is an iterative calculation process, initial values (pool of models) can be suggested (**Figure 5**).

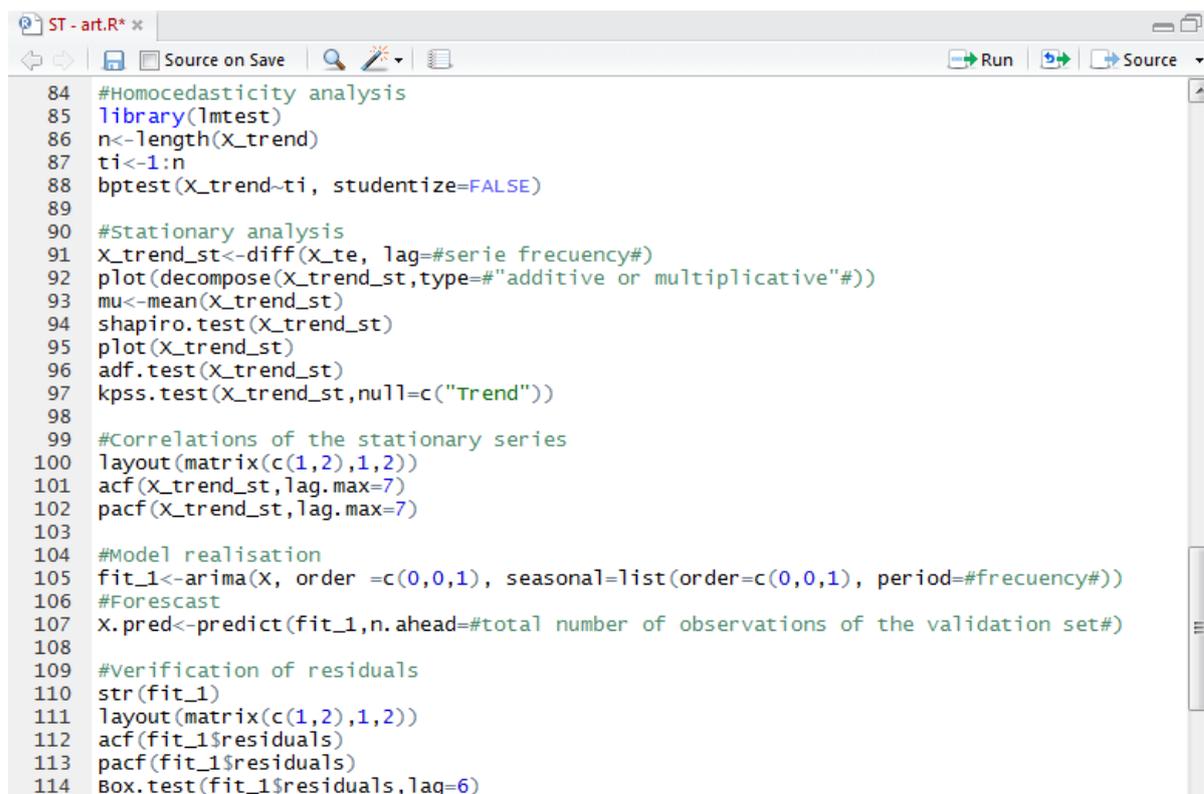


```

66
67 #Trend analysis
68 adf.test(X)
69 kpss.test(X,null=c("Trend"))
70 X_trend<-diff(X, lag=1)
71 #verification of normality from a graphical perspective
72 hist(X, prob="T")
73 lines(density(X), lwd=2)
74 mu<-mean(X)
75 sigma<-sd(X)
76 x<-seq(0, 10000, lenght=100)
77 y<-dnorm(x,mu, sigma)
78 lines(x, y, lwd=2, col="blue")
79 #verification of normality from shapiro test
80 shapiro.test(X)
81 #Trend graphics
82 plot(X_trend)
83 adf.test(X_trend)

```

**Figure 4.** An R Graphical User Interface (GUI) for step 2. Trend analysis.



```

84 #Homocedasticity analysis
85 library(lmtest)
86 n<-length(X_trend)
87 ti<-1:n
88 bptest(X_trend~ti, studentize=FALSE)
89
90 #Stationary analysis
91 X_trend_st<-diff(X_te, lag=#serie frequency#)
92 plot(decompose(X_trend_st,type="#additive or multiplicative#"))
93 mu<-mean(X_trend_st)
94 shapiro.test(X_trend_st)
95 plot(X_trend_st)
96 adf.test(X_trend_st)
97 kpss.test(X_trend_st,null=c("Trend"))
98
99 #Correlations of the stationary series
100 layout(matrix(c(1,2),1,2))
101 acf(X_trend_st,lag.max=7)
102 pacf(X_trend_st,lag.max=7)
103
104 #Model realisation
105 fit_1<-arima(X, order =c(0,0,1), seasonal=list(order=c(0,0,1), period=#frequency#))
106 #Forecast
107 X.pred<-predict(fit_1,n.ahead=#total number of observations of the validation set#)
108
109 #verification of residuals
110 str(fit_1)
111 layout(matrix(c(1,2),1,2))
112 acf(fit_1$residuals)
113 pacf(fit_1$residuals)
114 Box.test(fit_1$residuals,lag=6)

```

**Figure 5.** An R Graphical User Interface (GUI) for steps 3 - 7, 10. Homocedasticity analysis; stationarity analysis; model identification; estimation of the coefficients of the model; detailed error analysis; forecast.

7) Detailed error analysis: It is made from the verified differences between values observed empirically and estimated by the model for their final assessment. It is necessary to check an inconsistent regime of them and analyse the existence of significant errors. The Ljung-Box test is applied [19] (Figure 5).

8) Contrast of model validity: the model or models initially selected are quantified and valued using various statistical measures. The measures applied are:  $R^2$  (coefficient of determination), % SEP (standard error percentage),  $E_2$  (coefficient of efficiency), ARV (average relative variance), AIC (Akaike information criterion), RMSE (root mean square error) and MAE (mean absolute error) (Figure 6).

9) Model selection: based on the results of the previous steps, the model to work on is decided upon (Figure 6).

10) Forecast: the most optimal model will be used as the prediction base tool (Figure 5).

## 2.4. ARIMA Models

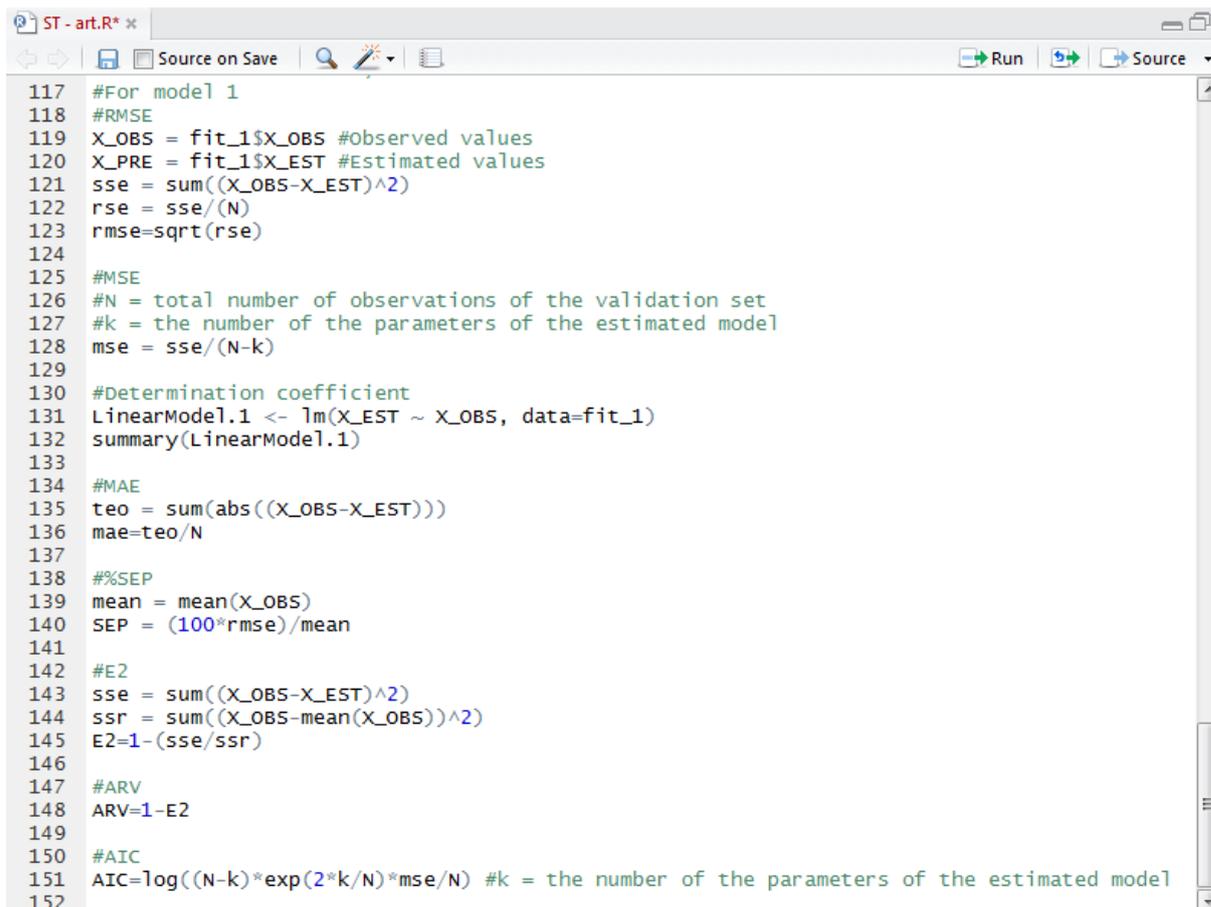
The univariate ARIMA models  $(p,d,q)$  [11] try to explain the behaviour of a time series from past observations of the series itself and from past forecast errors. The compact notation of the ARIMA models is as follows:

$$\text{ARIMA}(p,d,q) \quad (1)$$

where  $p$  is the number of autoregressive parameters,  $d$  is the number of differentiations for the series to be stationary, and  $q$  is the number of parameters of moving averages. The Box-Jenkins model  $(p,q)$  is represented by the following equation:

$$Y_t = \phi_0 + \phi_1 y_{t-1} + L + \phi_p y_{t-p} + a_t - \theta_1 a_{t-1} - L - \theta_q a_{t-q} \quad (2)$$

The autoregressive part (AR) of the model is  $\phi_1 y_{t-1} + L + \phi_p y_{t-p}$ , while the part of moving averages of the model (MA) is  $-\theta_1 a_{t-1} - L - \theta_q a_{t-q}$ . The coefficients of the parameters  $\phi_0$ ,  $\phi_1$ ,  $L$ ,  $\phi_p$ ,  $\theta_1$ ,  $\theta_q$  are determined



```

117 #For model 1
118 #RMSE
119 X_OBS = fit_1$X_OBS #Observed values
120 X_PRE = fit_1$X_EST #Estimated values
121 sse = sum((X_OBS-X_EST)^2)
122 rse = sse/(N)
123 rmse=sqrt(rse)
124
125 #MSE
126 #N = total number of observations of the validation set
127 #k = the number of the parameters of the estimated model
128 mse = sse/(N-k)
129
130 #Determination coefficient
131 LinearModel.1 <- lm(X_EST ~ X_OBS, data=fit_1)
132 summary(LinearModel.1)
133
134 #MAE
135 teo = sum(abs((X_OBS-X_EST)))
136 mae=teo/N
137
138 #%SEP
139 mean = mean(X_OBS)
140 SEP = (100*rmse)/mean
141
142 #E2
143 sse = sum((X_OBS-X_EST)^2)
144 sssr = sum((X_OBS-mean(X_OBS))^2)
145 E2=1-(sse/ssr)
146
147 #ARV
148 ARV=1-E2
149
150 #AIC
151 AIC=log((N-k)*exp(2*k/N)*mse/N) #k = the number of the parameters of the estimated model
152

```

Figure 6. An R Graphical User Interface (GUI) for steps 8 and 9. Contrast of model validity; model selection.

from the data, by means of any consistent statistic. The ARIMA models allow fitting the trend plus the stationarity in data. In this case, the model is noted as:

$$\text{ARIMA}(p, d, q)(P, D, Q)^S \quad (3)$$

where  $P$  is the number of autoregressive parameters in the seasonal part,  $D$  is the number of differentiations for the series to be seasonal in the seasonal part,  $Q$  is the number of parameters of moving averages in the seasonal part and  $S$  is the series frequency.

The Box-Jenkins method provides forecasts without any previous conditions, apart from being parsimonious with regard to coefficients [20]. Once the model has been found, forecasts and comparisons between actual and estimated data for observations from the past can be done immediately [21].

The identification of the parameters  $p$ ,  $q$ ,  $P$ ,  $Q$  and  $S$  is done by inspecting the autocorrelation function (ACF) and the partial autocorrelation function (PACF), taking into account differentiation and seasonal differentiation [22].

To create models, the most suitable values of  $p$ ,  $d$  and  $q$  were used, according to the measures of accuracy which are presented in the section of criteria for model selection. The parameters  $\phi$  and  $\theta$  are set through the use of the function minimisation procedures so that the square sum of residues be minimised.

The time series trend is studied applying the Dickey-Fuller [16] and KPSS [17] tests. The Dickey-Fuller test contrasts the null hypothesis that there is a unit root in the autoregressive polynomial (non-stationary series) against the alternative hypothesis that holds the opposite. The KPSS is another test with the same aim, but not exclusive of autoregressive models, supplementary of the former, which contrasts the null hypothesis that the series is stationary around a deterministic trend against the unit root alternative (non-stationary series). Homoscedasticity is studied through the Breusch-Pagan test [18], which contrasts the null hypothesis that holds heteroscedasticity exists against its nonexistence.

## 2.5. Model Selection Criteria

The correlation between the actual and forecast data for a variable ( $x$ ) is expressed by using the correlation coefficient. The coefficient of determination ( $R^2$ ) describes the proportion of total variation in the actual data, which can be explained by the model. The coefficient of determination shows a range of variation [0-1]. If  $R^2 = 1$ , it means a perfect linear fit, that is to say the proportion of total variation in the actual data is explained by the model. Instead if  $R^2 = 0$ , the model does not explain anything of the proportion of total variation in the actual data [23].

Other selection measures applied in R are the standard error of prediction percentage (% SEP) [24], the efficiency coefficient ( $E_2$ ) [25] [26], the average relative variance (ARV) [27] and the Akaike information criterion (AIC) [28]. The first four estimators are unbiased estimators which are used in order to check to what extent the model is able to explain the total variance of the data, while the AIC uses the maximum likelihood function to select the model which best fits data. Moreover, it is advisable to quantify the error in the same units as the studied variable.

These measures, or absolute error measures, include the root mean squared error (RMSE) and the mean absolute error (MAE), both expressed as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (x_t - \hat{x}_t)^2}{N}} \quad (4)$$

$$\text{MAE} = \frac{\sum_{i=1}^N |x_t - \hat{x}_t|}{N} \quad (5)$$

where  $x_t$  is the variable observed at moment  $t$ ,  $\hat{x}_t$  is the estimated variable at the same moment  $t$  and  $N$  is the total number of observations of the validation set.

The standard error of forecast percentage, % SEP, is defined as:

$$\% \text{SEP} = \frac{100}{\bar{x}} \text{RMSE} \quad (6)$$

where  $\bar{x}$  is the average of the variable observed of the validation set. The main advantage of %SEP is its

non-dimensionality, which allows to compare the forecasts of the different models on the same base.

The efficiency coefficient ( $E_2$ ) and the average relative variance (ARV) are used to verify how the model explains the total variance of data and to represent the proportion of the variation of the data observed considered for the model.  $E_2$  and ARV are defined as:

$$E_2 = 1 - \frac{\sum_{i=1}^N (x_t - \hat{x}_t)^2}{\sum_{i=1}^N (x_t - \bar{x}_t)^2}, \quad \text{ARV} = 1 - E_2 \quad (7)$$

The sensitivity to the atypical values due to squaring the terms of the difference is associated to  $E_2$  or to ARV. The Akaike information criterion (AIC) combines the maximum likelihood theory, theoretical information and information entropy [29], and is defined by the following equation [30] [31]:

$$\text{AIC} = \log \left( \frac{(N - k) * \text{MSE}}{N} e^{\frac{2k}{N}} \right) \quad (8)$$

where  $N$  is the total number of observations of the validation set,  $k$  is the number of the parameters of the estimated model, MSE is the mean square error estimated, which is defined by the following equation [30] [31]:

$$\text{MSE} = \sum_{i=1}^N \frac{(x_t - \hat{x}_t)^2}{N - k} \quad (9)$$

where  $N$  is the total number of observations of the validation set,  $k$  is the number of parameters of the estimated model and  $x_t$  is the variable observed at moment  $t$  and  $\hat{x}_t$  is the estimated variable at the same moment  $t$ . The AIC criterion takes into account the changes in the goodness of fit and the differences in the number of parameters between two models [32].

Depending on the fit, a model which explains a high variance level ( $R^2$ , ARV,  $E_2$ ) in the validation period is associated to low absolute error (RMSE, MAE), relative (% SEP) and Akaike (AIC) values. Hence, the hypothesis is validated that when using AIC the best model will be that which presents the lowest value, since its likelihood function will fit the data more accurately [28].

## 2.6. Application

The nature of information differs now from that of information in the past. Due to the vast amount of measuring devices (sensors, microphones, cameras, medical scanners, images, etc.), the data generated by these elements are the largest of the entire available information spectrum. For this reason, the analysis of the wealth of time series has been carried out in a continuous and frequent way [33] in order to obtain the prediction variables and thus to be able to warn behaviour in the environment these occur.

The analyses of time series take into account the degree of dependence between observations and allow to obtain valid inferences without violating basic assumptions of the statistical model or introducing variations in order to overlook this problem; this way, the model further fits the real behaviour of the series.

Since time series are currently employed in different and various fields of knowledge—telecommunications [34], fisheries [35], medicine [36], etc.—it is important to perform a script that allows to give a global and integrated vision on the treatment of time series grouping all the relevant information with the characteristics of the series and prediction models.

Treatment and analysis of time series using free software such as R presents advantages and disadvantages in comparison with private software. On the one hand R has been used in this work as a free and cross-platformer software, making it easy to work with different operating systems. As it has an open source, it is continuously updated by users, not to mention its great graphical power. On the other hand we are aware that the development of this script in the R programming environment presents a number of drawbacks, such as abundant but unstructured help information or packages and functions that make it difficult to locate specific information in a given search. Error messages do not show clearly where in the development of the script the bug is committed, which creates problems for users with little experience in this programming environment making the initiation tedious. R is a programming language in lines of commands, which does not use menus as other statistical pro-

grams (e.g., Statgraphics) interfaces. However this can also be an advantage since R advanced users are able to schedule the treatment and analysis of data, in order to understand the basis of the statistical development and data analysis.

To this aim the ST.R script has been created, whose main objective is the analysis and development of forecasting models for time series. It can be established that time series models allow to estimate the degree of significance of a level change which is operated as a result of the application of a treatment [37]. These models not only allow to obtain statistical inferences on treatment action, but also solve the problem of dependence inherent to this type of designs which use a single subject.

In this work, Excel has been used for the database structure management. We know that this system is not sufficiently solvent to support the current data productions [38]. Although Excel is satisfactory for time series management since this working field is univariate based, Excel has also the advantage of being user friendly and accessible for most users. Then this system is considered an efficient tool when it comes to structuring univariate time series.

### 3. Conclusion

In conclusion, the present script aims to be a useful and efficient tool to give a global and integrated vision on the time series treatment through the application of Data Mining based on ARIMA models. Introducing this script has made it possible to group all the most relevant information related to the series and prediction models characteristics in order to be able to optimise decision-making in research, in the sense of obtaining more robust and reliable results to support the study.

### Acknowledgements

We thank Sonia Páez-Mejías for the edition of the manuscript in English. We also wish to acknowledge Miguel Ángel Rozalén Soriano for the constructive comments and suggestions about Big Data and Data Mining. This study has been submitted to the V International Symposium on Marine Sciences (July, 2016). The authors are grateful to anonymous referees for their helpful comments and CACYTMAR (Centro Andaluz de Ciencia y Tecnología Marinas) for funding support.

### References

- [1] IBM (2015). [www-01.ibm.com/software/data/bigdata/what-is-big-data.html](http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html)
- [2] Einav, L. and Levin, J. (2014) Economics in the Age of Big Data. *Science*, **346**, 715-721. <http://dx.doi.org/10.1126/science.1243089>
- [3] Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014) The Parable of Google Flu: Traps in Big Data Analysis. *Science*, **343**, 1203-1205. <http://dx.doi.org/10.1126/science.1248506>
- [4] Fan, C., Xiao, F., Madsen, H. and Wang, D. (2015) Temporal Knowledge Discovery in Big BAS Data for Building Energy Management. *Energy and Buildings*, **109**, 75-89. <http://dx.doi.org/10.1016/j.enbuild.2015.09.060>
- [5] Vera-Baquero, A., Colomo-Palacios, R. and Molloy, O. (2016) Real-Time Business Activity Monitoring and Analysis of Process Performance on Big-Data Domains. *Telematics and Informatics*, **33**, 793-807. <http://dx.doi.org/10.1016/j.tele.2015.12.005>
- [6] Krishnan, K. (2013) *Data Warehousing in the Age of Big Data*. Newnes, Boston.
- [7] Inmon, W.H. and Linstedt, D. (2015) *Data Architecture: A Primer for the Data Scientist*. Morgan Kaufmann, Boston.
- [8] Rathod, R.R. and Garg, R.D. (2016) Regional Electricity Consumption Analysis for Consumers Using Data Mining Techniques and Consumer Meter Reading Data. *Electrical Power and Energy Systems*, **78**, 368-374. <http://dx.doi.org/10.1016/j.ijepes.2015.11.110>
- [9] Zhang, Z., Kusiak, A., Zeng, Y. and Wei, X. (2016) Modeling and Optimization of a Wastewater Pumping System with Data-Mining Methods. *Applied Energy*, **164**, 303-311. <http://dx.doi.org/10.1016/j.apenergy.2015.11.061>
- [10] Shaheen, M. and Khan, M.Z. (2016) A Method of Data Mining for Selection for Wind Turbines. *Renewable and Sustainable Energy Reviews*, **55**, 1225-1233. <http://dx.doi.org/10.1016/j.rser.2015.04.015>
- [11] Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.
- [12] Batareseh, F.A. and Latif, E.A. (2015) Assessing the Quality of Service Using Big Data Analytics: With Application to Healthcare. *Big Data Research*, **4**, 13-24. <http://dx.doi.org/10.1016/j.bdr.2015.10.001>

- [13] Legates, M.J. (1999) Evaluating the Use of Goodness of Fit Measures in Hydrologic and Hydroclimatic Model Validation. *Water Resources Research*, **35**, 233-241. <http://dx.doi.org/10.1029/1998WR900018>
- [14] Abrahart, R.J. and See, L. (2000) Comparing Neural Network and Autoregressive Moving Average Techniques for the Provision of Continuous River Flow Forecasts in Two Contrasting Catchments. *Hydrological Processes*, **14**, 2157-2172. [http://dx.doi.org/10.1002/1099-1085\(20000815/30\)14:11/12<2157::AID-HYP57>3.0.CO;2-S](http://dx.doi.org/10.1002/1099-1085(20000815/30)14:11/12<2157::AID-HYP57>3.0.CO;2-S)
- [15] R Documentation (2016) ARIMA Modelling of Time Series. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/arima.html>
- [16] Dickey, D.A. and Fuller, W.A. (1979) Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, **74**, 427-431.
- [17] Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shinb, Y. (1992) Testing the Null Hypothesis of Stationary against the Alternative of a Unit Root. *Journal of Econometrics*, **54**, 159-178. [http://dx.doi.org/10.1016/0304-4076\(92\)90104-Y](http://dx.doi.org/10.1016/0304-4076(92)90104-Y)
- [18] Breusch, T.S. and Pagan, A.R. (1979) A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, **47**, 1287-1294. <http://dx.doi.org/10.2307/1911963>
- [19] Ljung, G.M. and Box, G.E.P. (1978) On a Measure of Lack of Fit in Time Series Models. *Biometrika*, **65**, 297-303. <http://dx.doi.org/10.1093/biomet/65.2.297>
- [20] Chatfield, C. (2013) *The Analysis of Time Series: An Introduction*. CRC Press, Boca Raton.
- [21] Parreño, J., De la Fuente, D., Gómez, A. and Fernández, I. (2003) Previsión en el sector turístico en España con las metodologías Box-Jenkins y Redes neuronales. XIII Congreso Nacional ACEDE, Salamanca, España.
- [22] Holton, J. and Keating, B. (1996) *Previsiones en los negocios*. Irwin, Madrid.
- [23] Steel, R.G.D. and Torrie, J.H. (1960) *Principles and Procedures of Statistics with Special Reference to the Biological Sciences*. McGraw Hill, New York, 187-287.
- [24] Ventura, S., Silva, M., Pérez-Bendito, D. and Hervas, C. (1995) Artificial Neural Networks for Estimation of Kinetic Analytical Parameters. *Analytical Chemistry*, **67**, 1521-1525. <http://dx.doi.org/10.1021/ac00105a007>
- [25] Nash, J.E. and Sutcliffe, J.V. (1970) River Flow Forecasting through Conceptual Models Part I-A Discussion of Principles. *Journal of Hydrology*, **10**, 282-290. [http://dx.doi.org/10.1016/0022-1694\(70\)90255-6](http://dx.doi.org/10.1016/0022-1694(70)90255-6)
- [26] Kitanidis, P.K. and Bras, R.L. (1980) Real-Time Forecasting with a Conceptual Hydrologic Model: 2. Applications and Results. *Water Resources Research*, **16**, 1034-1044. <http://dx.doi.org/10.1029/WR016i006p01034>
- [27] Griño, R. (1992) Neural Networks for Univariate Time Series Forecasting and Their Application to Water Demand Prediction. *Neural Network World*, **2**, 437-450.
- [28] Akaike, H. (1974) A New Look at the Statistical Identification Model. *IEEE Transactions on Automatic Control*, **19**, 716-723. <http://dx.doi.org/10.1109/TAC.1974.1100705>
- [29] Motulsky, H.J. and Christopoulos, A. (2003) *Fitting Models to Biological Data Using Linear and Nonlinear Regression*. GraphPad Software Inc., San Diego, 351 p.
- [30] Diebold, F. (1999) *Elementos de Pronósticos*. International Thomson Editores, México, 106-128 p.
- [31] Giraldo Gómez, N.D. (2006) *Series de Tiempo con R*. Universidad Nacional de Colombia, Colombia.
- [32] Gaona, B. (2005) *Matrices de covarianza estructuradas en modelos con medidas repericas*. Tesis de maestría, Mayagüez, Puerto Rico.
- [33] Guyet, T. and Nicolas, H. (2016) Long Term Analysis of Time Series of Satellite Images. *Pattern Recognition Letters*, **70**, 17-23. <http://dx.doi.org/10.1016/j.patrec.2015.11.005>
- [34] Siluyele, I. and Jere, S. (2016) Using Box-Jenkins Models to Forecast Mobile Cellular Subscription. *Open Journal of Statistics*, **6**, 303-309. <http://dx.doi.org/10.4236/ojs.2016.62026>
- [35] Czerwinski, I.A., Gutiérrez-Estrada, J.C. and Hernando-Casal, J.A. (2007) Short-Term Forecasting of Halibut CPUE: Linear and Non-Linear Univariate Approaches. *Fisheries Research*, **86**, 120-128. <http://dx.doi.org/10.1016/j.fishres.2007.05.006>
- [36] Jere, S. and Moyo, E. (2016) Modelling Epidemiological Data Using Box-Jenkins Procedure. *Open Journal of Statistics*, **6**, 295-302. <http://dx.doi.org/10.4236/ojs.2016.62025>
- [37] Arnau, J. (1981) Uso de los modelos de series temporales como técnica de análisis de los diseños conductuales. *Anuario de psicología*, **25**, 20-34.
- [38] Maté Jiménez, C. (2014) Big data. Un nuevo paradigma de análisis de datos. *Anales de mecánica y electricidad*, 10-16.



**Submit or recommend next manuscript to SCIRP and we will provide best service for you:**

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.

A wide selection of journals (inclusive of 9 subjects, more than 200 journals)

Providing 24-hour high-quality service

User-friendly online submission system

Fair and swift peer-review system

Efficient typesetting and proofreading procedure

Display of the result of downloads and visits, as well as the number of cited articles

Maximum dissemination of your research work

Submit your manuscript at: <http://papersubmission.scirp.org/>