AUTOMATIC IDENTIFICATION OF ARIMA MODELS: A CASE STUDY IN THE SUPPLY CHAIN

Fernando Turrado García[1,2], Luis Javier García Villalba[1]

[1]Group of Analysis, Security and Systems (GASS)
Department of Software Engineering and Artificial Intelligence (DISIA)
School of Computer Science, Office 431, Universidad Complutense de Madrid (UCM)
Calle Profesor José García Santesmases s/n, Ciudad Universitaria, 28040 Madrid, Spain
Email: {fturrado, javiergv}@fdi.ucm.es

[2] Decide Soluciones
Calle Albasanz 76 Bajo
Madrid 28037, Spain
Email: fturrado@decidesoluciones.es
URL: http://www.decidesoluciones.es/

Abstract

In the management of the supply chain, one of the main problems is to be able to forecast customer behaviour effectively. At present, there are many models based on statistical or artificial intelligence techniques to generate forecasts for a time series (in the present case the customer purchases), but the identification of the model and its setting is still a manual labour. Due to the nature of this sector and the volume of data involved, manual processing of these time series and their models is not viable. This paper focusses on making an analysis of the difficulties encountered when setting ARIMA models automatically using support vector machines. It also proposes a procedure for the construction of a core set of training data.

Keywords - ARIMA, automatic model fitting, support vector machines, supply chain.

1 INTRODUCTION

In this retail sector domain, the analysis of time series composed by daily sales is a key factor for large corporations. Having a mechanism to accurately anticipate demand allows companies to define optimal strategies for purchasing merchandise from suppliers and minimize inventory costs ... But this task, the time series analysis, is a complex one that requires highly specialized profiles for it to be carried out. Consequently, finding an automatic way to perform it also reduces operational costs.

From another point of view, accurate forecasts also allow the evaluation of the commercial activities performance (promotions, events ...). This can be made by comparing the forecasted value with the real one. This allows an objective evaluation of the business performance of different promotions, campaigns, advertisements etc.

To the inherent difficulty in the study of the series, there is another complexity caused by the volume of data being handled. In [1]a study is shown of the same problem located at an international distributor. In that paper, each store (hypermarkets) offers consumers more than 10,000 different SKUs. Given that the company has 200 such centers, there are over 2 million sales time series to be analyzed.

2 PROBLEM DISCUSSION

Given a massive set of time series, the problem which needs to be solved is to define a method / automatic procedure that assignsan estimator to each element (a time series) in order to make forecasts with it. These estimators should be as accurate as possible.

In [2] a method is proposed which combines statistical techniques (ARIMA models for the calculation of forecasts) with artificial intelligence methods (support vector machines) for the classification and forecasting of the series. In that case, the resolution of the problem is made in two phases: a first one represented by the classification of the time series and a second in which statistical models are assigned to them.

The first phase focuses on the classification of the time series in groups that share the same statistical model. The main objective pursued in this stage is to reduce the number of sets needed to be analyzed when looking for statistical models. This classification is based on the degree of similarity of the results of the autocorrelation functions and partial autocorrelation (ACF and PACF) on each series. That is, if two series have similar values ACF and PACF then the same ARIMA model can be assigned to them. For classification using support vector machines (SVM), which are trained with self-generated dummy data.The results of this phase show the effectiveness of SVM as classifiers, the 98% of the elements taken from the original set were grouped into 12 categories.

However, due to the absence of training data that relate an ARIMA model with a category, the second phase is indicated to perform a manual task. The aim of this paper is to present the difficulties encountered when trying to generate a set of training data that enables the association of an ARIMA model for each category. It also shows a simple algorithm for generating a basic dataset for training the SVMs needed to identify ARIMA models.

2.1 Simulation of ARIMA models with R statistical software

The statistical software R [3] provides a utility function to generate time series given an ARIMA model. The following parameters must be provided in order to define the ARIMA model (only the required ones are detailed):

• The order P of the auto-regressive (AR) and its coefficient set.

• The Q of the order based on moving averages (MA) and its coefficient set.

• The order of differentiation D.

• The number of samples to generate.

After defining the parameter values (which identify the ARIMA model), pseudo-random data can be generated to train and test the SVMs. However, as shown at 2.2 section, for given values of P, D and Q, the choice of coefficients significantly alter the results of ACF and PACF functions applied to the generated series.

2.2 Generating coefficients

Considering the above, in order to build an artificial time series based on an ARIMA (P, D, Q), the algorithm will create a tuple of P elements to the auto-regressive and a tuple of Q items for the moving average part of the model.

A basic approach for the construction of such tuples can be to create combinations of elements from a given set of coefficient values.

However, a necessary condition for the series to be stationary is not having a unit root. Given a time series, check this condition is a complex task that in some cases involves procedures carried out heuristically, as can be seen in [4] and [5]. However, in this case, in which the coefficients are provided, that condition can be tested by calculating the roots of the complex polynomial associated (with another utility function present in R).

In our case, the basic set of parameters for the auto-regressive part is set as {-1, -0.9, -0.8, ... -0.1,0,0.1, ... 0.8,0.9,1} and {-2 - 1,0,1,2} for the moving averages part. Once applied the root unit filter, the following combinations for each model exist.

Table 1. Number of parameter combinations to study in each case.

P	D	Q	AR Comb.	MA Comb.	Total
0	0	0	1	1	1
0	0	1	1	5	5
0	0	2	1	25	25
0	0	3	1	125	125
0	0	4	1	625	625
1	0	0	19	1	19
1	0	1	19	5	95
1	0	2	19	25	475
1	0	3	19	125	2375
1	0	4	19	625	11875
2	0	0	291	1	291
2	0	1	291	5	1455
2	0	2	291	25	7275
2	0	3	291	125	36375
2	0	4	291	625	181875
...					

To demonstrate the variance introduced by the coefficients, the results of applying ACF & PACF functions to two basic ARIMA models can be found downwards. For each case, 1000 different time series were generated using the R function mentioned above. The table shows the following values: minimum, maximum and average results for ACF and PACF functions for each of the coordinates.

A. ARIMA(1,0,0)

Table 2. Results of ACF & PACF functions for coeff. Values -0.9,0.9 and 0.3

		P=-0,9			P=0,9			P=0,3		
	Lag	Min.	Max.	Mean	Min.	Max.	Mean	Min.	Max.	Mean
	1	1	1	1	1	1	1	1	1	1
ACF	2	-0,95	-0,74	-0,87	0,65	0,94	0,84	0,01	0,51	0,29
	3	0,53	0,91	0,75	0,44	0,90	0,71	-0,15	0,25	0,09
	4	-0,86	-0,40	-0,66	0,27	0,85	0,59	-0,17	0,27	0,03
	5	0,22	0,82	0,57	0,13	0,80	0,49	-0,21	0,28	-0,01
	6	-0,79	-0,04	-0,49	0,04	0,76	0,40	-0,28	0,22	-0,04
	7	-0,10	0,76	0,43	-0,07	0,72	0,34	-0,29	0,22	-0,02
	8	-0,75	0,18	-0,37	-0,18	0,68	0,28	-0,27	0,19	-0,01
	9	-0,17	0,72	0,32	-0,23	0,63	0,23	-0,20	0,22	0,01
	10	-0,71	0,22	-0,27	-0,25	0,58	0,19	-0,21	0,16	0,00
	11	-0,29	0,70	0,23	-0,27	0,56	0,14	-0,19	0,22	-0,02
PACF	12	-0,95	-0,74	-0,87	0,65	0,94	0,84	0,01	0,51	0,29
	13	-0,34	0,16	-0,03	-0,33	0,17	-0,03	-0,20	0,21	0,00
	14	-0,30	0,20	-0,02	-0,39	0,21	-0,02	-0,18	0,17	0,00
	15	-0,23	0,11	-0,03	-0,22	0,16	-0,01	-0,20	0,13	-0,03
	16	-0,26	0,18	-0,03	-0,24	0,15	-0,01	-0,22	0,10	-0,03
	17	-0,22	0,23	-0,03	-0,26	0,20	0,00	-0,25	0,23	0,00
	18	-0,27	0,18	-0,01	-0,21	0,16	-0,01	-0,22	0,19	0,00
	19	-0,26	0,15	-0,04	-0,27	0,25	-0,03	-0,26	0,17	0,00
	20	-0,18	0,17	0,00	-0,17	0,23	-0,01	-0,16	0,24	-0,01
	21	-0,23	0,17	-0,02	-0,22	0,13	-0,03	-0,21	0,16	-0,03

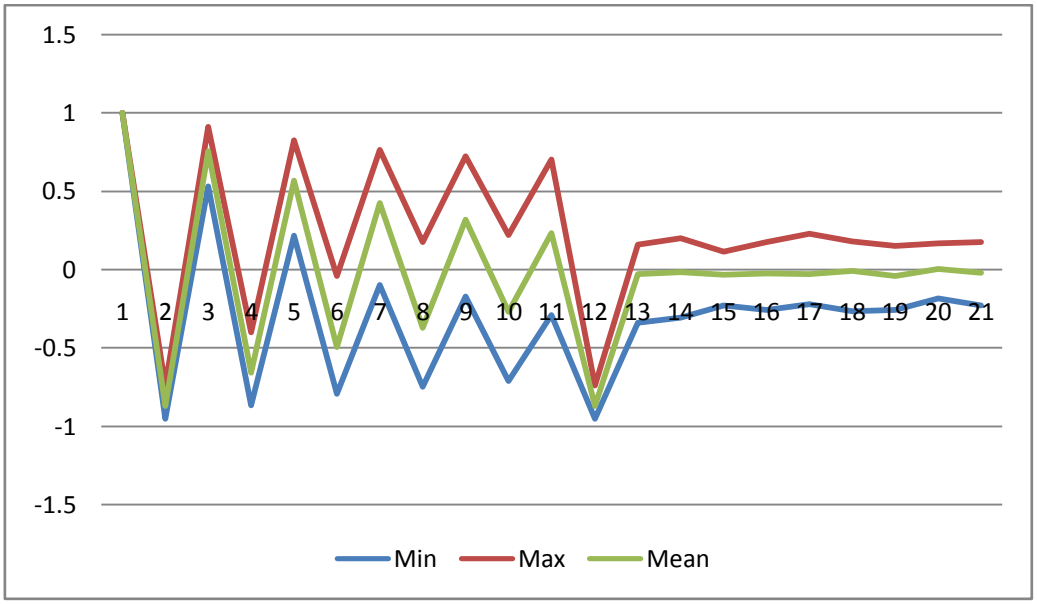

Figure 1. Graphic representation of ACF & PACF values where coefficient value is -0.9

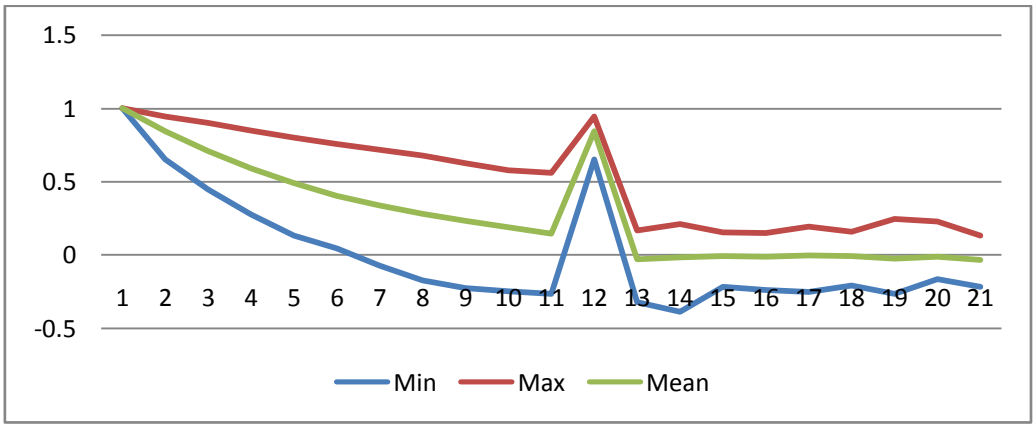

Figure 2. Graphic representation of ACF & PACF values where coefficient value is 0.9

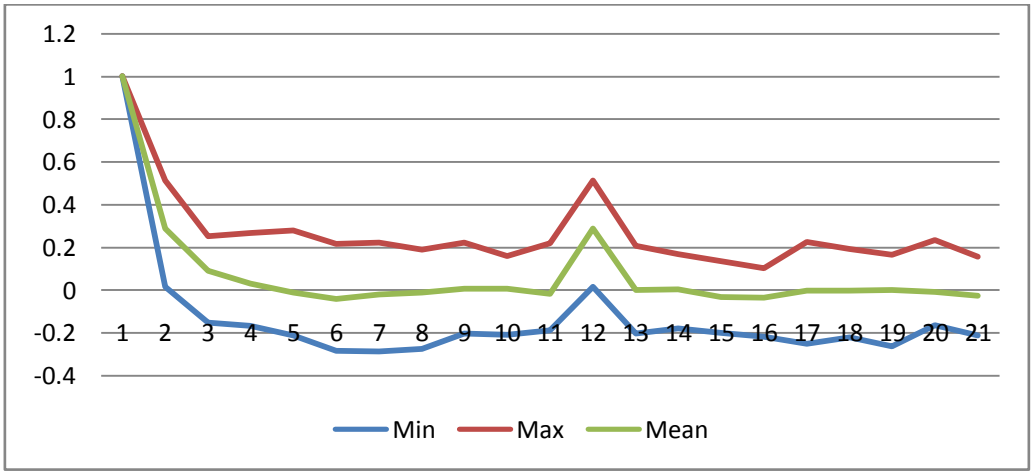

Figure 3. Graphic representation of ACF & PACF values where coefficient value is 0.3

B. *ARIMA(0,0,1)*

Table 3. Results of ACF & PACF functions for coeff. Values -2, 1 and 2

		P=-2			P=1			P=2		
	Lag	Min.	Max.	Mean	Min.	Max.	Mean	Min.	Max.	Mean
	1	1	1	1	1	1	1	1	1	1
ACF	2	-0,54	-0,25	-0,40	0,24	0,60	0,46	0,18	0,54	0,38
	3	-0,19	0,24	0,01	-0,26	0,23	-0,04	-0,26	0,24	-0,02
	4	-0,25	0,33	0,02	-0,30	0,23	-0,01	-0,19	0,17	-0,02
	5	-0,28	0,20	-0,04	-0,33	0,19	-0,03	-0,25	0,18	-0,02
	6	-0,25	0,27	0,01	-0,22	0,17	-0,04	-0,32	0,20	-0,03
	7	-0,23	0,24	0,00	-0,25	0,14	-0,04	-0,23	0,25	-0,02
	8	-0,25	0,31	-0,03	-0,33	0,21	-0,02	-0,24	0,19	-0,02
	9	-0,23	0,34	0,01	-0,26	0,17	-0,02	-0,30	0,26	-0,01
	10	-0,30	0,24	-0,01	-0,24	0,14	-0,03	-0,26	0,17	-0,03
	11	-0,23	0,21	-0,01	-0,30	0,18	-0,03	-0,32	0,22	-0,05
PACF	12	-0,54	-0,25	-0,40	0,24	0,60	0,46	0,18	0,54	0,38
	13	-0,36	0,03	-0,18	-0,48	-0,04	-0,32	-0,37	0,02	-0,20
	14	-0,24	0,17	-0,06	-0,07	0,42	0,22	-0,06	0,33	0,08
	15	-0,30	0,19	-0,08	-0,39	-0,01	-0,21	-0,32	0,12	-0,07
	16	-0,20	0,17	-0,04	-0,04	0,33	0,11	-0,17	0,19	0,01
	17	-0,33	0,11	-0,04	-0,35	0,05	-0,14	-0,13	0,22	-0,02
	18	-0,20	0,09	-0,05	-0,15	0,27	0,09	-0,29	0,20	-0,01
	19	-0,24	0,25	-0,04	-0,34	0,05	-0,12	-0,21	0,16	-0,01
	20	-0,29	0,18	-0,03	-0,17	0,24	0,08	-0,24	0,16	-0,03
	21	-0,23	0,18	-0,04	-0,25	0,06	-0,10	-0,20	0,11	-0,04

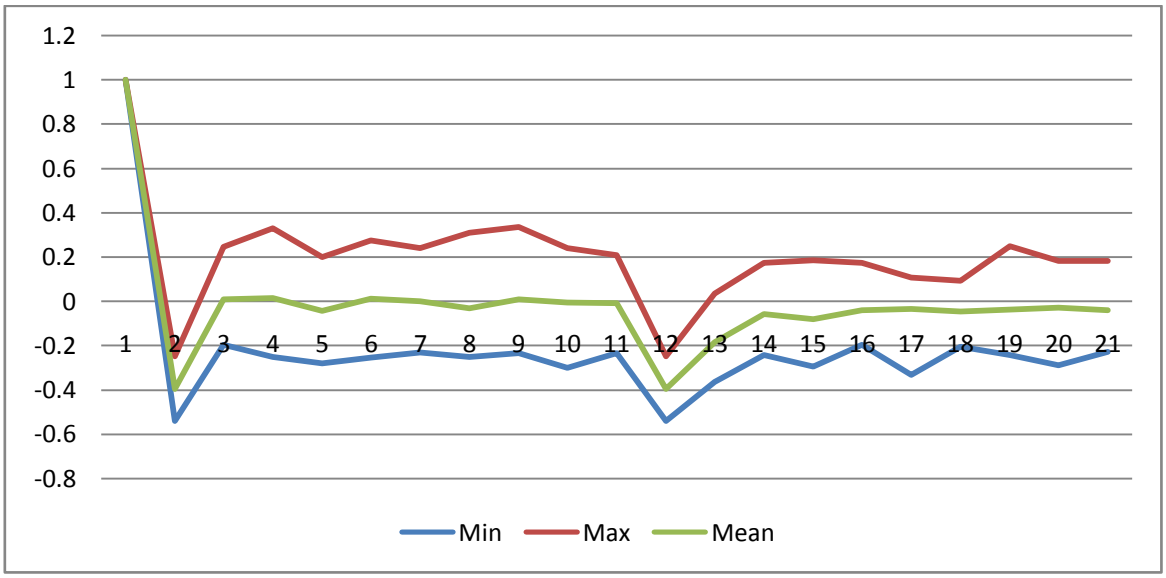

Figure 4. Graphic representation of ACF & PACF values where coefficient value is -2

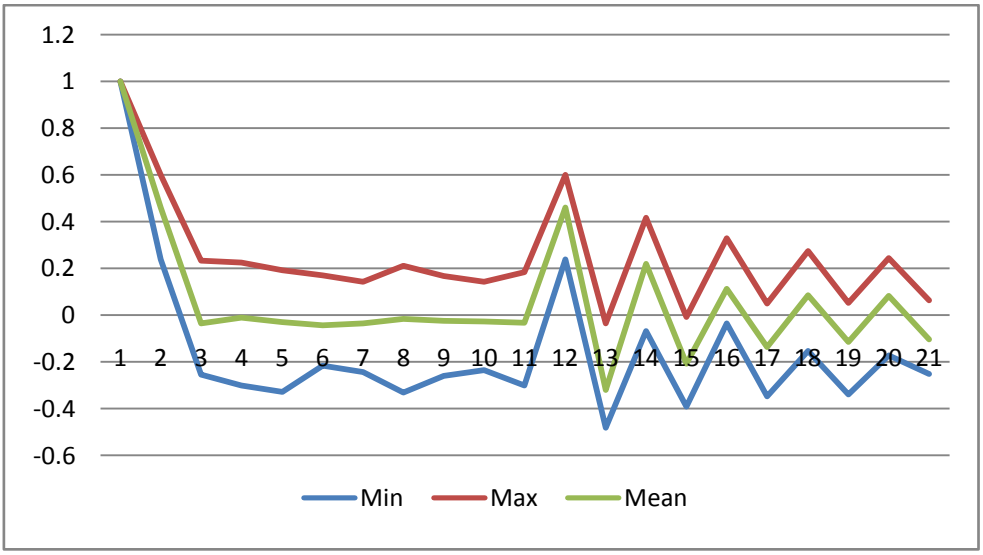

Figure 5. Graphic reprentation of ACF & PACF values where coefficient value is 1

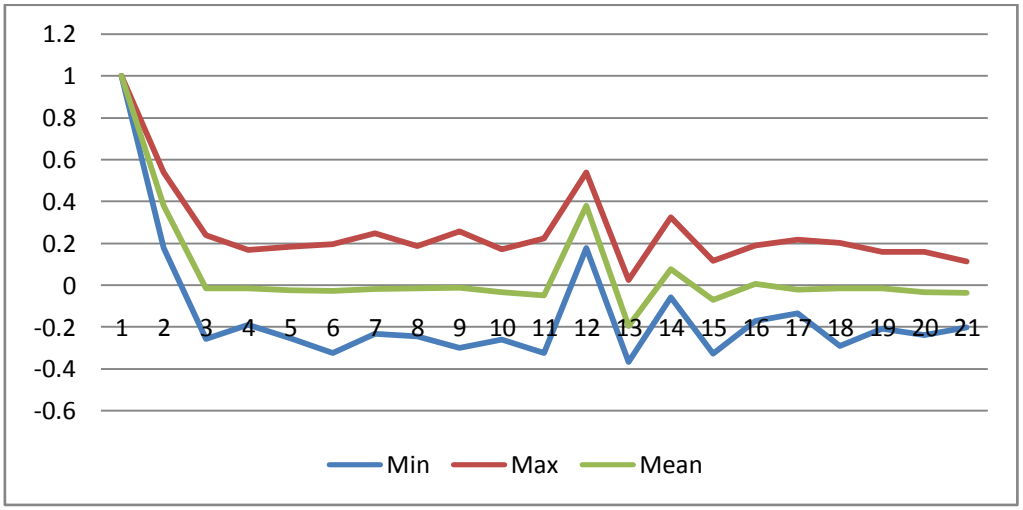

Figure 6. Graphic reprentation of ACF & PACF values where coefficient value is 2

2.3 Extension to seasonal ARIMA models

Construction of data sets for seasonal ARIMA models can be done in an analogous manner, as these models are defined as combination of two non-seasonal ARIMA models plus an additional parameter, the cycle length. Thus, a model of this type has the following parameters:

• P, p: Order of autoregressive terms (regular and seasonal, respectively).

• D, d: Differentiation order (regular and seasonal, respectively).

• Q, q: Order of the terms based on moving averages (regular and seasonal, respectively).

• s: cycle length

This time, the algorithm has to create combinations of P + Q + p + q +1 elements instead of combinations of P + Q elements needed in the non-seasonal models.

2.4 Training of Support Vector Machines

Due to the nature of ACF & PACF functions, all values passed to the SVM belong to the [-1,1] interval, so we can represent each time series as a vector of 2N coordinates (N for the ACF values and another N for the PACF values).

In the construction of the training data sets, supposing a use case where the support vector machines are used as binary classifiers, two data groups are created. The first group contains the data related to the time series defined by the ARIMA model and its coefficients, let it be called A. The second group is defined as the complementary of A in [-1,1]^2N.

At this point, it only remains to describe a method for generating the complementary group. To generate these data, we propose a simple method:Generate random vectors whose coordinates are outside the "corridor" defined by the maximum and minimum for each coordinate in A. That is, x (i) is contained in [-1, min y(i) {y in A }) U (max y(i) {y in A }, 1].

3 CONCLUSIONS AND FUTURE WORK

After analyzing the data generated for small models, there are a couple of aspects to be considered: the first one is that the sign of the coefficients changes the type of values ACF / PACF (as can be seen clearly in the case ARIMA (1,0,0)). The second is that the magnitude of the coefficients also modifies the result of these functions but to a lesser extent.

The presence of these two facts suggests that it will be necessary to train a SVM for each of the possible combinations. But as Table 1 shows, the number of combinations grows exponentially as the order of the model does.

To mitigate this growth, a data transformation will be needed that reduces or unify the differences caused by changes in the coefficients. What does not seem avoidable is having to generate a support vector machine for each combination of signs of the coefficients. That is, in the case ARIMA (2,0,0) would have 4: + +, + -, - + and - -.

To summarize, the proposed method allows a data set can be generated which allows the support vector machines to identify ARIMA models (seasonal or not). But it is an inefficient process since as the model complexity increases, the amount of resources necessary to recognize it grows exponentially.

ACKNOWLEDGMENTS

This work was supported by the Agencia Española de Cooperación Internacional para el Desarrollo (AECID, Spain) through Acción Integrada MAEC-AECID MEDITERRÁNEO A1/037528/11.

REFERENCES

[1] F. Turrado García, L. J. García Villalba, J. Portela.Intelligent System for Time Series Classification Using Support Vector Machines Applied to Supply-Chain, *Expert Systems with Applications*, Vol. 39, No. 12, pp. 10590-10599,September 2012.

[2] Box, George and Jenkins, Gwilym. Time Series Analysis: Forecasting and Control, San Francisco,*Holden-Day*,1970

[3] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2008.

[4] P. C.Phillips,and P.Perron. Testing for a Unit Root in Time Series Regression.*Biometrika*,Vol. *75*, No. 2, pp. 335-346.

[5] S. E. Said, and D. A. Dickey. Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order.*Biometrika*, Vol. 71, No. 3,pp. 599-607, 1984.