Use of ARIMA models for predicting milk production. Case study in UBPC ¨Maniabo¨, Las Tunas

Lyhen Sánchez[1], Gladis Cabanas[2],Yoandra Abad[3] and Verena Torres[1]
[1] *Instituto de Ciencia Animal, Apartado Postal 24, San José de las Lajas, Mayabeque, Cuba*
[2] *Facultad de Matemática y Computación, Universidad de la Habana, Ciudad de La Habana, Cuba*
[3] *Instituto de Medicina Tropical "Pedro Kourí", La Habana, Cuba*
Email: lyhen@ipk.sld.cu

This study was carried out to predict milk production, in a short term, by using ARIMA models, in a case study performed at the Basic Unit of Cooperative Production (UBPC, initials in Spanish) "Maniabo", Las Tunas, Cuba. For that purpose, data of monthly milk production during the period of 2000-2010 were used. Charts of horizontal plotting and Box–Plot were used for the descriptive analysis of the series and for detecting the tendency to polynomial softening. Using periodgram, the marked peaks in the periodical movements of the series and the significance of a frequency were confirmed, and seasonal nature was detected, which was described with seasonal indexes. Correlations with the differentiated series were determined and parameters p and q of model were estimated. The best fitted model was ARIMA (1, 0, 3) x (0, 1, 0) 12 with constant. Milk production for 2011 was predicted and validated.

Key words: *time series, ARIMA models, milk production*

The statistic research, applied to the agricultural and livestock field, deals with problems where the conditions for using a general linear model are not fulfilled, because controls of the variable are carried out historically and, the observations of the same variable are generally correlated, besides, the prediction is a type of extrapolation (Aguirre, 1994). With the application of time series models, the performance of a phenomenon can be predicted and described, which changes in time and shows dependence among the consecutive observations (Arellano, 2001).

According to Maté (1995), Cuban milk production has undergone variations for years which have brought about negative implications for the sector growth. Most of models for time series assume that factors provoking changes in the series are constant. Therefore, one of the first analyses is to isolate these factors for analyzing its performance (Blumberg, 1984).

The ARIMA (Autoregressive Integrated Moving Average) modeling, among the time series, was developed by George Box and Gwilym Jenkins during the 70´s from last century, and constituted a revolution in the analysis of time series.

It is known that since 2000, the UBPC "Maniabo" have not undergone important changes in milk production, therefore, the series produced liters from this year can be taken as objective of this study.

There are few studies in Cuba that use time series, particularly using ARIMA models in the agricultural and livestock field. That is why, this study tried to predict, in a short term, milk production of the UBPC "Maniabo" from Las Tunas using this type of model.

Materials and Methods

The study was carried out in UBPC "Maniabo", Las Tunas. Data used for the analysis was taken from the monthly milk production from the UBPC during the period of 2000-2010.

Charts of horizontal plotting and Box –Plot were used for the descriptive analysis of the series, and the polynomial softening for detecting its tendency (Coutín, 2003).

The periodgram (integrated and simple periodgram) were used to detect the significance of the seasonal component, and the seasonal indexes were found to study the performance of seasonal nature in a descriptive way (Maté, 2011).

Together with the series of seasonal difference, there are coefficients of partial and simple correlation, through which the parameters p, d and q, and P, D and Q from the ARIMA model to be fitted were estimated, through the notation ARIMA (p,d,q) (P,D,Q) k, where k is length of seasonal period.

The fitness goodness of the tested model was carried out through the statistical criteria of classical errors: Mean Error (ME), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE) and Square Mean Error (SME) (Torres *et al.*, 2012).

The selected model was validated through the verification of the randomness of residues with the test of sequences, and the residual normality of it (Maté, 1995).

The results were processed with statistical packages, R version 2.15.0, STATGRAPHICS Centurión XV.II and Microsoft Excel 2007.

Results and Discussion

According to criteria of Einsteinnet (2001), the series analyzed was considered as consistent, stable and did not presented abnormal values.

The initial descriptive analysis of the series (figure 1) shows remarkable seasonal nature, in which the highest yield in milk production appears between May and October. November, in spite of not being included in the rainy period for the country, achieved good milk productions because there is still humidity in the soils, which is a great help for growing pastures. The mean production of this month is very close to the general mean.

The highest variability appears in the months of higher mean, but not so much to make this visible in the coefficients of variation. In the case of rainy season, variability is higher during the central months of humid season, and due to the relation of rain with pastures, and theirs with milk production, variation of milk production, evidenced in these months, could be related to rain. These results coincide with those of a study carried out by Grubb, (1992).

Figure 2, which shows the temporal series of milk production during the analyzed period (2000-2010), suggests a decreasing tendency in the last years, which is not very marked or sustained for a considerable period of time. Therefore, a type of smoothening is applied to visualize better the movement of the series, and check the existence of tendency or not. The polynomial softening was applied for this purpose (figure 3), for n=1 and n=3, and comparison of fitness. The criterion that stated that the series did not present tendency was confirmed with the polynomial softening.

According to Colás *et al.* (2011), periodgram are the elements that validate the significance of the seasonal component and the existence of other periodical phenomena in the series performance (cyclic component).

In the periodgram, the seasonal movement, every 12 months (frequency 0.0833), is highly significant. It presents a peak in this frequency (figure 4).

The integrated periodgram of figure 5 shows that the segment representing the increase in the same frequency is longer than the wide of the most external stripe, which marks a 95% of confidence and the integrated periodgram do not detect other cycles as significant.

Taking into account the previous analysis, which detected the presence of seasonal nature and absence of tendency, the series can be modeled with an ARIMA model with a seasonal difference and without a non-seasonal difference. This coincides with studies carried out by Funes *et al.* (2009).

In the functions of autocorrelation and partial autocorrelation of the differentiated series, present in figure 6, the model orders were estimated to be $p \leq 1$ and

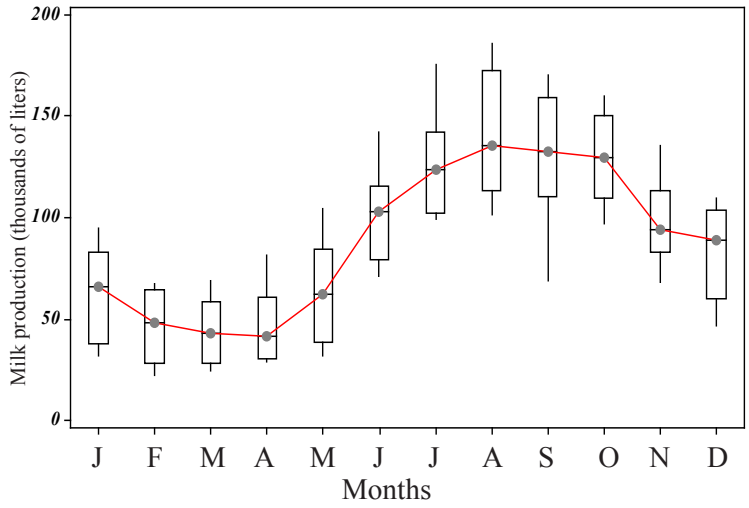

Figure 1. Descriptive analysis of the series of milk production per month.

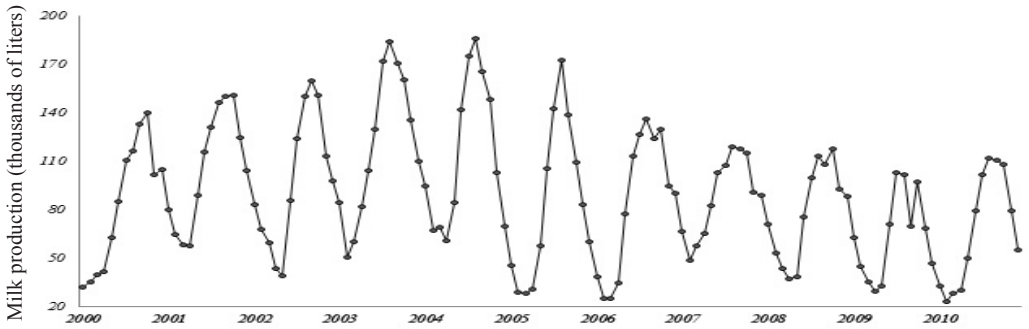

Figure 2. Performance of milk production in the UBPC "Maniabo" from 2000 to 2010.

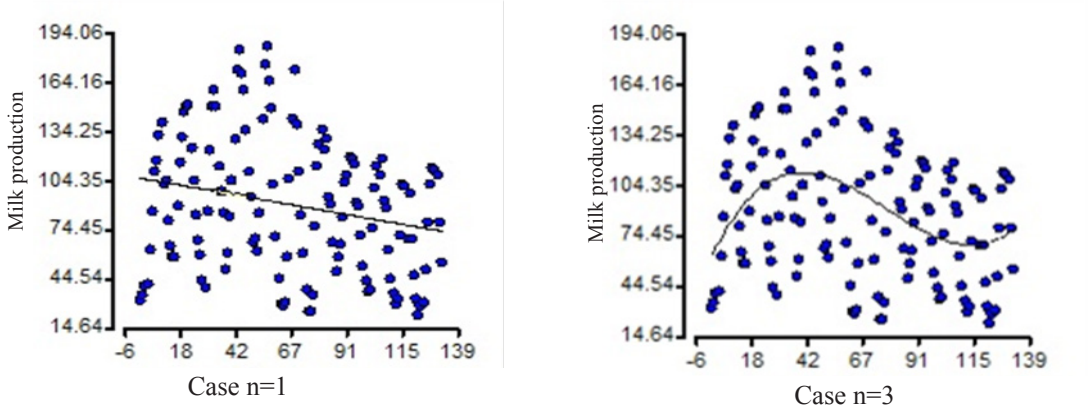

Figure 3. Tendency adjustment of production milk tendency through polynomial softening for n=1 and n=3.

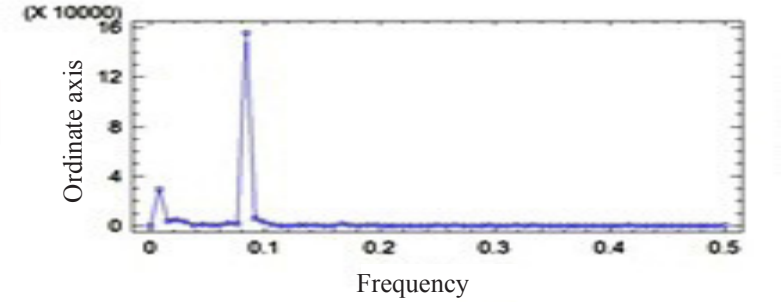

Figure 4. Periodgram of milk production series in the UBPC "Maniabo" from 2000 to 2010.

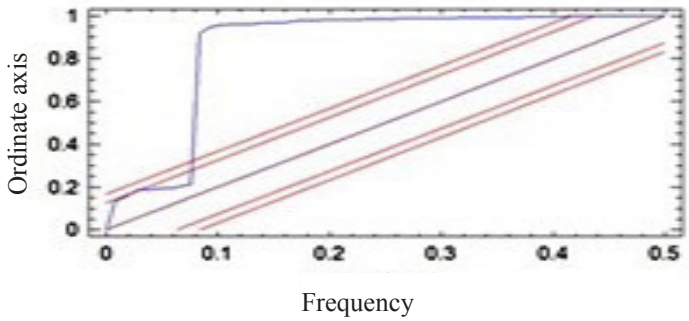

Figure 5. Integrated periodgram of milk production series in the UBPC "Maniabo" from 2000 to 2010.

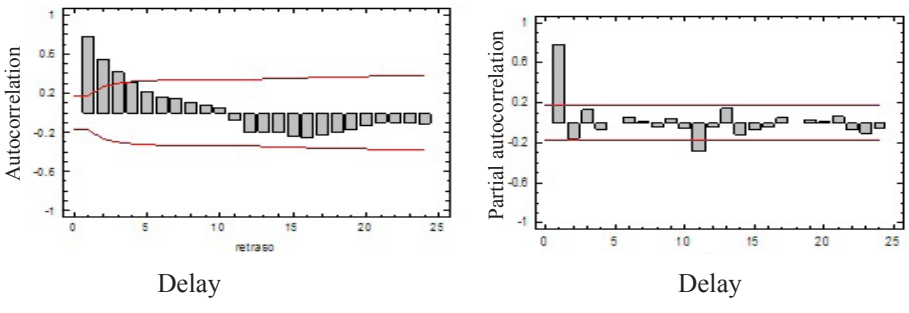

Figure 6. Functions of autocorrelation and partial autocorrelation for the differentiated series of milk production.

$q \leq 3$, and based on that fact, different ARIMA models are tested.

The following models were tested: (1) ARIMA(1,0,0) x(0,1,0)12, (2) ARIMA(1,0,1)x(0,1,0)12, (3) ARIMA (1,0,2)x(0,1,0)12, (4) ARIMA(1,0,3)x(0,1,0)12 and (5) ARIMA(0,0,3)x(0,1,0)12, all of them with constant.

The comparison of tested models was performed using the following table of errors (table 1).

As table 1 shows, model 4 is low in most of the evaluated parameters, that is why it is considered as the best fitted model. Regarding the bias analysis of the model, it tends to have values lower than the real ones, mainly in the peaks of maximum production, which is not considered as important according to the characteristics of Cuban milk production and the low different between prediction and actual values.

Table 2 shows the parameters of the chosen model, P values of the terms MA(1), MA(2) and MA(3) are statistically significant at 95% (lower than 0.05), the value of AR(1) at 90% and the estimated typical deviation of the white noise (error) equals 13.88.

In order to validate the model, tests of randomness (test of sequences) and normality (modified Shapiro-Wilks) of residues (figure 7) were carried out, which demonstrate that residues of milk production series, of the chosen model, are normal and uncorrelated. Besides, there are no changes of mean or variance maintained in time in the chart. Similar results were obtained by Castañeda *et al.* (2006) in a study on the productive performance of European crossbred cattle with Zebu in the tropical weather of Baja California, México. Colás *et al.* 2011 also found similar results in an epidemiological evaluation of bacterial respiratory processes of laying chickens in Cuba.

Then, the milk production for 2011 was predicted using the chosen model, with a trust rate for the prediction at 95% of the trust level (figure 8).

Table 3 shows actual and predicted results of milk

Table 1. Comparison of tested models for milk production series.

Model	Root of Mean Square Error (RMSE)	Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)	Mean Error (ME)	Mean Percentage Error (MPE)
(1)	22.42	17.56	25.56	-0.00	-6.14
(2)	15.51	11.68	16.66	-0.03	-3.46
(3)	14.08	10.79	14.84	-0.08	-2.64
(4)	13.82	10.55	14.12	-0.14	-2.33
(5)	19.42	15.02	22.47	-0.00	-3.64

Table 2. Parameters of the model ARIMA(1,0,3)x(0,1,0)12 with constant.

Parameters	Estimate	Standard Error	t	P-Value
AR(1)	0.00	0.00	-1.73	0.09
MA(1)	-0.99	0.09	-10.99	0.00
MA(2)	-0.53	0.12	-4.48	0.00
MA(3)	-0.20	0.09	-2.14	0.03
Media	-1.59	3.44	-0.46	0.65
Constante	-1.59			

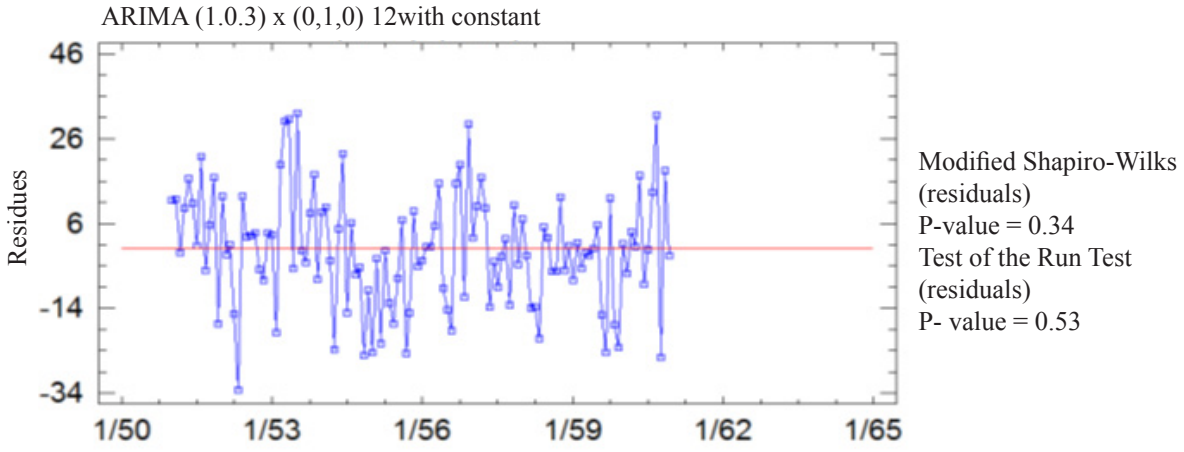

Figure 7. Analysis of supposed fulfillment in residuals from fitted ARIMA model

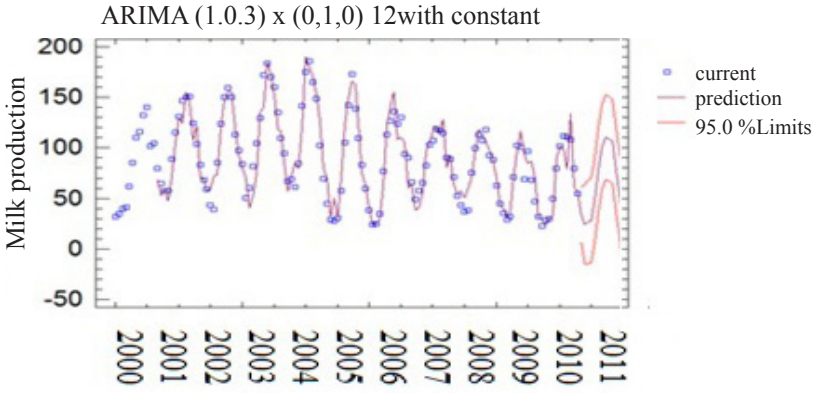

Figure 8. Prediction of milk production for 2011-2013.

Table 3. Actual and predicted values of milk production for 2011.

Year 2011	Actual	Prediction
January	36530	37208
February	25500	25678
March	29650	28225
April	30340	31370
May	49870	48974
June	79170	85769
July	101190	110039
August	111090	118844
September	106960	110453
October	106960	110468
November	78390	77282
December	53940	60116

production for 2011, which demonstrate that there are no significant variations among the prediction of 2011

Milk production in the UBPC had a marked seasonal nature performance, related to humid and dry periods of the year. This characteristic is repeated in any other UBPC of the country, due to the features of the region and the weather.

The prediction values of milk production in this UBPC for 2011 were obtained using ARIMA models, with absolute percentage errors under 15%, which is considered as a good approximation.

Their comparison to actual values for 2011, obtained later, confirmed the goodness of prediction and the possibility of predicting the series performance for next years.

It is demonstrated that, with the application of ARIMA modeling, the prediction of milk production in UBPCs can be achieved, and it helps in the process of decision-making in the agricultural and livestock field

References

Aguirre, A.J. 1994. Introducción al tratamiento de series temporales. Aplicación a las Ciencias de la Salud. . Departamento de Epidemiologia y Bioestadística. Escuela Nacional de Sanidad. Editorial. Díaz de Santos, Madrid. España. 585 p.

Arellano, M. 2001 Introducción al Análisis Clásico de Series de Tiempo. Available: http://ciberconta.unizar.es/LECCION/seriest/100.HTM [Consulted: January 12th, 2012]

Blumberg, J.C.1984. Comments on a simplified time-series analysis for evaluating treatment intervention. Applied Behavior Analysis. 17: 539-542.

Box, G. E. P. & Jenkins, G. M.1970. Time Series Analysis, Forecasting and Control. Editorial. Holden-Day, San Francisco.

Castañeda, M. O. G., Ávila, D. A. & Vinay, V. J. C. 2006. Comportamiento productivo de ganado cruzado europeo por cebú en clima tropical. XVI Reunión Internacional Sobre Producción de Carne y Leche en Clima Cálido. p: 195-199. Baja California, México.

Colás, M., Chávez, M. & Lamazares, D., 2011. Evaluación Epidemiológica de procesos respiratorios bacterianos en gallinas ponedoras. Revista de Salud Animal. 33: 69-75.

Coutín, Marie, Gisele. 2003. Las series temporales. Degree Thesis. Facultad de Contabilidad y Finanzas. Universidad de La Habana. 133 pp

Einsteinnet, 2001. Análisis Clásico De Series Temporales. Available: http://www.einsteinnet.com/econometria/seriestemp/acseriestemp. [Consulted: June 25th, 2010].

Funes, M., Lantiga, M. Van Keulen, H. 2009. AgroEcological Indicators (AEIs) for dairy and mixed farming systems

classification: Identifying alternatives for the Cuban livestock sector. Sustainable Agriculture. 33: 435

Grubb, H. 1992. Multivariate Time Series Analysis of Some Flour Price Data. Applied Statistics, 41: 95–107.

Maté, C.1995. Curso General Sobre STATGRAPHICS. Procedimientos. Métodos Estadísticos. Aplicaciones. Ejercicios Resueltos. Tomo II. Editorial Universidad Pontificia Comillas. Madrid, España.

Maté, C. 2011. A multivariate analysis approach to forecasts combination. Application to Foreign Exchange (FX) markets. Revista Colombiana de Estadística. 34: 347

Torres, V., Barbosa I., Meyer R. Noda A. & Sarduy L. 2012. Criteria of goodness of fit test in the selection of non-linear models for the description of biological performances. Cuban J. Agric. Sci. 46: 345

Received: February 15, 2013