Time series analysis of cow milk production at Andassa dairy farm, West Gojam Zone, Amhara Region, Ethiopia

Birhan Ambachew Taye¹ · Alemayehu Amsalu Alene¹ · Ashenafi Kalayu Nega¹ · Bantie Getnet Yirsaw¹

Received: 16 March 2020 / Accepted: 20 August 2020 / Published online: 3 September 2020
© The Author(s) 2020

Abstract

Milk production is an integral part of Andassa agricultural farming system, even though the area has potential for milk and dairy products, there is always a great demand for milk and milk products among people. However there are no long-term researches done on the area for forecasting the volume of milk production. Therefore this study was attempted to investigate the trends of actual yield of cow milk production and forecast the volume of milk. A time series study was conducted on the volume of cow's milk production in Andassa dairy farm, and here we used secondary data for this study. The study employed inferential statistics for the purpose of hypothesis testing, estimating the trend, to fit the appropriate model and forecasting. A total of 179 days of milk production were included in this study, and the results showed that the trend of milk yield of caws is decreasing and it is time dependent. The model that has been selected for forecasting volume of milk production is ARIMA (1, 2, 1) since the model was found to be a better model.

Keywords Milk production · Forecasting · Time series analysis · Ethiopia · Andasa

Background

Milk is used to hydrating dairy products such as butter, cheese and milk powered. There is always a great demand for milk and milk products among people. Most countries to fulfill the needs of the people expands their production system and techniques to increase the supply of milk production (Haenlein and Wendorff 2006). A number of technological innovations now permit automated daily monitoring of cow performance. So use of automated milk yield recording systems for early detection of diseases requires a statistical model to forecast expect performance and to compare forecast and actual performance. Previous researches on modeling milk production in cows had focused on fitting linear or

nonlinear deterministic models to daily, weekly, or monthly milk measurements from lactations sing either partial or complete lactation data sets (Jacobs and Siegford 2012).

It is essential that researchers and dairy development agents have to understand the existing situations, to design relevant development strategies that to the area (Tassew and Seifu 2009). Hence, milk production is an integral part of Andassa agricultural farming system, and the area has favorable climate and potential for milk and dairy products, so the level of its milk production is low when compared to some dairy sectors of tropical countries. So to design relevant development strategies that appropriate to the area, it is essential that to have in-depth understanding about the past behavior of the sector so that this will helps to evaluate the strengths and weaknesses of strategies adopted in the past and also provides the necessary basis for setting future growth targets. This study was designed to; examine daily trend analysis of milk production, to fit the appropriate model, to forecast the milk production for the future based on the previous recorded data.

✉ Birhan Ambachew Taye
workneshambachew@gmail.com

Alemayehu Amsalu Alene
alexso21@gmail.com

Ashenafi Kalayu Nega
kalayuashenafi@gmail.com

Bantie Getnet Yirsaw
bantiegetnet@gmail.com

¹ Department of Statistics, Woldia University, Woldia, Ethiopia

Data and methods

This study was conducted at Andassa dairy farm which is located around Bahir Dar in north western Ethiopia. Here in Andassa dairy farm, milk productions were recorded in litters per cow on daily basis and the data provided information on volume of daily milk recorded in litters per cow, so this study was employed only secondary data obtained from this dairy sector. Time series analysis method was employed for analyzing this study.

Time series method

Time series analysis comprises methods or processes that breakdown a series into components and explainable portions that allows trends to be identified, estimates and forecasts to be made (Kantz and Schreiber 2004). Basically time series analysis attempts to understand the underlying context of the data points through the use of a model to forecast future values based on known past values. Such time series models include MA, AR, ARIMA, GARCH, TARCH, EGARCH, FIGARCH, CGARCH and ARIMA among others but the main focus of this study is based on MA, AR, ARMA, and ARIMA models.

Test of randomness

This test is used to identify whether there is any systematic components in the time series data or the data are dependent in time t . The main concern is on the assumptions that the data are independent and identically distributed (Beck and Katz 2011). To do this, there are four common test of randomness. These are, rank test, turning point test, different sign test and phase length test. Here the researcher applies only different sign test (Figs. 1, 2, 3).

Trend analysis

A time series which extends to consistently throughout the entire period of time under considerations and it shows long-term tendency of data to grow or to decline is known as trend analysis (Golyandina and Zhigljavsky 2013). In this case trend analysis employed to the past trend of milk production and to forecast the future amount of milk product. The trend model that are usually considered as the linear trend, in which the mean of Y_t expected to change linearly with time as $E(Y_t) = \beta_0 + \beta_1 t$, or as quadratic function of time $E(Y_t) = \beta_0 + \beta_1 t^2$ or even possibly as an exponential functions of time such as $E(Y_t) = \beta_0 e^{\beta_1 t}$. But, here we consider on linear trend and estimate

Fig. 1 Average volume of milk production (litter) as series level

Fig. 2 Average volume of milk production of data in the first difference

based on least square estimation method, double moving average and double exponential smoothing methods. This paper used the least square estimations methods. In the mathematical form expressed as follows: (Figs. 4, 5, 6, 7).

$$Y_t = \beta_0 + \beta_1 t$$

Then, the values of β_0 and β_1 can be calculated from the following normal equations.

$$\sum Y_t = N\beta_0 + \beta_1 \sum t$$

$$\sum tY_t = \beta_0 \sum t + \beta_1 \sum t^2$$

By, rearrange the equations simply to calculate the unknown parameters of β_0 and β_1 .

where, the values of $\sum Y_t$, $\sum t$, $\sum tY_t$ and $\sum t^2$ are obtained from the given collected data.

Fig. 3 Trend analysis of average milk production

Fig. 4 Stationary trend analysis of average milk production

Fig. 5 Moving average plot

Fig. 6 Autocorrelation plot **Fig. 7** Partial autocorrelation plot

Stationary time series

Stationary is a critical assumption in time series models. Stationary implies homogeneity in the sense that the series behaves in a similar way regardless of time, which means that its statistical properties do not change over time. More precisely, stationary means that the data fluctuate about a mean or a constant level. Many time series models assume equilibrium processes. However in practice, a much weaker definition of stationary, often referred to a weak stationary is used. The condition of weak stationary requires that the elements of the time series should have a common finite expected value and that the auto covariance of two elements should depend only on their temporal separation. Generally stationary time series is one whose statistical properties such as mean, variance and autocorrelation are all constant over time. Mathematically, these conditions are: Mean = μ and variance = σ^2 are constant for all values of t (Mietinen et al. 2016).

Stationary through differencing

The first difference of a time series is the series of changes from one period to the next. If $Y(t)$ denotes the value of the time series Y at period t , then the first difference of Y at period t is equal to $Y(t) - Y(t - 1)$, this implies that,

$$\Delta Y_t = Y_t - Y_{t-1}, \text{ where } \Delta Y_t \text{ is the first difference of } Y_t$$

$\Delta^2 Y_t = Y_t - Y_{t-1} - Y_{t-2}$, where $\Delta^2 Y_t$ is the second difference of Y_t .

General linear processes

Let $\{Y_t\}$ denotes the observed time series and also $\{e_t\}$ represents an observed white noise series, that is sequence of identically distributed with zero mean and independent random variable. If the right side of this expression is truly an infinite series, then certain conditions must be placed on the Ψ weights for the right hand side to the meaningful mathematical expression. For our process, it suffices to assume that,

$$\sum_{i=1}^{\infty} \Psi_i^2 < \infty$$

We should also note that since $\{e_t\}$ is an unobservable, there is no loss in the overview of the above equation if we assume that the coefficient on e_t is 1; effectively $\Psi_0 = 1$ (Box et al. 2015).

Moving average process

In the situation where only a finite number of Ψ weights are non-zero, we called it a moving average process. In this case, we change the notation slightly and we can write as

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q}, \quad (1)$$

this is moving average of order q and abbreviated as MA (q). The terminology of moving average arises from the fact that Y_t is obtained by applying the weights $1, -\theta_1, -\theta_2, \dots, -\theta_q$ to the variables, $e_t, e_{t-1}, e_{t-2}, \dots, e_{t-q}$ and then moving the weights and applying to, $e_{t+1}, e_t, e_{t-1}, \dots, e_{t-q+1}$ to obtain Y_{t+1} and so on.

For the general MA (q) process $Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q}$, similar calculations shows that

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2) \sigma_e^2$$

$$\text{And } \rho_k = \frac{-e_k + \theta_1 e_{k+1} + \theta_2 e_{k+2} + \dots + \theta_q e_{k+q}}{1 + \theta_1^2 + \theta_2^2 + \dots + \theta_q^2}$$

Otherwise 0 for $k > q$, where the numerator of ρ_q is just $-\theta_q$. The autocorrelation function “cuts off” after lag q , and its shape can be almost anything for the earlier lags (Fan and Yao 2008).

Autoregressive process

Autoregressive models are based on the idea that the current value of the series, x_t , can be explained as a function of past values, $x_{t-1}, x_{t-2}, \dots, x_{t-p}$, where “ p ” determines the number of steps into the past needed to forecast the current value. An autoregressive of order “ p ” can be written as:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

where x_t is stationary series, $\phi_1, \phi_2, \dots, \phi_p$ are the parameters of the AR ($\phi_p \neq 0$) (Lütkepohl et al. 2004).

Autocorrelation function (ACF)

It is known that the key properties of any random variable are the two moments, namely its mean and the variance, and as we assumed in the introduction that a time series is a realization of a stochastic process this also applies for any time series. But in the context of time series analysis, the relationships between observations in different time-periods play also a very

important role. These relationships across time can be captured by the time series correlation, respectively (Brockwell and Davis 2016).

Partial autocorrelation function (PACF)

The partial autocorrelation function (π_k), where $k \geq 2$, is defined as the partial correlation between Y_t and Y_{t-k} under holding the random variables in between Y_u , where $t-k < u < t$, constant. The partial autocorrelation plot or partial correlogram is also commonly used for model identification in Box and Jenkins models (Inoue 2008).

ARMA model

ARMA model is a mixed model in which the series is partly auto regressive and partly moving average, we obtain a quite general time series model in general the model is:-

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \dots + \Phi_p Y_{t-p} + e_t + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q}$$

We say that $\{Y_t\}$ is a mixed auto regressive moving average process of order p and q , respectively, we abbreviate the name to ARMA (p, q) (Tsay and Tiao 1984).

Box Jenkins modelling method

Box Jenkins model is one of the classes of model to choose the systematic approach to identifying the correct model form (Hipel et al. 1977). There are statistical tests for validity of the model using this test it is possible to identify the best appropriate Box Jenkins model, to fitting the data on the milk production at Andassa dairy farm. To use Box Jenkins method for the time series, data must be stationary after one or differencing for the reason to assuming stationery if to identify and estimate a statistical model, which can be interpreted as having the generated for the sample data. After finishing the process of stationary we used autocorrelation and partial autocorrelation functions to identify the model which is the good fit to the collected data. From this fitted model parameters are estimated by minimizing the residuals sum of squares. Finally, using these parameters, it is easy to calculate the increments or decrements of milk production at Andassa dairy farm. Therefore, in this study we identified an appropriate Box-Jenkins process or model, fitting to the data and then using the fitted model for application based on the data that comes from Andassa dairy farm.

Result and discussion

Descriptive summary

A common first step in data analysis is to summarize information about variables in the data using descriptive statistics such as the minimum, maximum, averages, variances, quartiles and percentiles, of variables set. We were collected secondarily data of the daily milk production of caw at Andassa dairy farm. There are 60 caws in the farm, but from this number of caws we selected 20 caws which have the same day of milking and we took the average daily milk production for 179 days of these caws.

As we look from the above average daily milk production table the minimum and maximum record of milk production at Andassa dairy farm is 715.0 and 2295.0, respectively, observed in the day 11/12/2017 and 11/2/2017 respectively, and the mean for the average daily milk production is 1416.9.

Test of randomness using different sign test

Test of hypothesis,

H0: The data of daily recorded in the average milk production are random.

H1: The data of daily recorded in the average milk production are not random.

Test statistics

$$Z_{cal} = \frac{w - E(w)}{\sqrt{\text{Var}(w)}}$$

where $E(w) = \frac{1}{2}(N - 1)$

$$\text{var}(w) = \frac{1}{12}(N + 1).$$

From thus we have the number of points increase $W = 81$

$$E(w) = \frac{1}{2}(N - 1) = \frac{1}{2}(179 - 1) = 89.$$

$$\text{var}(w) = \frac{1}{12}(N + 1) = \frac{1}{12}(179 + 1) = 15.$$

$$\sqrt{\text{Var}(w)} = \sqrt{15} = 3.87.$$

$$\text{Therefore: } Z_{cal} = \frac{81 - 89}{3.87} = -2.0672.$$

Note: level of significance use in this case is $\alpha = 0.05$.

Test rule:

We reject the null hypothesis since $|Z_{cal}| = 2.0672 > Z_{\frac{\alpha}{2}} = 1.96$. Thus, at five percent level of significance, the data of the average daily milk production is not random. This shows there is systematic pattern on the data and also it indicates, disseminations of the data are not balanced and it must be transformed in to random form using different methods.

Stationary time series

Stationary is a critical assumption in a time series models. Stationary implies homogeneity in the sense that the series behaves in a similar way regard less of time, which means that its statistical properties do not change over time. This can be judged by looking its time plot. And the time plot appears similar at different points along the time axis, and the time series plot is as shown below.

As we see from the above graph of average daily milk production of caw (Y), it is decreasing through time (t) and it is not stationary, so we must be transform in to stationary form of time series plot using differencing method.

Then after transforming the data by differencing in lag two it become stationary series as shown below.

As we see from the above time series plot the series fluctuates about the mean. This means that it does not rend away or drift away from the mean and the time plots appears similar at different points. Because of this reason the time series is stationary meaning, the series fluctuate of movement in the average daily milk production of caw is not far away from one to the other.

Estimation of trend analysis

Trend and difference stationary

A random walk with or without a drift can be transformed to a stationary process by differencing (subtracting Y_{t-1} from Y_t , taking the difference $Y_t - Y_{t-1}$) correspondingly to $Y_t - Y_{t-1} = e_t$ or $Y_t = Y_{t-1}\alpha + e_t$ and then the process becomes difference-stationary. The disadvantage of differencing is that the process loses one observation each time when difference is taken.

The trend analysis plot shows it is not stationary, and then it also needs differencing to be stationary. After differencing the data by lag two, the trend analysis becomes stationary as shown in the graph bellow.

As the fitted trend shows, the amount of average daily milk production is decreasing. The slope of the trend for 179 days data is -16.066 which is the rate at which the amount average milk per day are going to decrease. This also indicates that, the amount of average daily milk decreases from time to time.

Table 1 Descriptive statistics on average daily milk production

Variable	<i>N</i>	<i>N</i> *	Mean	SE mean	St DEV	Minimum	Q1	Median	Q3	maximum
Average	179	0	1416.9	28.4	380.0	715.0	1180.0	1370.0	1780.0	2295.0

Moving average

After the data are transformed in to stationary form using differencing the observations, the moving average plot becomes as shown below.

Box-Jenkins model

Most statistical methodologies are developed by assuming that the observations are random but the data in which successive observations correlated in such a case methods designed to export *ARIMA* (p, d, q). *ARIMA* models are, in theory, the most general class of models for forecasting a time series which can be stationeries by transformations such as differencing and lagging. To identify the appropriate *ARIMA* model for a time series, we begin by identifying the order(s) of differencing needing to stationary the series.

Auto correlation function of daily milk production

It is known that the key properties of any random variable are the two moments, namely its mean and the variance. But in the context of time series analysis, as the definition of a time series indicates, the relationships between observations in different time-periods play a vital role. These relationships across time can be captured by the time series correlation. The plot of the ACF over time is used to identify the order and degree of non-stationary of an *ARIMA* model.

From the above autocorrelations graph there is lagged number of the series and here we must test the *AR* model to check the model adequacy. That is, the average milk production under the autocorrelations graph show there are one points outside the lower bound that is *AR*(1).

Partial autocorrelation function: for average daily milk production

The partial autocorrelation plot or partial correlogram is also commonly used for model identification in Box and Jenkins models.

Similarly, from the above partial autocorrelation graph there is one lagged number. So we must test the *MA* model to check the model adequacy. That is, the average daily milk production under the partial autocorrelations graph shows there is a point outside the lower bounds that is *MA*

Table 2 Accuracy measures

MAPE	MAD	MSD
234.4	81.0	16,358.8

For all those three measures, the smaller value is better fit of the model, i.e., $MAD=81.0$ is better fit of the model

Table 3 Final estimates of parameters

Type	Coef	SE Coef	<i>T</i>	<i>P</i>
AR 1	− 0.2066	0.0742	− 2.79	0.006
MA 1	0.9977	0.0004	2352.08	0.000
Constant	0.2512	0.1447	1.74	0.084

(1). Meaning, disseminations' of the average daily milk production is not balance or equal.

ARMA process

ARMA is the combination of the autoregressive and moving average processes, so we have *AR*(1) and *MA* (1), we obtain a quite general time series model *ARMA*(1,1). (Tables 1, 2).

ARIMA process

One implication of above is that *AR* (p) process with a unit root can always be written as in the form *AR* ($p - 1$) process in differences. Such processes are often called auto-regressive integrated moving average processes *ARIMA* (p, d, q) where the d refers to how many times the data is differenced before it is an *ARMA*(p, q). Therefore the *ARIMA* model becomes:-

ARIMA (p, d, q) = *ARIMA* (1,2,1).

Differencing: 2 regular differences.

Number of observations: Original series 177, after differencing 175.

Residuals: $SS=2,732,779$ (back forecasts excluded).

$MS=15,888$ $DF=172$.

As we have seen from the MINITAB output (Table 3) the *ARIMA* model (1, 2, 1) the equations is described as follows.

$$Y_t = -0.2066Y_{t-1} + 0.9977e_{t-1} + e_t$$

Table 4 Modified box-pierce (ljung-box) chi-square statistic

Lag	Chi-square	DF	P-Value
12	67.1	9	0.000
24	92.0	21	0.000
36	102.9	33	0.000
48	109.0	45	0.000

Table 5 Confidence interval for the forecasted values

Period	Forecast	Lower	Upper
180	26.663	0	273.768
181	54.100	0	369.879
182	82.662	0	460.574
183	111.242	0	541.606
184	114.856	0	598.253
185	158.258	0	625.356

Testing of parameter

The final estimates are those that minimize the sum of squared errors to a point where no other estimates can be found that yield smaller sum of squared errors. A model should have significant parameter as shown in the MINITAB output. ARIMA (1, 2, 1) have a p -value which is less than the level of significance ($\alpha=0.05$). This means that the parameters are significantly different from zero, and have minimum sum squared error. Then, it has statistically significant parameters. So, the model is adequate.

Forecasting

After identification of the model and estimation of the parameters, the next step is forecasting with the appropriate model for the average daily milk production of cow at Andassa dairy farm. The appropriate model for this data is ARIMA (1, 2, 1) to obtain the point of forecast. The final model must be rewritten in the original form is described as follows.

$$Y_t = -0.2066 Y_{t-1} + 0.9977 e_{t-1} + e_t$$

95 Percent limits.

We can check the accuracy of the forecasted value using the above 95% confidence interval described as follows. All the forecast values are found between the lower and the upper interval then we can say that forecasted value is accurate (Table 4, 5).

Conclusions

The study was conducted using the data from secondary source. Based on the analysis conducted in this research on the average daily milk production, the following ideas are summarized. Based on the time series plot, the milk production is time dependent. The amount milk production at Andassa dairy farm the mean amount of the milk production is decreasing. The 179 day shows that the high variability in the amount of daily milk production than other days, this means that the amount milk is highly different from day to day. The minimum and maximum record milk production at Andassa dairy farm is 715.0 and 2295 observed in the day 11/12/2017 and 11/2/2017 respectively. The amount of milk is decreasing, since the slope of the trends is -16.066 . Generally daily milk production under the autocorrelations and partial autocorrelations graph shows it does not include the whole observation in the upper and lower bounds. Meaning, disseminations' of the milk production is not balance at Andassa dairy farm. For ARIMA (1, 2, 1) is a parameters which has a p value which is less than the level of significance ($\alpha=0.05$). This means the parameter is significantly different from zero, and has minimum sum squared error. Then, the parameter is statistically significant.

Recommendations

After observing all results of the study, the following recommendations can be made.

From the result of this study the amount of daily milk is decreasing from day to day. Even if the study is limited at Andassa dairy farm; the concerned body should take the following remedial actions.

Concerned bodies are recommended to take the remedial action that decreases the milk production. Since this decreasing of milk production may come from different factors so this factors must be study for the future to gain high amount of milk.

Regarding the dairy sector managers, data should be well organized and reported to have role in further studies so that the organization will beneficial.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Beck N, Katz JN (2011) modeling dynamics in time-series–cross-section political economy data. *Ann Rev Polit Sci* 14:331–352
- Box G, Jenkins E (2015) *Time series analysis: forecasting and control*. Wiley, Hoboken
- Brockwell P, Davis P (2016) *Introduction to time series and forecasting*. Springer, Berlin
- Fan J, Yao Q (2008) *Nonlinear time series: nonparametric and parametric methods*. Springer, Berlin
- Golyandina N, Zhigljavsky A (2013) *Singular spectrum analysis for time series*. Springer, Berlin
- Haenlein Y, Wendorff W (2006) Sheep milk. *Handbook of milk of non-bovine mammals*. Blackwell, Oxford, pp 137–194
- Hipel KW, Mcleod AL, Lennox WC (1977) Advances in box-jenkins modeling: model construction. *Water Resour Res* 13:567–575
- Inoue A (2008) Arand ma representation of partial autocorrelation functions, with applications. *Prob Theor Relat Fields* 140:523–551
- Jacobs JA, Siegford J (2012) Invited review: the impact of automatic milking systems on dairy cow management, behavior, health, and welfare. *J Dairy Sci* 95:2227–2247
- Kantz H, Schreiber T (2004) *nonlinear time series analysis*. Cambridge University Press, Cambridge
- Lütkepohl K, Phillips M (2004) *Applied time series econometrics*. Cambridge University Press, Cambridge
- Miettinen J, Illner K, Nordhausen K (2016) Separation of uncorrelated stationary time series using autocovariance matrices. *J Time Ser Anal* 37:337–354
- Tassew A, Seifu E (2009) Smallholder dairy production system and emergence of dairy cooperatives in bahir dar zuria and mecha woredas, northwestern ethiopia. *World J Dairy Food Sci* 4(2):185–192
- Tsay RS, Tiao C (1984) Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary arma models. *J Am Stat Assoc* 79:84–96

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.