Employment Of Arima Models For Prediction Of Sugar Cane In Sergipe, Brazil

Suzana Leitão Russo; Iracema Aragão Gomes Machado, Leda Valéria Santana, Jonas Pedro Fabris

Post-Graduate Program in Intellectual Property Science
Department of Statistic
Fabris Industry Federal University of Sergipe, São Cristóvão, Sergipe, Brazil

ABSTRACT
Sugar cane is considered a grass, native to Southeast Asia; it belonging to the genus Saccharin L., fits easily in tropical regions, hot and humid, the temperature prevalent is between 19 º and 32 º C and where the rains are well distributed, with accumulated rainfall above 1000 mm per year (EMBRAPA, 2011). Agriculture sugarcane was from the sixteenth century, the most important sector of the Brazilian colonial economy, and the culture of sugar cane responsible for the support base of the economy and the colonization of Brazil for centuries and remains to this day as a major product of the Brazilian economy. Among the crops grown in Brazil, sugarcane is the third in relation to acreage, second crop of soybeans and corn. Brazil is the world leader in the production of cane sugar, and processed approximately 569 million tons in 2008/2009, about 90% of the total in the main producing region of the country, the Mid-South, and 10% in Northeast. The importance of cane sugar may be related to the diversity of its use, can be used fresh, in the form of fodder, as animal feed or as raw material for the manufacture of brown sugar, molasses, alcohol, sugar and brandy. Due to the immense number of ethanol producers in Brazil, one can not treat sugarcane, as just one product, but as the main type of biomass energy, the entire basis for sugarcane agribusiness, represented by 350 industries of sugar and alcohol and one million direct and indirect jobs in Brazil (EMBRAPA, 2011). This article aims to analyze the production of this raw material, through descriptive analysis of data using graphs and tables, verifying the presence of autocorrelation. It will also apply the ARIMA models to determine the best forecasting for the short term of the sugar cane, using the MAPE for choosing the best model for the series under study. After several studies found that the ARIMA (2,1,0) with the lowest MAPE (10.73%), considered the most suitable among the models that were found to generate forecasts.

INTRODUCTION

The cultivation of sugar cane was probably domesticated in New Guinea, being introduced in Brazil in 1532 in the expedition of Martim Afonso de Souza donee Captaincy of São Vicente. The Portuguese realizing the potential producer of Brazil's land decided to invest even more, making the country the world's largest producer (MORAES, 2004).

Among the crops grown in Brazil, sugar cane occupies third place in relation to the area planted, behind soybean and corn only. Brazil is the world leader in the production of cane sugar, and processed approximately 569 million tons in 2008/2009, about 90% of the total in the main producing region of the country, the Mid-South, and 10% in Northeast (UNICA).

Sugarcane contributes to the sustainability of the planet and the fight against global warming, since it is considered the most efficient feedstock for ethanol production, obtained from sugar cane juice, and bioelectricity, obtained from biomass formed by the marc (residue fibrous generated after juice extraction) and straw (tops and leaves) of sugar cane (UNICA, 2011).

Among the institutions that perform research to develop new varieties of sugar cane genetically improved, have been the Inter-University Network for the Development of Sugarcane Industry in Brazil (RIDESA), founded in 1991, which accounts for cultivars planted in more than 50% of the total Brazilian cane sugar. Its base consists of 31 stations, located in the states where the culture of cane sugar has a higher expression (RIDESA, 2011).

This study will develop a model to forecast the production of cane sugar in the coming years. Since this raw material which for centuries has solidified its importance in the economy.
Theoretical Conceptions:
Sugar cane:
The industry of cane sugar has long been the basis of the Brazilian economy. From 1532 when the first
seedlings arrived in the country for more than two centuries sugar was the main Brazilian products. For nearly
40 years, began the transformation of the industry (MACEDO E SOUSA, 2010).
The main stages of the cane fall into two stations with different characteristics, the tropical climate that
provides germination, tillring and vegetative growth, followed by a period of cold and dry, to promote the matu-
ration and consequent accumulation of sucrose in stems (EMBRAPA, 2011).
The importance of cane sugar can relate to the diversity of its use and may be used in nature, in the form of
fodder, as animal feed or as raw material for the manufacture of brown sugar, molasses, alcohol, sugar and
brandy.
The environmental impacts of cane sugar have been discussed in various forums, with important progress
in the last ten years. The harvest is the most delicate step of the production process of sugar cane, because of the
fires (EMBRAPA, 2011).

Time series:
The time series analysis aims to: investigate the mechanism generating the time series; to forecast future
values of the series, to describe the behavior of the series; seek relevant periodicities in the data. A model that
describes a series does not necessarily lead to a procedure (or formula) prediction. You need to specify a func-
tion-loss, beyond the model, to get the procedure. A function-loss, which is often used, is the mean square error,
although on some occasions, other criteria or loss functions are more appropriate (MORETTIN AND TOLOI,
2006).

Autocorrelação:
The autocorrelation is a measure of dependency between observations Same series separated by a given
range named retardation.

Be a time series $Y_t$. The ratio between the covariance ($Y_t, Y_{t-k}$) and variance ($Y_t$) defines a autocorrelation
coefficient simple ($r_k$), while the sequence of $r_k$ values is called autocorrelation function simple (FAS)
(MATOS, 2000).
The graphical representation of this function is called correlogram. Formally, the autocorrelation coefficients
simple between $Y_t$ and their $Y_{t-k}$ lagged values, are defined by:

$$r_k = \frac{\text{cov}(Y_t, Y_{t-k})}{\text{var}(Y_t)} = \frac{\sum_{t=1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2}$$ (1)

We can see the existence of unit root if the values of the autocorrelation function begin near to unit and de-
cline slowly and gradually as increases the distance (number of lags, k) between the two sets of observations to
which they concern, calling himself, not stationary and follows a random walk. If these coefficients decline ra-
pidly as this distance increases, there is a series of characteristics of stationary (MORETTIN AND TOLOI,
2006).

Stationary Processes:
A process is considered stationary if its statistical characteristics do not change with time. Therefore, the
characteristics $Y_{(t+n)}$, for all n, will be the same as $Y_t$. (MORETTIN AND TOLOI, 2006).

Processes Non-Stationary:
Where a series submit over time variation in your parameters have if then a series non-stationary, which
when submitted to differentiation process becomes stationary (MORRETIN AND TOLOI, 2006).

Time series models:
Autoregressive models - AR(p):
The class of models purely autoregressive is defined by:

$$Y_t = \frac{a_t}{\phi_p(B)}$$ (2)
where $\phi_p(B)$ has $p$ coefficients. The AR (p) assumes that the result is the weighted sum of its $p$ past values than white noise. The condition of stationarity of the AR (p) states that all the $p$ roots of the characteristic equation fall outside the unit circle (RUSSO, ET AL, 2009; SHUKLA, 2011).

**Moving average models - MA(q):**

According to Russo, et al (2009) and Padhan (2012), the class of moving averages models is defined by

$$Y_t = \theta_q(B)\alpha_t$$ (3)

where $\theta_q(B)$ has $q$ coefficients. The models MA (q) resulting from the linear combination of random shocks that occurred during the current and past periods. The invertibility condition requires that all roots of the characteristic equation fall outside the unit circle.

**Autoregressive and moving average models - ARMA (p,q):**

The class of models, autoregressive-moving average is of type

$$Y_t = \frac{\theta_q(B)\alpha_t}{\phi_p(B)}$$ (4)

where $\phi_p(B)$ has $p$ coefficients and $\theta_q(B)$ has $q$ coefficients. With a combination of models AR (p) and MA (q), it is expected that the models ARMA (p,q) be models extremely parsimonious, using few coefficients to explain the same serie. From the standpoint of adjustment, it is very important because you can adjust more quickly. The condition of stationary and invertibility of a ARMA (p, q) require that all $p$ roots of $f(B) = 0$ and all the $q$ roots of $q(B) = 0$ fall outside the unit circle (RUSSO, ET AL, 2009).

**Autoregressive Integrated Moving Averages Models - ARIMA (p,d,q):**

The class of autoregressive-integrated-moving-average models are defined by the equation,

$$Y_t = \frac{\theta_q(B)\alpha_t}{\phi_p(B)(1-B)^d}$$ (5)

to an integrator positive $d$. Made the differentiation of the series $d$ times necessary to make it stationary, then the ARIMA(p, d, q) model can be adjusted through the ARMA(p,q) model (RUSSO, ET AL, 2009).

**Sazonal Model – SARIMA:**

According to Fischer (1982), the appearance of some short-term cyclical behavior is called seasonality. For a full treatment about series of time, need to characterize and eliminate this cyclic function of time to become get the condition of stationarity.

Seasonality means a tendency to repeat a certain behavior of the variable that occurs with some regularity in time. That is, are those series that have variations of a similar amount of time to another, characterized by showing high serial correlation between observations of the variable spaced by the period of seasonality, and, of course, the serial correlation between the next observations.

Similar to the process ARIMA (p,d,q) this process develops the model in one of three basic forms of description of each value of $Y_t$, and applies the same procedures developed for a model where the seasonal component is not present. After establishing the value of the variable in period $t+ h$, then applies the expectancy operator. Forecast errors, confidence intervals and updating are treated similarly to the ARIMA model (FISCHER, 1982).

**A Metodologia ARIMA:**

This method for the prediction is based on the setting called tentative ARIMA models, has a flexible modeling methodology that forecasts are made from the current and past values of these series. Therefore, describing both the stationary behavior as the non-stationary zero. ARIMA models are able to describe the process of generating a variety of series for forecasters (corresponding to the filters) without taking into account the economic relations, for example that generated the series (MORRETIN E TOLOI, 2006).

The determination of the best model for "Box and Jenkins" methodology following this steps (LEROTY, 2006; RUSSO, et al, 2013):
Identification:
Identification is the most critical phase of the "Box and Jenkins" methodology, it is possible that several researchers to identify different models for the same series, using different criteria of choice (FAC, FACP, Akaike, etc.). Typically, the models should be parsimonious. The study analyzes the FAC and FACP, and attempts to identify the model. The process seeks to determine the order of \((p,d,q)\), based on the behavior of the Autocorrelation Functions (ACF) and Partial Autocorrelation (FACP), as well as their respective correlograms.

Estimation:
After identifying the best model should then adjust and examine it. The adjusted models are compared using several criteria. One of the criteria is the of parsimony, in which it appears that the incorporation of coefficients additional improves the degree of adjustment (increases the R2 and reduces the sum of squared residuals) model, but you reduces the degrees of freedom. One of ways to improve the degree of adjustment of this model to time series data is to include lags additional in Cases AR \((p)\), MA \((q)\), ARMA \((p, q)\) and ARIMA. The inclusion of additional lags implies increasing the number of repressors, which leads to a reduction in the sum of squared residuals estimated. Currently, there are several criteria for selection of models that generate a trade-off between reductions in the sum of squared residuals and estimated a more parsimonious model.

Generally, when working with lagged variables are lost about the time series under study. Therefore, to compare alternative models (or competitors) should remain fixed number of information used for all models compared.

Checking:
Aspiring to know the efficacy of the model found, takes place waste analysis. If the residuals are autocorrelated, then the dynamics of the series is not completely explained by the coefficients of the fitted model. It should be excluded from the process of choosing the model(s) with this feature. An analysis of existence (or not) of serial autocorrelation of waste is made based on the functions of autocorrelation and partial autocorrelation of waste and their respective correlograms. It is noteworthy that, when estimating a model, it is desired that the error produced by it have characteristic "white noise" that is, this will be independent and identically distributed (i.i.d. condition).

Forecast:
Predictions can be ex-ante, made to calculate future values of short-term variable in the study. Or, ex-post held to generate values within the sample period. The better these last, the more efficient the model estimated. We choose the best model through the lower Mean Absolute Percentage Error (MAPE). It is a formal measure of the quality of forecasts ex-post. Therefore, the lower value of the MAPE is the best fit of forecasts of the model to time series data.

Methods:
The database used was acquired at the site of the Economically Research Institute (IPEA) and correspond to the period from 1990 to 2007 the production of sugar cane in the state of Sergipe (Brazil). It will be checked for the presence of the autocorrelation, then Box and Jenkins models will be determine, through the MAPE for the best model in the series on study, in order to make a prediction for the production of cane sugar in the following years.

Characterization of variable sugar cane - is a plant belonging to the genus Saccharin L. is a multi-species hybrid, receiving the designation Saccharin spp. The species of cane sugar are from Southeast Asia. The importance of sugar cane can be attributed to its multiple uses and can be used fresh, in the form of fodder, feed, or as raw material for the manufacture of brown sugar, molasses, rum, sugar and alcohol (EMBRAPA).

The software used in this paper was STATISTCA 11, which is licensed by Federal University of Sergipe.

RESULTS AND DISCUSSIONS

Figure 1 shows the behavior of the variable under study, illustrating the growth of production of cane sugar during the years 1900 to 2007 in the state of Sergipe (Brazil).

It is observed in Figure 1, the growth of cane sugar remained stable between 1900 and 1918 in the 20 notices that their productivity increases, but its rise is realized in the early 70's when it was released the Pro-Alcohol program, aiming to reduce energy dependency of the country from large investments in production and subsidies for the development of a consumer market of alcohol.
Making Brazil the only one able to resolve the crisis of international oil supplies, with the production and use of an alternative fuel, through government incentives.

**Table 1: Summary statistics**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>104299175</td>
</tr>
<tr>
<td>Median</td>
<td>39319344</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>129488020</td>
</tr>
<tr>
<td>CV%</td>
<td>124.1506</td>
</tr>
<tr>
<td>Coefficient Asymmetry</td>
<td>1.433265</td>
</tr>
</tbody>
</table>

The production of sugar cane has an average of 104299175 tons / year. It can also be observed through the median, 50% of sugarcane production is exceeding 39319344 tons / year. Analyzing the diversion perceiving themselves that productivity fluctuates throughout the years, however in most years this oscillation is positive (fact that can be observed by Figure 1). Being that the presence of the high values is justified due to production measure be made in ton.

In Figure 2, there is the normality test of Kolmogorov-Smirnov and Shapiro-Wilk noting that one can not reject the hypothesis that the data do not follow a normal distribution, because the test p-value that is less than 0.05.

**Fig. 1:** Chart of the Production of Sugar Cane

**Fig. 2:** Histogram of the observed data
Verifies through of Figure 3 (a) that some lags are out the control limits, indicating that the observed data are autocorrelated, event that can be confirmed when analyzes the Figure 3 (b).

![Autocorrelation Function](image1) ![Partial Autocorrelation Function](image2)

**Fig. 3:** (a) Autocorrelation Function (b) Partial Autocorrelation Function

We calculated several models until you find the best model for the series. Here we show only the models ARIMA (1,1,0) and ARIMA (2,1,0).

The series for be analyzed by ARIMA model should be stationary, or must have mean, variance and autocorrelation constant over time. Usually there is a need to take differences in series to achieve such condition. It was necessary to differentiate in series for the removal of the autocorrelation. The Figure 4 shows de first difference of the series.

![Graph of the Difference](image3)

**Fig. 4:** Graph of the Difference

To the ARIMA (1,1,0) model, was also generated the chart that identify the presence of autocorrelation and partial autocorrelation. It was found that despite the model not be the most appropriate, the autocorrelation presented himself better due the lags being within the control limits, Figure 5 (a and b).
According to Table 2, it is evident that the first model presents a lower error. So, have another event confirming that the ARIMA (2,1,0) is the best model to forecast the production of sugar cane.

Table 2: Parameters of models

<table>
<thead>
<tr>
<th>Models</th>
<th>Parameters</th>
<th>Estimate</th>
<th>Standard-Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1,0)</td>
<td>P(1)</td>
<td>0.48801</td>
<td>0.1040</td>
</tr>
<tr>
<td>(2,1,0)</td>
<td>P(1)</td>
<td>0.39729</td>
<td>0.1004</td>
</tr>
<tr>
<td></td>
<td>P(2)</td>
<td>0.44002</td>
<td>0.1131</td>
</tr>
</tbody>
</table>

Checking the correlograms, Figure 6 (a and b) of the ARIMA (2,1,0), we find that the autocorrelation is corrected, showing that most of the lags are among the control limits.

Table 3: Observed and predict values of the models

<table>
<thead>
<tr>
<th>Predict</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1,0)</td>
<td>(2,1,0)</td>
</tr>
<tr>
<td>374196694</td>
<td>380369644</td>
</tr>
<tr>
<td>378982739</td>
<td>395561365</td>
</tr>
<tr>
<td>381318373</td>
<td>408628573</td>
</tr>
<tr>
<td>382458185</td>
<td>420504767</td>
</tr>
<tr>
<td>383014424</td>
<td>430972947</td>
</tr>
</tbody>
</table>

Data were submitted to two ARIMA models, as shown below with their respective MAPE.
Table 4: MAPE of the models

<table>
<thead>
<tr>
<th>Model</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA (1,1,0)</td>
<td>18.93</td>
</tr>
<tr>
<td>ARIMA (2,1,0)</td>
<td>10.73</td>
</tr>
</tbody>
</table>

Considering that the best model should have the lowest MAPE, so we can note that the best model found was ARIMA (2,1,0) with MAPE (10.73) represents the best model to generate forecasts. Figure 7 is illustrated the graphic forecast, it is clear that the production of sugar cane tends to grow in coming years.

Fig. 7: Graph of forecast production from sugar cane.

Conclusions:

The cultivation of sugarcane responsible for the support base of the economy and the colonization of our country over the centuries remains to this day as a major product of the Brazilian economy, its widespread use encourages studies that are a source of knowledge for investment in sector.

The models for time series forecasting used in this study provide information through statistical analysis, to predict the future production of the raw material as a basis for large plans. The predictive analysis is extremely important because it is through them that are made possible projects that minimize negative externalities.

The aim of this study was to predict the production of sugar cane in the state of Sergipe, through the time series models. The data were subjected to procedures that generated models, which showed different statistics, and may make a comparison between them. And so, choose the best model. Through the correlogram found that the observed data were autocorrelated, and after modeling this correlation has been corrected. Even in the model presented is not suitable because the MAPE.

With the application of time series, it was possible to predict the production of cane sugar for the following years, through the ARIMA (2,1,0), verifying that this production tends to increase. Due to investments in new technologies, increase internal productivity. The discovery of the usefulness of the raw material for other purposes such as ethanol production, thermoelectricity, biogas.

REFERENCES


