Application of statistical methods of time-series for estimate and predict of the food gap in Yemen

Douaik Ahmed¹, Youssfi Elkettani², and Abdulbakee Kasem²

¹Institut National de Recherche Agronomique de Rabat, Morocco
²Department of Mathematics Faculty of Sciences, University Ibn Tofail, Kenitra, Morocco

Abstract: In this paper we provided modeling for the food gap in Yemen. We have studied this model by descriptive and analytical studies and formulated a model for the food gap, we estimated its parameters and predicted for the coming ten years using the Box and Jenkins methodology of time series analysis. Then, we compared this methodology to the exponential smoothing and simple regression methods.

We found the following main results for the three time series regarding the food gap:
1. ARIMA (1, 1, 1) model to predict the price of food importation series.
2. Brown exponential smoothing model to predict the price of food exportation series.
3. ARIMA (1, 1, 1) model to predict the price of food production series.

Through the results, we concluded that food production will not meet the local demand for food, where of the equation: local demand consumption of food = food importation + food production - food exportation. The ratio of production to consumption is expected to reach 29.3 % in 2015 and to continue to decline to reach 28.8 % in 2020.

Keywords: Time Series, Food security, Forecasting.

Introduction

Forecasting is a decision-making tool used by many Governments and businesses to help in budgeting, planning, and estimating future growth. In the simplest terms, forecasting is the attempt to predict future outcomes based on past events and management insight. The problem of food security is a very important subject which must be mobilized government officials and decision-makers to this fact then. Yemen is one of the the developing countries where local demand for food is growing exponentially, this means that it suffers from a huge lack to cover all the population needs of foodstuffs. Therefore, we applied statistical methods for time series by compared between the Box–Jenkins method which discovered in 1970 by Scientists George Box and Gwilym Jenkins (autoregressive moving average ARMA or ARIMA models), exponential smoothing (trend, seasonal) which is used a lot in the short-term forecast and simple linear regression, to estimate and of the food deficit for three annual data time series from 1961 to 2010 taken of the Organization’s site of Food and Agriculture (FAO) and the Central Bureau of Statistics in Yemen.

Theoretical Formulation

2.1 Simple Regression Analysis

A relation between the dependent variable \( Y \) and the independent variable \( X \) can be modeled as a function of type \( Y = \alpha + \beta X \), \( Y \) : dependent variable (explained), \( X \) : independent variable (explicative).
2.2 Model of Time Series

2.2.1 Definition

[1] A time series model for the observed data \{x_t\} is a specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables \{X_t\} of which \{x_t\} is postulated to be a realization.

2.2.2 Choose the Correct Model for Prediction

For the comparison of time series models by using the methodology of Box and Jenkins, exponential smoothing and simple regression, we used as a measure of the quality of forecasts MAPE (Mean Absolute Percentage Error). This is the relative error made by a predictive model. The following equation shows how we calculate this criterion:

\[
MAPE = \left(\frac{100}{n} \sum_{t=1}^{n} \left| \frac{D_t - P_t}{P_t} \right| \right)
\]

Where: \(y_t\) : initial observation, \(p_t\) : forecasting the initial observation and \(n\) : number of the initial observations [2]. Other criteria for comparison of models:

BIC (Bayesian Information Criterion, 1978), \(R^2\), significance of the parameters, residue analysis: (Ljung - Box, Durbin-Watson). The BIC is formally defined as

\[
BIC = -2 \ln \hat{L} + K \ln(n)
\]

where:
• \(x\) = the observed data;
• \(\theta\) = the parameters of the model;
• \(n\) = the number of data points in \(x\), the number of observations, or equivalently, the sample size;
• \(k\) = the number of free parameters to be estimated. If the model under consideration is a linear regression, \(k\) is the number of regressors, including the intercept;
• \(\hat{L}\) = the maximized value of the likelihood function of the model \(M\), \(\hat{L} = p(x | \hat{\theta}, M)\), where \(\hat{\theta}\) are the parameter values that maximize the likelihood function.

\(R^2\) (the coefficient of multiple determination): The elevation of the linear fit of the regression equation between the dependent variable \(Y\) and the set of explanatory variable is determined by the coefficient of determination \(R^2 = \frac{SCR}{SCT}\) with \(0 \leq R^2 \leq 1\) where \(SCR = \sum(Y_i - \bar{y})^2\), \(SCT = \sum(Y_i - \bar{y})^2\), \(\hat{Y}_i\) is the estimated values, \(Y_i\) is the real values, \(\bar{y}\) is the average of real values.

2.2.3 Forecast by Exponential Smoothing

Series with Trend Component Without Seasonal

\[
\hat{y}_{t+h} = a_t + b_t h, h = 1, 2, 3, \ldots \text{(According to the number required to provide)} \tag{1}
\]

The values \(a_t\) and \(b_t\) are constantly updated by the following:

\[
a_t = \alpha x_t + (1 - \alpha)(a_{t-1} + b_{t-1}) \quad \text{and} \quad b_t = \alpha x_t - a_{t-1} + (1 - \alpha)b_{t-1} \tag{2}
\]

This prediction model is known as Holt’s model name. A special case of Holt model, referred to as Brown model or dual exponential smoothing is obtained when the smoothing constant \(\alpha_1\) and \(\alpha_2\) are linked to the same parameter \(\alpha\), by the relations:

\[
\alpha_1 = \alpha(2 - \alpha) \quad \text{(3)} \quad \text{and} \quad \alpha_2 = \alpha(\alpha - 2) \quad \text{(4)}
\]
For these two models, we need to give initial values \( a_0 \) and \( b_0 \) to produce forecasts. Thus we takes \( b_0 \) equal to simple linear regression coefficient calculated on the basis of the first five series of values, thereafter, \( a_0 \) is deduced by the relation:

\[
a_0 = y_1 - b_0.
\]

### 2.2.4 Forecast by the methodology of Box and Jenkins

Box and Jenkins based on the notion of ARIMA process. This technique has four steps: identification, Estimation, Diagnostic and Forecast. The model consists of two parts: the frist one is autoregression (AR) and the second is moving average (MA) and model is generally noted ARMA \((p, q)\), where \( p \) is the order of the AR part and \( q \) the order of the party MA, a moving average and autoregressive model of order \((p, q)\) (abbreviated as ARMA \((p, q)\)) is a discrete time processes \((X_t, t \in N)\) verifying:

\[
X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \alpha_t - \theta_1 X_{t-1} - \ldots - \theta_q X_{t-q}
\]

where \( \phi_i \) and \( \theta_i \) parameters are constants and the terms of errors \( \alpha_t \) are independent of the process.

### Autoregression model and Moving average model

An autoregressive model of order \( p \), abbreviated AR \((p)\) is written:

\[
Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \alpha_t
\]

and the notation MA \((q)\) refers to the moving average model of order \( q \) is written:

\[
Y_t = \alpha_t - \theta_1 Y_{t-1} - \theta_2 Y_{t-2} - \ldots - \theta_q Y_{t-q}
\]

where \( \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q \) are the model parameters, \( \alpha_t \) an white noise that define is a series of random observation noncorrelation and sometimes assume that it is a series of independent random variables and have the same distribution, zero mean and constant variation such that

1. \( E(\alpha_t) = m, \forall t \),
2. \( Var(\alpha_t) = \sigma^2, \forall t \),
3. \( cov(\alpha_t, \alpha_{t+h}) = \gamma(h), \forall t, h > 0 \).

Additional constraints on the parameters are necessary to ensure the stationarity. A process is stationary in the second if:

1. \( \forall t \in \mathbb{Z}, E(X_t^2) < \infty \),
2. \( \forall t \in \mathbb{Z}, E(X_t^2) = m, \forall t \in \mathbb{Z}, \forall h \in \mathbb{Z}, Cov(X_t, X_{t+h}) = \gamma(h) \).

The model requires that a original series \( Y_t \) is stationary, but if it is not, it must be converted to a stationary taking the necessary differences, in this case, the ARMA model is written ARIMA (autoregressive integrated moving average). This model is characterized by three types of parameters: an autoregression \((p)\), an integration \((d)\) and a moving average \((q)\) and is symbolized as follows: ARIMA \((p, d, q)\) \([6]\). The Box-Jenkins methodology is summarized in four steps to found the best model for estimating and forecasting time series. These steps are sometimes nested with one another.

1. Identification: This is the most important step is to select the rank which model consists of it, first Know \( d \) through determine the degree of integration with study the stability of original time series. If the series is not stationary, for example having a general trend, we take the first difference, the series can become stationary after a number of differences, we can use several methods to know stationary of the series as the basic unit root test of Dickey-Fuller test and the expanded Dickey-Fuller \((1979, 1981, 1986)\) \([7]\). Then we know \( p \) and \( q \), we test the significance of the autocorrelation function, through \( SE(r_k) = \frac{1}{\sqrt{n}} \) and test the hypothesis \( H_0 : \rho(k) = 0 \) against \( H_1 : \rho(k) \neq 0 \) for \( k = 1, 2, \ldots \) using \( |Z| = \frac{r(k)}{SE(r(k))} \), \( H_0 \) is rejected if \( |Z| > 2 \) and we deduce that \( \rho(k) \) is significantly different from zero. The standard error for each of the autocorrelation coefficients must be less than \( 2*SE(r_k) \), \( k > q \) where \( SE(r_k) = \sqrt{\frac{1}{n} + 2\sum_{j=1}^{q} r^2(j)}, k > q \). To test the significance of the partial autocorrelation function is
calculated: \( SE(\phi_k) = \frac{1}{\sqrt{n}} \)

and test the hypothesis \( H_0 : \phi(kk) = 0 \) against \( H_1 : \phi(kk) \neq 0 \) for \( k = 1, 2, ..., \) using \( |Z| = \frac{|\phi(kk)|}{SE(\phi(kk))} \). \( H_0 \) is rejected if \( |Z| > 2 \) and we deduce that \( \phi(kk) \) is significantly different from zero. The standard error for each of the autocorrelation coefficients must be less than \( 2*SE(\phi(kk)) \), \( k > p \) where \( SE(\phi(kk)) = \frac{1}{\sqrt{n}} \). It also can be identified through graph ACF, PACF.

2. Estimation: This usually involves the use of a least squares estimation process.

3. Diagnostic: The diagnostic of model is generally to do many tests and checks. The most important are: the stationary, study of residues and statistics.

4. Forecast: After identifying the model (determination of p, d and q), estimated parameters and diagnosed, it is used for the preparation of forecasts.

3 APPLICATION

3.1 THE PRICE SERIES OF VALUES (IMPORTS, EXPORTS AND PRODUCTION) FOOD

3.1.1 GRAPH SERIES

![Graph series of imports food from 1961 to 2010](image)

Fig. 1. Graph series of imports food from 1961 to 2010

Through the graph figure, we observed a general upward trend over the period, this means that the series is not stationary.
3.1.2 AutoCorrelation and Autocorrelation Partial

We examine the autocorrelation and partial autocorrelation function in figures (2) and (2-1) we observed that the estimated autocorrelation parameter decreases exponentially towards zero while that only the first partial autocorrelation parameter is not significant. To confirm the previous results we execute the Dickey-Fuller test and observed in Figure (3) that the series is not stationary.

We note that the series is stationary. We deduce that $d = 1$ in the ARIMA model $(p, d, q)$.

Identification of Model for Wheat Production Series

Although it appears that each partial autocorrelation parameter after the second parameter is not significantly different from zero at $\alpha = 0.05$ but the autocorrelation function is gradually decreasing towards zero, this may be sufficient evidence that the random process is AR (1). For ensure we test the following statistical hypothesis: $H_0 : \phi_1 = 0$ ; $H_0 : \phi_1 \neq 0$,

$$SE(\phi_1) = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{50}} = 0.141, \quad Z = \frac{\phi_1}{SE(\phi_1)} = \frac{0.924}{0.141} = 6.4 > 2,$$

we deduce that the first partial autocorrelation parameter is not significantly different from zero at $\alpha = 0.05$. We examining the autocorrelation partial parameters, we find that $\phi_k < 0.282 \text{ for each } k = 2,3,\ldots$. Thus, can say that there is no reason for the partial autocorrelation function does not stop after the first parameter that supports the possibility of using the AR (1) and therefore $ARIMA(1,1,0)$.
Then we compared between ARIMA models with the exponential smoothing (Holt, Brown) and simple regression. For import food series and product food series we get the same results. We get the following results.

**Table 1. Comparison of model**

<table>
<thead>
<tr>
<th>Model</th>
<th>R-deux</th>
<th>Bic</th>
<th>MAPE</th>
<th>Luing</th>
<th>Sig du Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA (1,1,0)</td>
<td>0.92</td>
<td>24.1</td>
<td>16.9</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>ARIMA (0,1,0)</td>
<td>0.92</td>
<td>24</td>
<td>16.8</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>ARIMA (1,1,1)</td>
<td>0.93</td>
<td>24.2</td>
<td>17</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>Brown</td>
<td>0.89</td>
<td>24.4</td>
<td>17</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>Holt</td>
<td>0.93</td>
<td>24.1</td>
<td>16.4</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>0.70</td>
<td>----</td>
<td>106</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>R-deux</th>
<th>Bic</th>
<th>MAPE</th>
<th>Luing</th>
<th>Sig du Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA (1,1,0)</td>
<td>0.91</td>
<td>19</td>
<td>25.6</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>ARIMA (0,1,0)</td>
<td>0.92</td>
<td>18.8</td>
<td>26.7</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>ARIMA (2,1,0)</td>
<td>0.91</td>
<td>19.1</td>
<td>24.7</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>Brown</td>
<td>0.80</td>
<td>19.4</td>
<td>29.8</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Holt</td>
<td>0.92</td>
<td>18.9</td>
<td>26.5</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>0.92</td>
<td>----</td>
<td>93</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2. Comparison of model**

<table>
<thead>
<tr>
<th>Model</th>
<th>R-deux</th>
<th>Bic</th>
<th>MAPE</th>
<th>Luing</th>
<th>Sig du Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA (1,1,0)</td>
<td>0.99</td>
<td>21.7</td>
<td>5.7</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>ARIMA (0,1,0)</td>
<td>0.98</td>
<td>21.8</td>
<td>4.8</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>ARIMA (1,1,1)</td>
<td>0.99</td>
<td>21.9</td>
<td>5.3</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Brown</td>
<td>0.99</td>
<td>21.5</td>
<td>4.9</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Holt</td>
<td>0.99</td>
<td>21.6</td>
<td>4.6</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>0.87</td>
<td>----</td>
<td>15.7</td>
<td>0.12</td>
<td></td>
</tr>
</tbody>
</table>
We take 40 observation of the original series and forecast for the next ten years, then compare between models by MAPE and choose the best model. The results were as follows:

1. The \( ARIMA(1,1,1) \) for modeling the series of price values of food imports.
2. Brown’s exponential smoothing model for modeling the series of price value of food exports.
3. The \( ARIMA(1,1,1) \) for modeling the series of price values of food products.

### 3.2 Tests of Residues

We test the best model:

- Graphic residues confidence limits, \( ACF, PACF \)
- Graphic dispersion of points in parallel form residuals around zero \( ACF, PACF \)
- Ljung-Box value is significant
- If the model realizes the previous tests, we use it to forecast.

### 3.3 Forecasting

Then we use the previous models to calculate the forecast from 2011 to 2020 and the results were as follows:

*Table 3. Forecasting price value of the time series (exports, imports and production)*
4 CONCLUSION

• Through forecasting results we observed that food imports will continue to increase up to 92 % compared to food exports.
• According to forecasting results of the price value of food exports over the next 10 years, it seems that the value of export prices will increase slowly, which means that the Yemeni economy will remain as an importing economy at least during the current decade.
• We predicted to continue to decrease in the second decade of the current century (2011-2020) to 39 %.

REFERENCES

[8] Lahoussaine Baamal, Cours de Prévision par la Méthodologie de Box et Jenkins, Université Ibn Tofail, Kénitra, (2012).