"1st International Symposium on Computing in Informatics and Mathematics (ISCIM 2011)"
in Collaboration between EPOKA University and "Aleksandër Moisiu" University of Durrës
on June 2-4 2011, Tirana-Durres, ALBANIA.

An ARIMA birth number per month model for Albanian population

Eralda DHAMO[1], Llukan PUKA[2]

[1] *University of Tirana, Faculty of Natural Science, Department of Mathematics*
Email: eralda.dhamo@unitir.edu.al
[2] *University of Tirana, Faculty of Natural Science, Department of Mathematics*
Email: lpuka2001@yahoo.co.uk

ABSTRACT

Population projection models play a significant role in analyzing current demographic processes, as well as in forecasting their future development. In this study, we examine the number of birth per month, over the twenty-four year period 1985-2008 in Albania. We use data from INSTAT Albania official site. From a careful observation of the data, we see that there are months where the number of births is clearly higher, as well as others, where it seems to be lower. A decreasing trend is also evident. In our study, we use the Box-Jenkins methodology and *R* software as a programming language to consider process seasonal patterns. We model data by an ARIMA (p, d, q) stochastic process and we use it to forecast the number of births in Albania on the future. The model can be useful for governmental or nongovernmental agencies as well as insurances companies interested on birth number evolution in Albania.

Keywords*: Birth number, Forecasts, ARIMA, Seasonality.*

1. Introduction and motivation

During the 20-th century number of births per month has increased. Usually this phenomenon is considered in a pessimist way: according to statistics we are more than our ancestors. This increase in population is associated with the use of resources and threat to the ecosystem. Scientists note that the consequences of this rate growth will be reflected at the level of carbon dioxide in the atmosphere, global warming and pollution. Other sectors of the economy (employment, poverty, etc.), or the social (marriage, religion, etc.) will be affected by these changes.

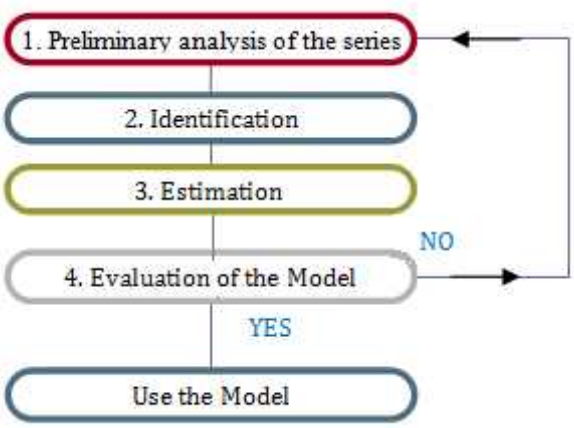
In this paper, we investigate the trend of number of births per month using the data from INSTAT Albania (from January 1985 to December 2008) as an important potential factor for the economical and social development of the country.

2. The model - ARIMA methodology

Number of births per month in Albania can be modeled as a stochastic process and consequently we can use the standard Box and Jenkins methodology (identification,

estimation, diagnostics and forecasting), [1, 2] , to generate an appropriate ARIMA(p, d, q) model for number of births per month in Albanian population.

The procedure of the ARIMA model goes through different iterative phases. Box and Jenkins propose the following methodology:



1) Preliminary analysis of the series and possible transformation

The monthly data were collected for the period January 1, 1985 to December 31, 2008 in total 288 observations (Fig.1).

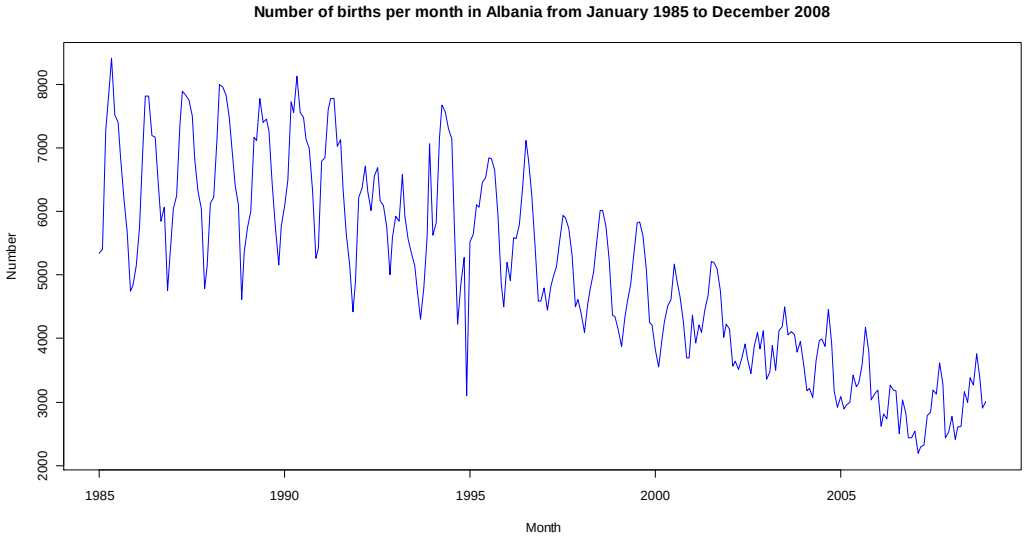


Figure 1. Number of births per month in Albania from January 1985 to December 2008

a) Detecting stationarity

For an ARIMA model the series needs to be stationary; we can transform a time series in a stationary time series by differencing until it becomes stationary. The plot and the autocorrelogram (ACF) help to investigate and fix the differencing level. The best value of *d* is the one that gives rise to a rapid decrease of the ACF towards zero. As seen from the following graphs, our time series is not stationary.

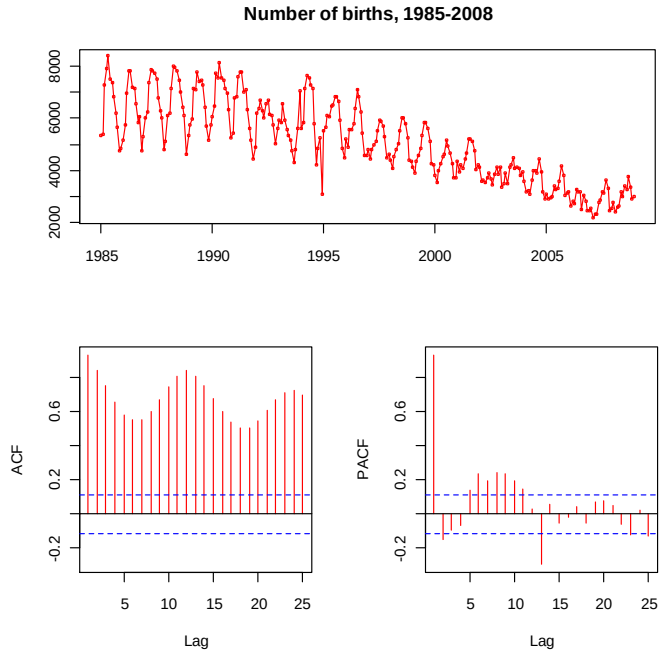


Figure 2. ACF and PACF of number of birth per month

Autocorrelation and Partial Autocorrelation values are significantly important (Figure 2).

b) Detecting seasonality

Many time series contain a seasonal phenomenon that repeats itself after a regular period of time. This phenomenon is evident in our time series. Seasonality, or periodicity, can usually be assessed from an *autocorrelation plo*t, a *seasonal subseries plot*, or a *spectral plot*.

As a first step to see for seasonality, we consider the evolution of the average monthly number of births across 24 years (Figure 3).

It seems, the averages are near to each other, except the last months of the year where the average number of birth seems to be lower.

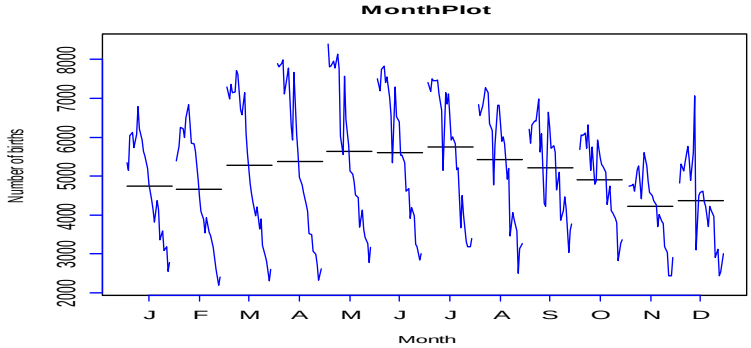


Figure 3. Number of birth for each month of 24 years

c) Differencing for stationarity and seasonality

Figure 1, show clearly that our time series, of births number per month, has a trend. By first differencing the data, $\Delta y_t = y_t - y_{t-1}$, the graph of the new series is shown in Figure 5.

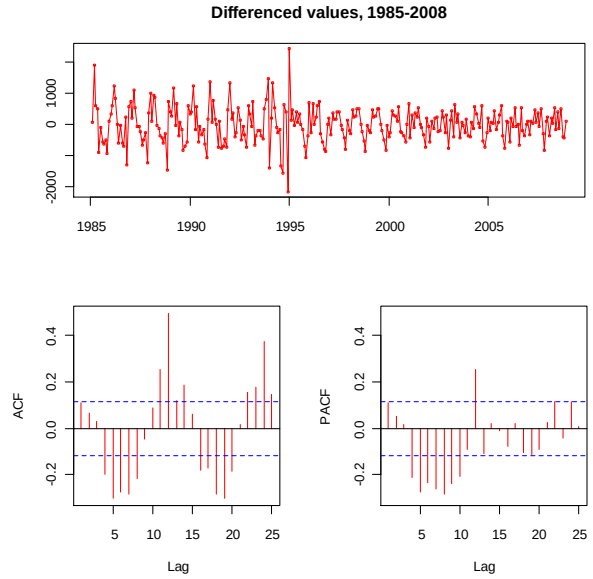


Figure 5. ACF and PACF of number of birth per month for the differenced time series

Later, we inspect the autocorrelation function of the first-differenced data for significant piks at the seasonal frequencies to discover the possible presence or absence of seasonality in the original data. In fact, the reciprocal autocorrelation values are relatively reduced, respectively to the same lag in two series. The same result can be seen looking the *lag plot* of the original data and the differenced data (Figure 6 and Figure 7).

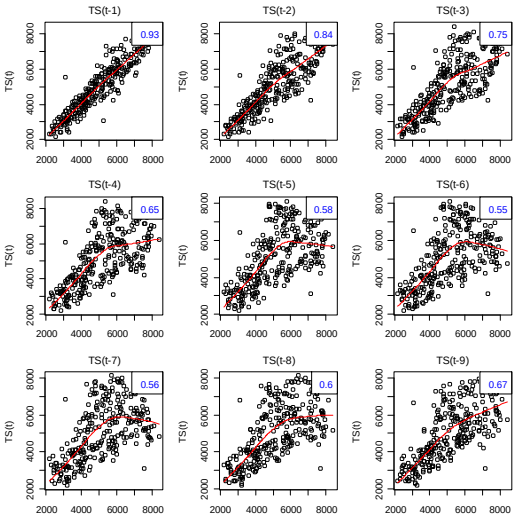


Figure 6. Lag plot of the original data

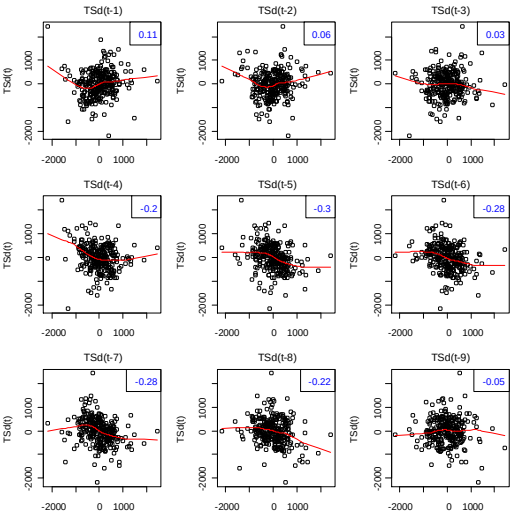


Figure 7. Lag plot of the differenced data

2) Estimation

The trend in our data was fitted by a linear regression line. The results are shown in Figure 8.

The regress equation line is: $B = 6864.724 - 18.758t$, B is the number of birth per month, t is the time. Multiple R-squared: *0.758*, Adjusted R-squared: *0.7569.*

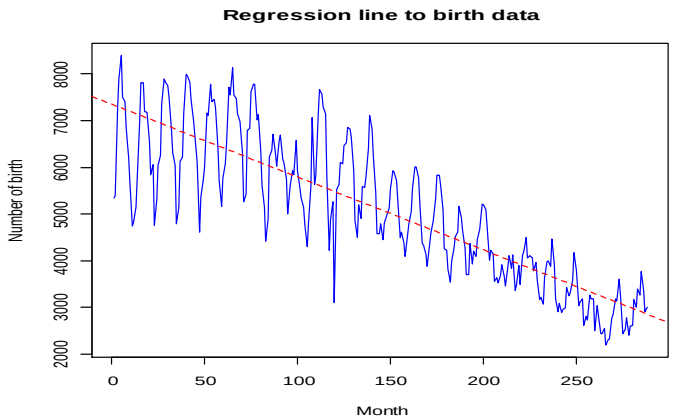


Figure 8. Linear Regression of time series

At the evaluation phase, the aim is to detect seasonality, if it exists, and to identify the order for the seasonal autoregressive and seasonal moving average terms. By inspecting the shape of ACF and PACF we predict that the model can be an ARIMA (p, d, q) and considering seasonality, it can be an ARIMA (p, d, q) $(P, D, Q)_m$ model, m is the seasonal frequency.

Based on the model of Box and Jenkins (1970), the seasonal autoregressive integrated moving average model is given by:

$$f(X)=b_1 \otimes X_1 \oplus b_2 \otimes X_2 \oplus .. \oplus b_n \otimes X_n \oplus c = w.X \oplus c. \qquad (1)$$

Where,

s = seasonal lag,
coefficient for AR process,
= coefficient for seasonal AR process,
= coefficient for MA process,
= coefficient for seasonal MA process.

B is the backward shift operator, ${}_s^D = (1 - B^s)^D$ and ${}^d = (1- B)^d$, w_t is an uncorrelated random variable with mean zero and constant variance, [3].

In R, we get a SARIMA (Seasonal Autoregressive Integrated Moving Average) model.

The proposed model is:

ARIMA(2,1,2)(1,0,1)[12]

and the coefficients of the model are:

s = 12, = 0.3319, = 0.2136, =0.9552, =-0.5477, = -0.4098, =-0.6757

The information criteria values and error measurements are:

AIC= 4299.67, AICc= 4300.08, BIC= 4325.29

ME= -31.60, RMSE= 413.46, MAE= 287.99, MPE= -1.25, MAPE= 5.95, MASE= 0.67

We selected the one with the lowest values of error measurements or information criteria between the proposed models, [4].

3) Diagnostic checking

One way to check if the model is satisfactory is to analyze the residuals. The nature of birth data seems to be seasonal. Starting with the normality test for the residuals, we investigate the residuals histogram. In addition, a normal probability plot or a Q-Q plot is done to identify departures from normality. We also inspect the sample autocorrelations of the residuals and check at the correlation structure of the residuals by ploting the autocorrelation versus h and the error bounds of $\pm 2/\sqrt{n}$, [5].

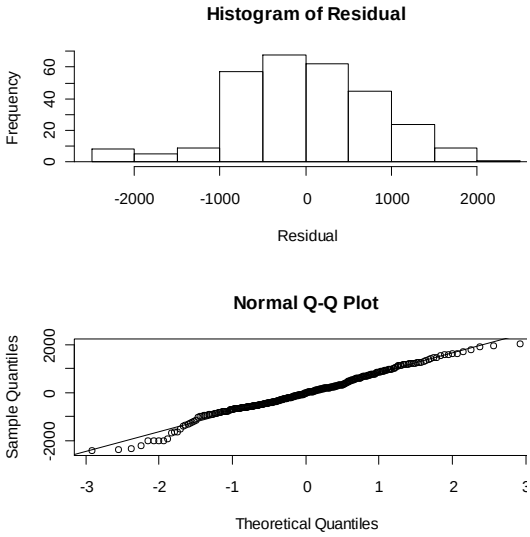


Figure 9.a Residuals of the regression model to birth data 1985-2008

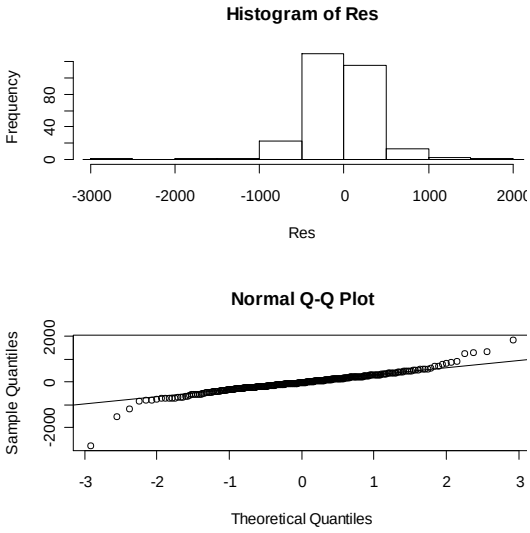


Figure 9.b Residuals of the ARIMA model to birth data 1985-20

We see, the residuals are nearly close to normality, except for a few extreme values in the tails. The ARIMA model seems to fit better our data.

In addition to autocorrelations plot, we perform a general test that takes into consideration the magnitudes of all autocorrelations. As an example, it may be the case that, individually, each value of autocorrelation at lag -*h* is small in magnitude, say, each one is just slightly less that $2/\sqrt{n}$ in magnitude, but, in group, the values are large. The test is done by Ljung–Box–Pierce Q-statistic:

$$Q = n(n+2)\sum_{h=1}^{H} \frac{\circ^2(h)}{n-h} \qquad (2)$$

The value H in (2) is chosen somewhat arbitrarily, typically, $H = 20$. Under the null hypothesis of model adequacy, asymptotically , $Q \sim ¡^2_{1-\sim, H-p-q}$. Thus, we would reject the null hypothesis at level α if the value of Q exceeds the $(1-\alpha)$-quantile of the $¡^2_{H-p-q}$ distribution, [5].

Also the *ACF* and *p-values* of the Ljung-Box statistic for the residuals of the ARIMA model, are within the confidence interval, so we have no evidence to reject the model.

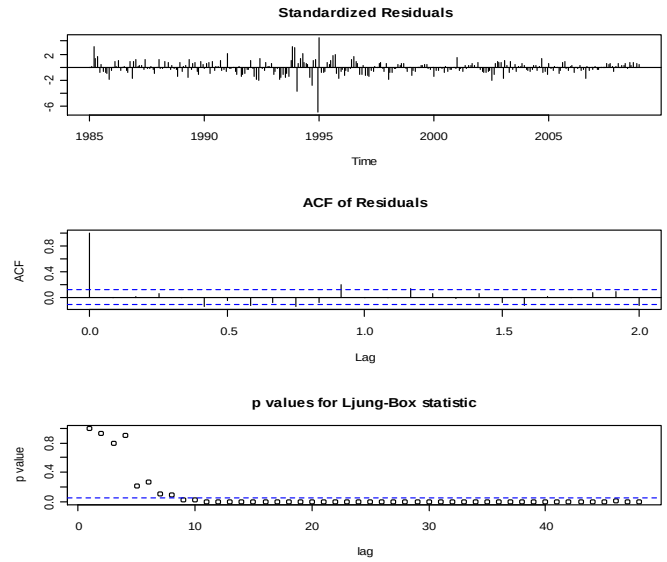


Figure 10. Ljung-Box statistic and ACF of residuals

Looking at the lag plot of the residuals of the ARIMA model, we see that the autocorrelations and partial autocorrelations values are statistically equal to zero.

4) Forecasting

The main purpose of estimating the birth models is to use them as input to a demographic model that can be used for forecasting the population of Albania.

In our SARIMA model we consider the seasonal patterns of birth data and, using some accuracy measurements, we choose the one with the best information criteria or lowest error measurements. In Figure 12, we show the forecasted values and the confidence intervals acording to three periods: 1985-2008 (288 observations), 1990-2008 (228 observations) and 2000-2008 (96 observations).

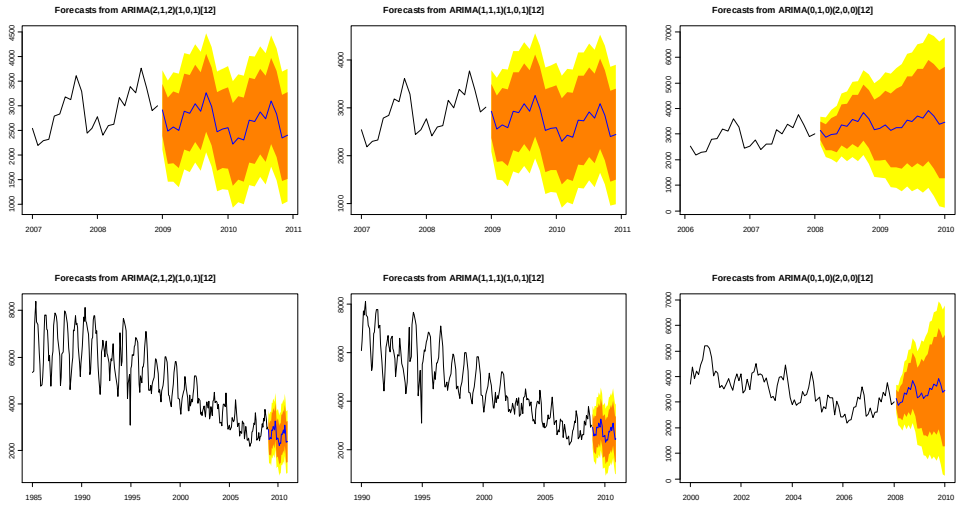


Figure 12. Forecasts from ARIMA model

CONCLUSION

The model presented in this paper and the preceding one, [4], represents only a first step on the investigation that could be done using ARIMA models to the demographic data concerning the Albanian population. As a useful methodology for forecast, Box and Jenkins methodology require at the same time skills and experience.

In this paper we show the usefulness of the Box and Jenkins methodology to study and forecast the evolution of birth number in Albania from January 1985 to December 2008. This number has decreased significantly during all this period. But at the same time the evolution of this indicator in the last part of the period show that the decreasing trend is slower: during the period from January 2000 to December 2008, the birth number process seems to be stationary and the forecasting process can be more useful. Further studies and other data, related to birth number or social and economic factors, are important for forecasting the development of vital models in Albania.

REFERENCES

[1] Box, G. E. P. and Jenkins, G. (1976) *Time Series Analysis: Forecasting and Control*, Oakland, CA: Holden-Day, (revised edn, 1976).

[2] Hamilton, J. D. (1994). *Time Series Analysis*, Princeton, NJ: Princeton Univ. Press

[3] Shumway H. R. and Stoffer S. D. (2006) *Time Series Analysis and Its Applications With R examples.* Springer Second edition, ISBN: 978-0-387-75958-6

[4] Dhamo, E. and Puka, Ll. (2010) Using the R-package to forecast time series: ARIMA models and Application.*INTERNATIONAL CONFERENCE Economic & Social Challenges and Problems 2010 Facing Impact of Global Crisis*, Tirana, Albania.

[5] Ljung G. M. and Box G. E. P. (1978) On a Measure of a Lack of Fit in Time Series Models *Biometrika* **65**: 297–303.

[6] Dhamo, E. and Puka, Ll. (2010) Një vështrim mbi disa kode të paketës forecast në R, *Buletini i Shkencave Natyrore*, Tiranë, **10**: 5-18.