Combining Econometric and Time Series Models to Project Albanian Population (Projection for Years 2015- 2023)

Eralda Dhamo (Gjika), Oriana Zaçaj, Edionada Gjika

***Abstract*— Albania is a small country in the Balkan region but with a key geographical position in the demographic movements in the region and in Europe. In the recent years the political and economic changes in Europe have significantly affected demographic indicators of the country. The population projection is one of the important issues in this moment. The number of births is decreasing, by other hand number of emigrants is increasing rapidly. At this moment population forecasting techniques are seen with interest. In this work we study many socioeconomic variables that may affect the total number of population in Albania. We propose two models which may be used to projected population in the upcoming years. The models combine multiple regression and time series models. The most important variables in the model were selected based on many indicators of the model (AIC, MAPE, MSE etc.) and graphical tests. The final model was selected based on several measures of accuracy and bias, and formal statistical tests of differences in errors by technique.**

***Index Terms*— population projection, time series, regression, socioeconomics, accuracy**

I. INTRODUCTION

The first round of Census in Albania was realized in 1989, then again in 2001 and 2011. 1990s were the years in which a significant movement of emigrants and immigrants were observed. These movements continued in 2000 and involved the country by a significant reduction of the population. Low fertility rate was observed in these year's due to reduction of number of births. The reduction of number of births was affected by economic and social factors. Based on INSTAT Albania publications the total population had a peak of 3,182 thousand in 1989, it declined to 3,069 thousand in 2001 and 2,907 in 2011, corresponding to an annual decrease of 0.3 and 0.8%, respectively.

After the year 2010 the global crisis touched Albania. Its effect was observed not only in the economic situation of the country but also in aggravation of demographic situation ([10]). Visa liberalization was also an important factor that influenced further movements of population to European countries. These movements have disturbed the government and other institutions which began to take precautions in education, employment, civil status etc.

The classic cohort component method is widely used by official organisms for population projections and is defined by the following equation:

$$P_t = B_t - D_t + M_t \qquad (1)$$

where P_t denotes the population at time t, B_t denotes births at time t, D_t denotes deaths at time t and M_t denotes the net migration in the period ($t-1$, t). This equation is completed by other equations based on fertility rate, mortality rate and net migration rate by age and sex. In this work we have worked on other models which take into consideration not only demographic variables.

Many forecasting methodologies are projection techniques that were developed prior to 1960. S. K. Smith (1997) [1], makes a detailed analysis of the techniques proposed in years on the population projection. He analyzes in his paper the simplicity versus complexity, evidence regarding forecast accuracy, the costs and benefits of disaggregation, and the potential benefits of combining projections.

The term *projection* is typically defined by demographers as the numerical outcome of a set of techniques and assumptions regarding future trends, whereas a *forecast* is the specific projection the analyst believes is most likely to provide an accurate prediction of future population change. Many authors propose a projection horizons ranging from five to twenty years to achieve good projections, ([2], [8], [12]). Other studies ([3], [6]) demonstrate that in some circumstances a particular simple technique may tend to perform better than less accurate techniques or complex techniques [7]. Issues such as the costs and benefits in projection models are important when we evaluate the prediction model.

Ahlburg (1995) [5], called for more research on the benefits of combining projections from different models to create population forecasts. Makridakis *et al*. ([4], [6]) suggest that combining forecasts is more likely to improve accuracy when no individual forecasting method has been established.

Our study is based on these guidelines and intends to show that the combination of multiple regression techniques and time series models can offer satisfactory population projections for the upcoming years. We analyze a combination of multiple regression with ARIMA models using a wide range of explanatory variables. There are in total 11 explanatory variables from other disciplines that are not typically involved in formal population forecasting efforts and which can have significant effects on population change. The relationships between the population and its demographic characteristics, social and economic phenomena is a strong relationship that should be considered (Sanderson WC, 1998). Data are taken form INSTAT Albania and are registered yearly from 1990 to 2013. We have categorized the variables into three main groups: Demographic, Social and

Revised Version Manuscript Received on November 21, 2015.

Eralda Dhamo (Gjika), Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Tirana, Albania.

Oriana Zaçaj, Department of Mathematics, Faculty of Mathematics and Physic Engineering, Polytechnic University Tirana, Albania.

Edionada Gjika, Department of Biotechnology, Faculty of Natural Science, University of Tirana, Tirana, Albania.

Economic variables.

***Demographic variables*:** Number of births, Number of deaths, Number of immigrants

Social variables: Number of divorces, Number of abortions , Number of marriages, Average age of marriage (for women and man)

***Economic variables*:** Unemployment rate, Average salary, Number of graduates.

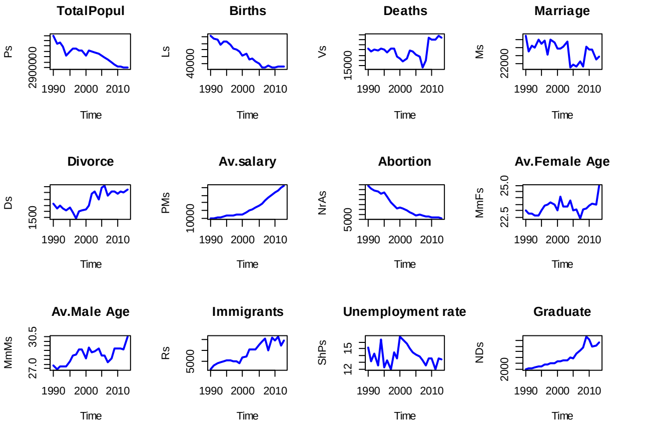


Fig.1. Time series of economic social and demographic variables

Fig.1 show the time series of 11 explanatory variables and the total population for years 1990 to 2013. As can be clearly seen some variables indicate ascending or descending trend while others have the presence casualty. Total population time series also shows a clear descending trend. We expect that the final model include the variables with visible trend in their behavior.

II. ECONOMETRIC MODELS ADOPTED

Three models have been chosen for comparative analysis of the results. We have proposed three models using the observation of 11 explanatory variables, yearly from 1990 to 2013. The first model we study is a linear regression model:

$$Y=\beta_0+\beta_1*X1+\ldots+\beta n*Xn+\varepsilon \qquad (2)$$

It assumes linear dependency of total population and explanatory variables. The second model taken under consideration is a log-linear regression model which assumes log-linear dependency:

$$ln(Y)=\beta_0+\beta_1*ln(X_1)+\ldots+\beta_n*ln(X_n)+\varepsilon \qquad (3)$$

Time series is a set of observations ordered in time. This analysis deals with observations that are collected over equally spaced, discrete time intervals so the third model considered is a SARIMA model:

$$\Phi_p(B^s)\phi(B)\nabla_s^D\nabla^d X_t=\alpha+\Theta_Q(B^s)\theta(B)w_t \qquad (4)$$

where,

s = seasonal lag,

ϕ = coefficient for AR process,

Φ = coefficient for seasonal AR process,

θ = coefficient for MA process,

Θ = coefficient for seasonal MA process.

B is the backward shift operator,$\nabla_s{}^D= (1 – B^s)^D$ and $\nabla^d = (1-B)^d$, w_t is an uncorrelated random variable with mean zero and constant variance.

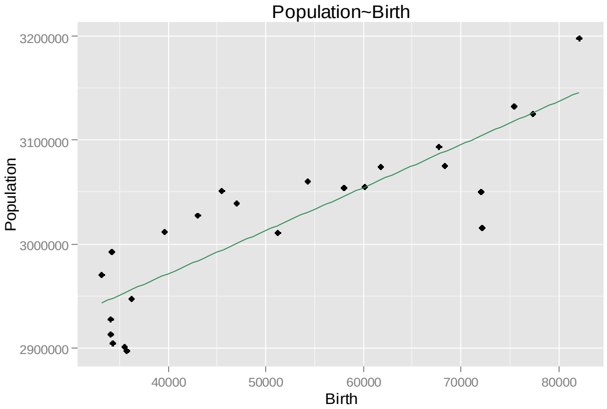
As can be observed the first two models use the explanatory variables to project the total population and the third model only uses the time series of total population. The first two models are selected based on: Akaike information criterion (AIC), Scatter plot, Accuracy indicators (MSE,MAPE, MAD etc.) , R-square, T-test (ANOVA) etc.

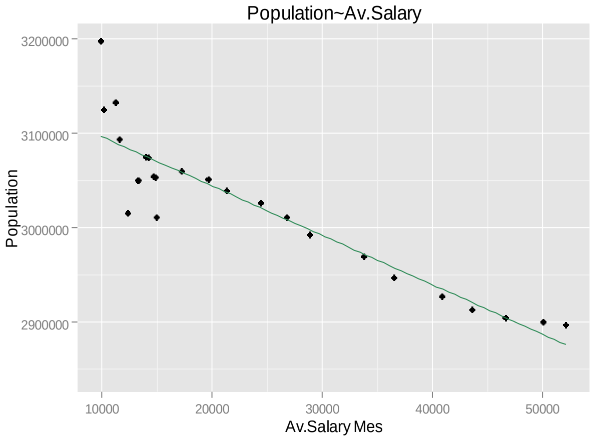
Table 1 Pearson correlation coefficient between total population and explanatory variables

Corr Coef	Birth	Death	Div.	Marr	Av.Sal	Abort
Pop.	**0.86**	-0.38	-0.63	0.47	**-0.91**	**0.84**
	F.Age	**M.Age**	**Immig**	**Unempl**	**Grad.**	
Pop.	-0.34	-0.61	-0.86	0.23	**-0.9**	

Before we start building regression models we consider the dependence between the number of total population and the explanatory variables calculating the Pearson correlation coefficient. As can be seen from table 1, number of population seems to have high dependency with number of births (*correlation=0.86*), average salary (*correlation=-0.91*), abortion (correlation=0.84) and number of graduated (*correlation=-0.90*). These variables are those we mostly expect to be in the final regression model.

Fig. 2 show the scatter plot for some explanatory variables and number of total populations. The relation between these variables was shown also by the value of Pearson correlation coefficient.





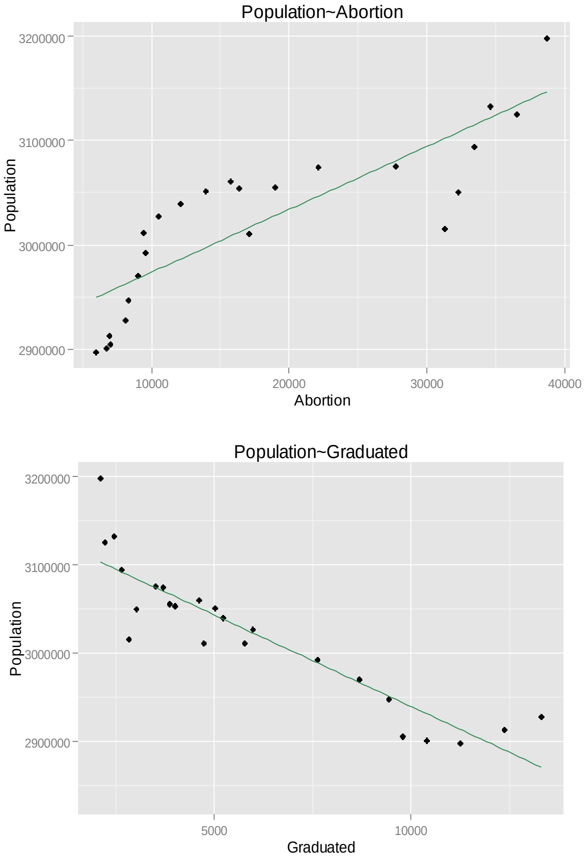


Fig.2. Scatter plot of total population and explanatory variables

Each of the regression model launched including 11 variables and eliminating them one by one based on their importance in total population. Relying on lower value of Akaike Information Criteria and higher value of R-square value, we select as the “best” models the following:

Linear model:

*Population=2.457e+06+3.194e+01*Divorce-4.849*Av.Salary+3.823*Abortion+1.831e+04*Av.men.Age*

Table 1 Analysis of Variance for Linear model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Divorce	1	5.8168e+10	5.8168e+10	76.1767	4.491e-08	***
Average Sal	1	6.4268e+10	6.4268e+10	84.1645	2.071e-08	***
Abortion	1	5.8376e+09	5.8376e+09	7.6448	0.01233	*
Av.men Age	1	2.1187e+09	2.1187e+09	2.7746	0.11217	
Residuals	19	1.4508e+10	7.6360e+08			

Multiple R-squared: 0.8999, Adjusted R-squared: 0.8788
F-statistic: 42.69 on 4 and 19 DF, p-value: 3.056e-09
AIC=495.28

It can be noticed that the effect of *Average salary* and *number of divorces* on total population is stronger than *number of abortions* and *average age of men* in marriages. This can be explained as a consequence of economic and social welfare of families to reduce the number of children and consequently the number of births in the country. R-squared value is considerable *0.8999*.

Log-Linear model:

*Population=exp(15.19229+0.03816*log(Divorce)-0.03722*log(Marriage)-0.05897*log(Av.salary)+0.17101*log(Av.men.Age)-0.01275*log(Immigrants)-0.02588*log(Unemploy.rate))*

Table 2 Analysis of Variance for Log-Linear model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
log(Divorce)	1	0.0058743	0.0058743	95.0078	2.250e-08	***
log(Marriage)	1	0.0005227	0.0005227	8.4544	0.009803	**
log(Average sal)	1	0.0079190	0.0079190	128.0775	2.453e-09	***
log(Av.men Age)	1	0.0001277	0.0001277	2.0657	0.168801	
log(Immigrants)	1	0.0002667	0.0002667	4.3127	0.053304	.
log(Unemploy.rate)	1	0.0000984	0.0000984	1.5908	0.224239	
Residuals	17	0.0010511	0.0000618			

Multiple R-squared: 0.9337, Adjusted R-squared: 0.9103
F-statistic: 39.92 on 6 and 17 DF, p-value: 4.229e-09
AIC=-226.86

In both models we have observed that even without the variable of *average age in marriage* for men the model has the same value of accuracy measures. Multiple R-squared was not affected in value by the removal of the variable. So, this variable may be optional to be considered or not.

III. TIME SERIES FORECAST OF IMPORTANT EXPLANATORY VARIABLES

Box and Jenkins technique has been shown to give relatively accurate forecasts although it goes through simple steps it is simpler and more effective than other contemporary models. In the forecasting step for each of the predictive variables we have used the Box and Jenkins iterative three steps procedure in R software [11].

Model selection, parameter estimation and model checking was conducted for time series of: number of marriages, number of divorces, number of abortions, average salary, average age of men in marriages, number of immigrants, unemployment rate.

The models obtained by this methodology in R software were used to predict the values for 10 upcoming years. In lack of space Fig. 3 show the ARIMA model for two of the explanatory variables (Average salary and number of divorces).

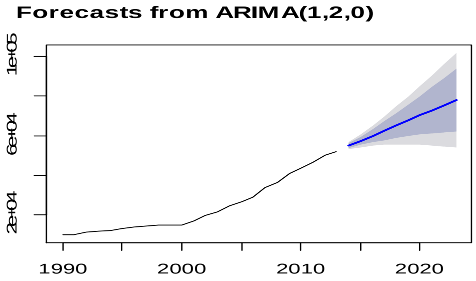


Fig.3.a Forecasted values and confidence intervals for average salary

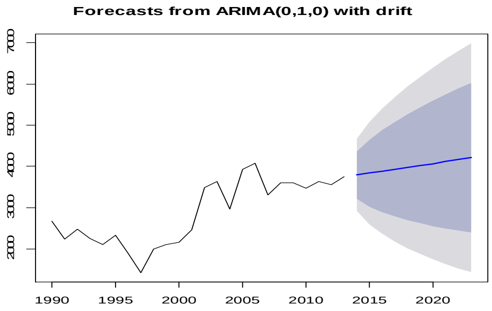


Fig.3.b Forecasted values and confidence intervals for number of divorces

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

For the number of total number of population the time series model was an ARIMA(0,1,0) model with drift. Fig. 4 show the forecasted values and confidence intervals 85% and 95%.

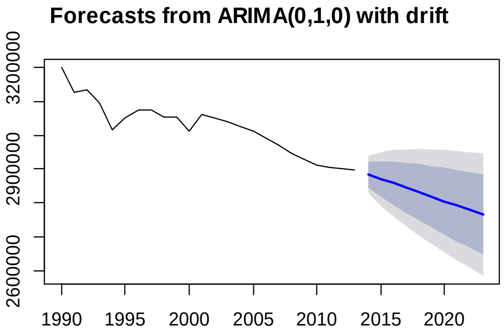


Fig.4. Forecasted values and confidence intervals for number of total population

The time series of total population shows a clear decreasing trend, so the ARIMA model maintains the trend and offers a justify decreasing trend.

A reasonable degree of accuracy has been achieved through consistency of data. We have compared the three projection models (linear regression; log-linear regression and ARIMA model) based on accuracy measure (AIC, R-square, graphical test) and we may say that the log-linear model is a good choice for predicting the total population in upcoming years. Both models show a descending trend of total population. This can be seen also from graphical view of the forecasted values, Fig.5.

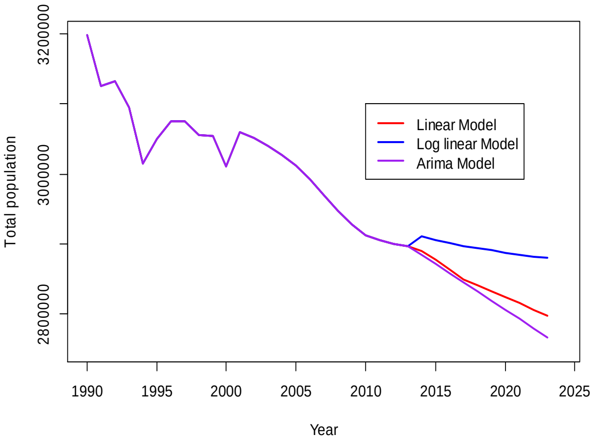


Fig.5. Total population data and forecasted values by the models

IV. Conclusion

Previous studies of population projection in Albania has made no effort to build models based on the combination of several known prediction techniques. This study attempts to provide an overview of effectiveness of combining techniques to better projection of total population in Albania. We built three models, two of them are regression models with social, demographic and economics explanatory factors. We observed that the regression methods produce more accurate projections than the extrapolation techniques (ARIMA). Although the population redistribution process may experience different patterns and may be affected by various factors it also affect the regression projection accuracy. So, a change on explanatory variables can have greater impact on number of total population and then a new regression model should be considered. The population projection is of interest for governmental institutions, INSTAT Albania or academic units to build strategy to solve potential problems of decreasing total number of population. The results obtained in this work are based on real data published by INSTAT Albania.

References

1. Smith S. K. (1997), Further thoughts on simplicity and complexity in population projection models ,International Journal of Forecasting 13 (1997) 557-565 , Bureau of Economic and Business Research.
2. Long J.F. (1995), Complexity. accuracy, and utility of official population projections. Mathematical Population Studies 5. 203-216
3. Mahmoud, E.. (1984). Accuracy in forecasting: A survey. Journal of Forecasting 3. 139-159.
4. Makridakis. S., Hibon. M.. 1979. Accuracy of forecasting: An empirical investigation. Journal of the Royal Statistical Society A 142.97-145,
5. Ahlburg, D.A., 1995. Simple versus complex models: Evaluation. accuracy and combining. Mathematical Population Studies 5, 281-20
6. Makridakis. S.. Winkler, R.L.. 1983. Averages of forecasts: Some empirical results. Management Science 29, 987-996.
7. Sanderson WC. (1998) “Knowledge Can Improve Forecasts: A Review of Selected Socioeconomic Population Projection Models” Population and Development Review. 1998;24 (Suppl.):88–117.
8. Meng Yi., Wang Zizheng., Sia Wai Leng (2011): Study of Mathematical Models for Population Projection;. Singapore 259-978
9. McLennan A.(2006): Population Growth in a Closed System, Oxford
10. Xhaja B., Dhamo E., (2011): Population projections: methodological issues and challenges in Albanian Population, 6TH ANNUAL MEETING OF INSTITUTE ALB-SHKENCA, Prishtina, 1-4 September,2011,Kosovo
11. Shumway H. R. & Stoffer S. D. (2006): Time Series Analysis and Its Applications With R examples. Springer Second edition, ISBN: 978-0-387-75958-6
12. Meng Yi, Wang Zizheng, Sia Wai Leng (2011): Study of Mathematical Models for Population Projection;. Singapore 259 978.

Eralda Dhamo (Gjika) Received the University degree in Mathematics from University of Tirana, Faculty of Natural Science, Department of Mathematics’ in 2006; the Master degree in Probability and Statistics Application at Department of Mathematics, University of Tirana in 2009 and the PhD title, June 2014, Dep. Of Applied Mathematics, Faculty of Natural Science UT. Research interests include (but are not limited) on: time series analysis, models, forecast, similarity techniques and dimensionality reduction.

Oriana Zaçaj Received the University degree in Mathematics from University of Tirana, Faculty of Natural Science, Department of Mathematics’ in 2000; PhD title, November 2015, Department of Mathematics, Faculty of Mathematics and Physic Engineering, Polytechnic University Research interests include (but are not limited) on: actuarial science, demography, time series analysis, models, forecast.

Edionada Gjika Received the Bachelor degree in Biology, (2011) and MSc. Degree in Biotechnology in 2013, Faculty of Natural Science,. Research interests include (but are not limited) on: demography and population studies.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.