Application of Stochastic Regression Models: ARIMA (p, d, q)- HW Algorithm Approach for Human Population Forecasting

Muhammad Ilyas*, Shaheen Abbas, S. Akhter Raza *****
Wajid Ali****, S. Masood Raza*******

*Department of Mathematics, Government College University Hyderabad, Sindh, Pakistan
** Laboratory for Applied Mathematics and Data Analysis (LAMDA)
Mathematical Sciences Research Centre, Federal Urdu University of Arts, Sciences and Technology, Karachi, Pakistan
***Department of Computer Science Federal Urdu University of Arts, Sciences and Technology, Karachi, Pakistan.
Department of Mathematical Sciences Karakoram International University Gilgit. Pakistan
*****Department of Physics Federal Urdu University of Arts, Sciences and Technology, Karachi, Pakistan

Abstract- Population forecasting plays a constructive role in altering population policy and promoting the development of social, economic, natural disasters (rainfalls, cyclone, flooding and earthquakes) and pandemic of infectious disease such as seasonal dengue, malaria and novel Coronavirus (2019-nCoV/SARS-CoV-2) endeavors. the city of Karachi has challenged frequent problems due to uncontrolled population dynamics, morphological pattern, their socio-health, and climate impacts on seasonal disease.In this research, we have found a suitable stochastic Auto Regressive Integrating Moving Average (ARIMA (*p, d, q*)) model by using diagnostic checking to overcome the aberrant of the Karachi city annual population data intervals from 1951 to 2015 respectively.The population data of Karachi city in 2015 is also verified , The results analysis are shown that the actual fitting outcome of the model is appropriate, Finally, comparing the accurateness of the ARIMA model, Holt-Winters (Non-Seasonal) algorithm, linear - exponential trends on the base of Normality check MAPE, Geary's α and Sample Kurtosis b_2 statistic test for population forecasting of the time intervals are 15 years and the forecast prospect range from 5 to 15 years. The results show that the ARIMA (1,2,1) seem to similar Holt-winters forecasting can extend appropriate results and is fitted for population forecasting.In the future, these results will be more helpful for investigate the epidemiological trend of the occurrence and pandemic of COVID-19.In addition, the stochastic analysis approach will be employed to increase the perception of data analysis, thus providing scientists more efficiently to the urban environment in the Karachi region.

Index Terms- ARIMA (*p, d, q*)
Holt-Winters algorithm, Diagnostic checking, autocorrelation coefficients (ACF), partial autocorrelation coefficients (PACF).

I. INTRODUCTION

The Stochastic Autoregressive integrated moving average (ARIMA) model is used to model the population data [1-3]. The order of an ARIMA model is frequently used by notation ARIMA (p, d, q). The pure ARIMA is mathematically express as,

$$W_t = \frac{\mu+\theta(B)}{\emptyset(B)} a_t \qquad (1)$$

$W_t$is indicate the data time series of observed data (Yt) , at shown the independent random error , μ is the expected value (mean) in B is the back shift operator that is $BX_t = X_{t-1}$ and $\emptyset(B)$is the auto regressive operator which is presented as a mathematical expressions in the back shift operator $\emptyset(B) = 1 - \emptyset_1 B - \cdots - \emptyset_p B^p$, $\theta(B)$ is the moving average operator denoted[4] as an expression in the back-shift operator [5, 6]. The ARIMA notation with constant term as inscribed $\emptyset(B)(W_t - \mu) = \theta(B)a_t$, else $\emptyset(B)W_t = constant + \theta(B)a_t$, here,$constant = \emptyset(B)\mu = \mu - \emptyset_1\mu - \emptyset_2\mu - \cdots - \emptyset_p\mu$.

The general ARIMA model with data time series also called ARIMA is written as

$$W_t = \mu + \sum_i \frac{\omega_i B}{\delta_i B} B^{k_i} X_{i,t} + \frac{\theta(B)}{\emptyset(B)} a_t \qquad (2)$$

Where,$X_{i,t}$is ith time series or a difference of the time series at time t, k_i ist the time lag or (delay) for the effect if the ith time series, $\omega_i(B)$ is the numerator time delay for the effect other ith time series and $\delta_i(B)$ is denominator of the transfer function for ith time series [7]. The model can be written more closely is

$$W_t = \mu + \sum_i \Psi_i B \; X_{i,t} + n_t \qquad (3)$$

Where $\Psi_i B$ is the transfer function time series modeled as a ratio of ω *and* δ expression: $\Psi_i(B) = (\omega_i(B)/\,\delta_i(B))B^{k_i}$ is the noise series: $n_t = (\theta(B)/\,\phi(B))\,a_t$, this is the factor Modeled

$$W_t = \mu + \frac{\theta_1(\mathrm{B})\theta_2(\mathrm{B})}{\emptyset_1(B)\emptyset_2(B)} a_t \qquad (4)$$

Where $\emptyset_1(B)\emptyset_2(B) = \phi(B)\theta_1(\mathrm{B})\theta_2(\mathrm{B}) = \theta(B), \phi$ and θ is indicated autoregressive and moving average orders of the

ARIMA model [8, 9,11]. The first part, it has an integrated (I) and the component (d) which represents the order of differencing to be performed on the series to conquer stationarity. The second part of ARIMA consists of an Auto Regressive Moving Average model for the series provide the stationary through differentiation. The Auto Regressive Moving Average (ARMA) part is auxiliary decomposed into Auto Regressive and Moving Average components [11]. The Auto Regressive (AR) parts are confined the correlation among the present values of the time series data and some of its past values, AR (1) and AR(2) denotes that the existing observation is correlated with instant past values at time [11]. The moving Average (MA) parts are represent the duration of the influence of a random shock, MA (1) and MA(2) indicates that a shock on the value of the data series at time t is correlated with the astonish at time t = 1 and t = 2. The Auto Correlation Functions (ACF) and Partial Autocorrelation Functions (PACF) are applied to estimate the values of *p* and *q*. The ACF and PACF are constructed using actual, difference and transformed data. The Box – Jenkins transformation is applied to transform the population data in this research. For the selection of the adequate model validation statistics are suggested specifically, diagnostic checking test and the Bayesian Information Criterion (BIC) and Akaike information criterion (AIC) in the view of adequate forecasted error tests etc. These statistics are computed for each applicant model and the model having smallest values of errors with BIC is suggested as an adequate ARIMA model assuming that it is to be closest to the unknown certainty by which the series is generated [12] After selecting an adequate ARIMA (*p, d, q*) stochastic process and we used to estimate the number population in Karachi on the future along with confidence interval. The fitted population is plotted on the same graph are depicts to examine the model adequacy.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

Population protuberance for several developing countries including Pakistan could be quite a challenging task for the demographers typically due to lack of availability of enough reliable data. The intense population trends in the city of Karachi of the yearly time interval from 1951 to 2015 are discussed in this research. The population census data records from 1951 to 1998 have been taken from Federal Bureau of Statistics Pakistan, the missing Data values have been found by using Interpolation method. This paper is used ten AR(2) stochastic ARIMA models to examine the dynamic of population fluctuations of the Karachi city in per year. subsequently, we determined the accuracy of the nonseason Holt-Winters algorithm, as well as trend extrapolation techniques and ARIMA time series models for population forecasting. As a basis of normality check test, we also evaluated the accuracy of the ARIMA and Holt-winters algorithm for the forecast horizon varied from 5 to 15 years.

1.1. The Stochastic ARIMA (p, d, q) Modelling

The Box-Jenkins Auto regressive modeling procedure is involved an iterative three-stage process of model selection, parameter estimation and model diagnostic checking [2,13,15]. The annual population time series in the city of Karachi can be modeled as a stochastic process, The procedures of the ARIMA (*p, d,* q) model is expressed the different iterative phases [9,14, 16]. Box and Jenkins propose the ARIMA methodology are involved the four main steps such as: Model identification, Model Estimation, Model Diagnostic Checking, fitting, and forecasting depict in figure 1,

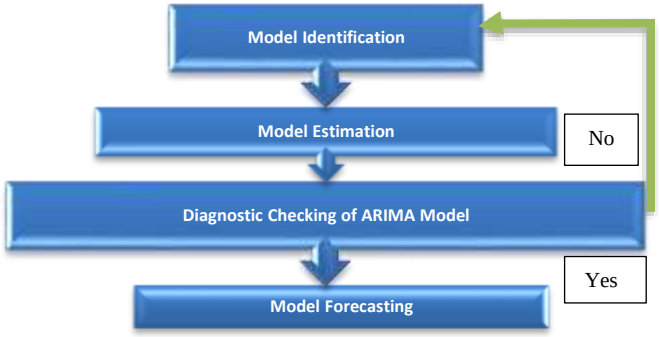


Figure 1: The Methodology of an ARIMA (*p, d, q*).

2.1.1 Model Identification

The primary step in the process of modeling is to check for the stationarity of the time series data. This is done by observing the graph of the data or autocorrelation and the partial autocorrelation functions [9, 11].

2.1.2 Model Estimation

The estimation of the ARIMA model's parameters is acquired from the values of the fitted autocorrelation of the differenced of time series data values. These initial values are used for fitting of least square estimates. In performing it is declared the not all parameters in the models are significant. In the mathematical appearance, as[9, 11].

$$\left|\frac{parametrs}{1.96 \times standard\ error}\right| > 1 \qquad (5)$$

The ratio is suggested annoying a model in which some of the parameters are set to zero [22]. The normality test for the residuals, are investigate the ACF and PACF residuals plots, residuals histogram, residual Probability plots and residual Quintile plots etc. As well as the plot of the autocorrelation and partial autocorrelation functions of the residuals from the tentatively identified ARIMA (*p, d , q*) models are depicted in result and discussion.

2.1.3 Diagnostic Checking

The diagnostic checking is a procedure that is used to ensure residuals of the fitted ARIMA models. The residual is ought to fulfill the model assumption of being independent and normally distributed. If these assumptions are not fulfilled, then another model is chosen for the time series [2, 13, 17, 18]. It is concerned through the diagnostic checking test for the fitness of ARIMA model. As, the Residual plots of the ACF and PACF, it can appear that all points are randomly distributed, and it is completed that there is an irregular pattern of the values which is indicated that the model is adequate [2,13, 14]. The autocorrelations of the individual residuals are very small values and verify at the correlation structure of the residuals because of graphs the autocorrelation of the significance error bounds of ±**2**/√**n**, [6, 13].

There are various diagnostic criteria are used to evaluate the accuracy of each model. The R-Square error (coefficient of Determination), Mean Square Error (MSE), Mean Absolute Error (MAE), Root mean square error (RMSE), Mean absolute percentage error (MAPE) are employed to measure the accuracy of ARIMA model and forecasting error of each model. These statistical measures of data sample predictions are determined by:

Mean Square Error:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2 \qquad (6)$$

Mean square error is a procedure for estimating the average of the squares error,it is compute data points are close a fitted regression line . Where $\hat{y}_t, y_t$ are actual and predicted population data values at time *t*, *n* is the number of observation in selected time period.

Root Mean Square Error**:** RMSE

$$= \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \qquad (7)$$

Root Mean Square Error is presently square rooot of the mean square error,that is probably the most simply interpreted statistic,RMSE is the distance on averge of a data point from the fitted regression line.

Mean Absolute Error:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t| \qquad (8)$$

Mean Absolute Error is a measure of prediction accuracy of a forecasting method, how close forecasts (predictions) are to the ultimate outcome of data points.

Mean Absolute Percentage Error(MAPE):

$$\text{MAPE}= \frac{100}{\text{n}}\sum_{\text{t=1}}^{\text{n}}\frac{|y_t - \hat{y}_t|}{|y_t|} \qquad (9)$$

Mean absolute percentage error evaluate the quality of the fit, while removing the scale effect and not relatively penalizing bigger errors. $\hat{y}_t, y_t$ are actual and predicted population data values at time t, n is the number of observation in selected time period.

Forecasted Error (EF):

$$EF = \frac{Actual\ values - Forecasted\ values}{Actual\ values} \qquad (10)$$

Forecasted Error is evaluated the difference between the actual values and predicted (forecast value) of any time series data.

Consequence of more diagnostic criteria statistics, such as Bayesian Information Criterion [BIC] and Akaike Information Criterion for selection of most adequate ARIMA forecasting.

To evaluate the forecasting adequacy of the AR (*p*) models for nonstationary data also comparison the non- seasonal Holt-Winters Algorithm forecasting. Forecasting of ARIMA model using Bayesian Information Criterion and AIC.

2.1.4 Forecasting using ARIMA

As we know, an ARIMA (*p*, *d*, *q*) model is expressed three parts, first is order of autoregressive term, second is differencing degree, and last part is the order of moving average term. An autoregressive AR(*p*) is a linear regression of the present value of the time series compared to past values of the time series, the time series is non-stationary then differencing (*d*)is a form of transformation and a moving average (q) term is a linear regression of the present value of the time series in contradiction of white noise[3]. We have considered ten ARIMA models for this study. Here, the most appropriate ARIMA model for forecasting of future time series chosen by Bayesian Information Criterion (BIC). Our selected all ARIMA models are fitted to data from the actual period and the BIC and AIC value are obtained for each ARIMA model fit. The BIC has been generally used for model identification and forecasting in time series and Linear exponential regression. It can, however, be applied quite widely to any set of maximum likelihood-based models [11].

$$BIC = ln\hat{\sigma}^2 + \frac{(p+q)\ln(\ln(n))}{n} \qquad (11)$$

Where *n* implies the sample size of the series and σ^2 is the mean squared error of the ARIMA model fit to the series, *p* is the order of AR parameter while *q* is the order of MA parameter. The BIC is situating, in extent, on the likelihood function, and it is directly associated to Akaike information criterion (AIC) [11,14]. While, the adequacy of fitting models, it is possible to increase the like-hood by manipulating the values of parameters, while stack so may result in over fitting. The Akaike Information Criterion (AIC) are determined the most fitting model by significant value of AIC :

$$AIC = ln\hat{\sigma}^2 + \frac{2(p+q)}{n} \qquad (12)$$

Where *n* implies the sample size of the series and σ^2 is the mean squared error of the ARIMA model fit to the series, *p* is the order of Auto Regressive (AR) parameter while *q* is the order of Moving Average (MA) parameter. Thus, the most adequate ARIMA model shown the smallest AIC value for forecasts were based on that model. The mathematical ARIMA model forecasted formula for Akaike's Information Criterion is as follows:

$$-2LL + 2p \qquad (13)$$

Where, LL is representing the log-likelihood value and p is the number of parameters fitted in the model. The Holt-Winters forecasting is probably to investigate the most adequate forecasts for annual population data than ARIMA modelling.

2. The Holt-Winters Algorithm Forecasting

The Holt-winters algorithm is mainly appropriate for stochastic data series that have a non-seasonality and stationery for linear trend model and exponential smoothing requiring the forecasts in the form [3**,** 20]

$$P_n Y_{n+h} = \hat{m}_n, \quad h = 1, 2, ... \qquad (14)$$

then, m_n is constant and the exponential smoothing forecast of Y_{n+h} based on the observed values. The simple clue is to permit for a time- varying trend via requiring the forecasts to have the mathematical expression [3].

$$P_t\, Y_{t+h} = \hat{\dot{a}}_t + \hat{\dot{b}}_t h, \quad h = 1,2,3,.. \qquad (15)$$

Where, $\hat{a}_t$ and $\hat{b}_t$ is the estimated level and slope estimated trend at time t. The Holt-Winters proposed method for obtaining the quantities $\hat{a}_t$ and $\hat{b}_t$ in above mention (1). Representing in the $\hat{Y}_{n+1}$,one-step forecast $P_n Y_{n+1}$, So we have mathematical form:

$\hat{Y}_{n+1} = \hat{a}_n + \hat{b}_n$ (16)

Similarly, the exponential smoothing, we now take the estimated level at time n+1 is the linear form of the observed and forecast values at time $n+1$, i.e.

$$\hat{a}_{n+1} = \alpha\hat{Y}_{n+1} + (1+\alpha)(\hat{a}_n + \hat{b}_n) \qquad (17)$$

Then the estimated slop at time n+1 as linear form of $\hat{a}_{n+1} - \hat{a}_n$ and the estimated value of slope $\hat{b}_n$ at time n. Thus,

$\hat{b}_{n+1} = \beta(\hat{a}_{n+1} - \hat{a}_n) + (1-\beta)\hat{b}_n$
(18)

By solving equation (17) and (18). we require the initial conditions. The natural selection is to procedure

$\hat{a}_2 = Y_2$ and $\hat{b}_2 = Y_2 - Y_1$

Then the equation (17) and (18) can be solved sequentially for $\hat{a}_i$ and $\hat{b}_i$, i = 3...,n, and the forecasts (19) then have the form:

$$P_n Y_{n+h} = \hat{a}_n + \hat{b}_n h, \quad h = 1,2,3,.... \qquad (19)$$

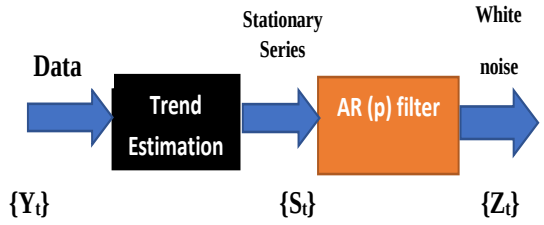
Forecasting values are depending in the (17) and (19) the smoothing parameters α and β. These can also give arbitrarily values (between 0 and 1). Consequently, one-step sum of squares minimize errors are obtained

Forecasted sum of squares error

$= \sum_{i-3}^{n} (y_i - p_{i-1}\, y_{i)2}$ (20)

3.1 The Holt-Winters and ARIMA Forecasting

The Holt-winters (HWS) algorithm is one of the forecasting techniques, which is suited for series that have no seasonality of ARIMA models. This algorithm has two main steps, one-step HWS algorithm is handle time series data in which there are both trend and exponential smoothing variation ,second –step of algorithm are generate forecasts of time series data containing a trend plus noise[19,20].



As we know, the non-seasonal Holt-Winters forecasted is expressed:

$H_t = \alpha A_t + (1 - \alpha)(H_{t-1} + T_{t-1})$ $0 \le \alpha \le 1$ (21)

$T_t = \gamma(H_t - H_{t-1}) + (1 - \gamma)T_{t-1}$ $0 \le \gamma \le 1$ (22)

Where, smoothed value and constant is H_t, α and γ, the trend value and actual value is T_t and A_t of the time series. For this research, Holt-winter Forecasted with ARIMA, α and γ are calculated in minimizing the squared prediction error and the initial values for H and T are:

$H_3 = A_2$ (23)

$T_3 = Y_2 - Y_1$ (24)

The Holt-winters exponential smoothing by parameter in obtained:

$\hat{m}_t = \alpha X_t + (1 - \alpha \hat{m}_{t-1}) \quad t = 2, \ldots\ldots\ldots, n,$ (25)

$P_n Y_{n+h} = \hat{m}_n, \quad h = 1, 2, ...$ (26)

The Holt-winters and ARIMA content the Forecasts relations in mathematical expression:

$P_n Y_{n+1} = Y_n - (1-\alpha)(Y_n - Y_{n-1} Y_n), \quad n \ge 2$ (27)

$Y_t = Y_{t-1} + Z_t - (1-\alpha)Z_{t-1}, \{Z_t\} \sim WN(0, \sigma^2).$ (28)

The exponential smoothing parameter α for forecasting can be interpretation as fitting an associate of an ARIMA model with two-parameters (28) to the data and using the large-sample forecast set in $P_0 Y_1 = Y_1$.

Similarly, it can be shown that Holt–Winters forecasting as fitting value of the ARIMA model with three-parameters

$(1-\beta)^2 Y_t = Z_t - (2 - \alpha - \alpha\beta)Z_{t-1} + (1-\alpha)\, Z_{t-2}$ (29)

The Holt-Winters forecasting method is determined to exponential smoothing in any time series data. Here, we have also compared the Holt-winters Forecasted results with Extrapolation linear and exponential trend Forecasting.

3.*2 Extrapolation Trend Forecasting*

For the more verification of Holt-winters (HWS) algorithm, we have considered the linear and exponential extrapolation forecasting. The linear approach accepts that the population will increase or decrease in the upcoming time interval. The exponential technique obtains that the population will raise or decline exponentially in imminent era as in the present interval. In mathematical terms, the linear and exponential techniques can be expressed

Linear Forecasting:

$F_{Ly} = P_i + (l / A)(P_i - P_f)$ (30)

Exponential Forecasting:

$F_{Ey} = P_i \exp[(\ln(P_i / P_f) / A)\, l]$ (31)

where l is the length of the forecast interval, A is the length of the actual time interval, F_{Ly} is the forecasted year for linear forecasted value, F_{Ey} is the forecasted year for exponential forecasted value, P_i is the population in the initial year and P_f is the population of the final year.

3.3 Normality Check Test for forecasting

When a model is selected a normality check test are finalized to explore that the model is adequate for forecasting. The Geary's α statistic test is indicated that the consistent residuals cannot follow a normal distribution and later the described confidence intervals did not signify the actual values [20, 21].

The mathematical expression is given below:

$$\propto = \frac{{}^{1}/_{n} \sum_{i=1}^{n} |x_i - \bar{x}|}{\sqrt{{}^{1}/_{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad (32)$$

where n is the residuals in the forecasted series being modeled, X_i the value of residual i and X is the mean value of the residuals.

Mean Absolute Percentage Error (MAPE) is computed by dividing the difference between actual value A_t and forecasted value F_t (known as the forecasting error) by the actual value A_t, where i is the series, t is the forecast period and m is the forecasting method.

$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^{n} \frac{|A_t - \hat{F}_t|}{|A_t|} \qquad (33)$$

Mean absolute percentage error evaluate the quality of the fit, while removing the scale effect and not relatively penalizing bigger errors. Since above mention diagnostic checking for ARIMA models and normality check tests, we have explored the adequate ARIMA model, afterwards, by comparing the ARIMA, Holt-winters and Extrapolation Trend forecasting methods and

one best fitted forecasts are calculated and suggest for population forecasting of the city of Karachi.

III. WRITE DOWN YOUR STUDIES AND FINDINGS

Having discussed some basic concepts and hypothetical basis of time series that are enable us to analyze the data. We now present a step-by-step analysis of our population data time series. In Figure.2 illustrates the population trend of the natural logarithm population of Karachi from 1951 to 2015, respectively.

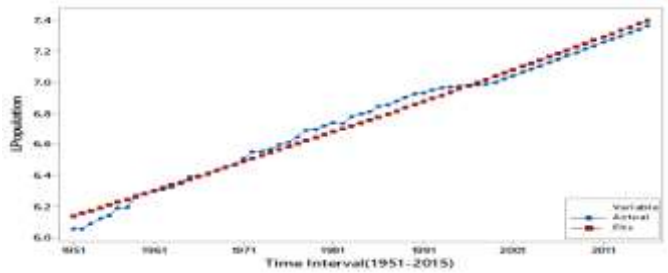


Figure 2: The Logarithmic population time series plot of Karachi during 1951-2015

In Figure 3 and 4 depicts the trend of the population after taking the first and second differencing of the natural logarithm of the population time series data. Certainly, the first differencing plot of the trend of actual population is unusual as the first difference plot of trend of logarithmic population except the second (2^{nd}) differencing plot is approximately stationary. Thus, the 2^{nd}differenced logarithmic population time series is being for onward analysis.

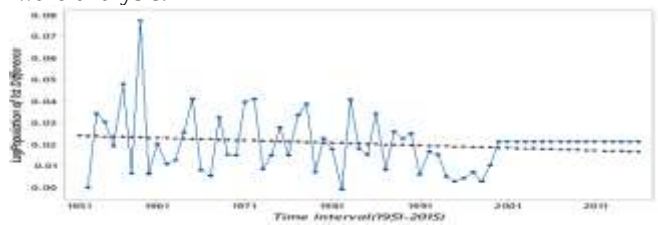


Figure 3: Logarithmic Population Trend time series plot of Karachi after 1st Differencing

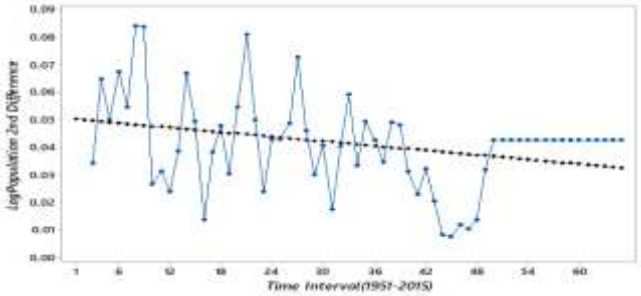


Figure 4: Logarithmic Population Trend time series plot of Karachi after 2^{nd} Differencing

In the figure.5 and 6 depicts of ACF and PACF plots of the second (2^{nd}) difference of logarithm of the population are shown all the points at different lags are within the 95% confidence limits, it is an indication that the selected skimping model might be without moving average components. In PACF graph are depicts all points at different lags of the figure are within the 95% confidence limits except two points, one at lag 1 and second at lag 5.

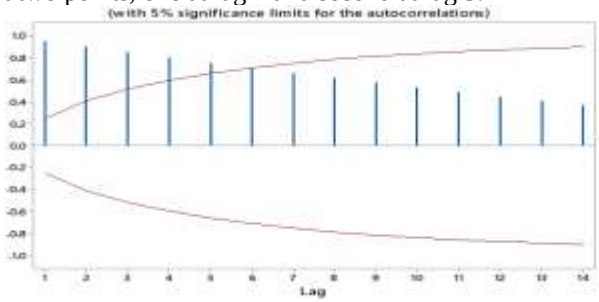


Figure5: ACF Plot of differencing logarithmic Population

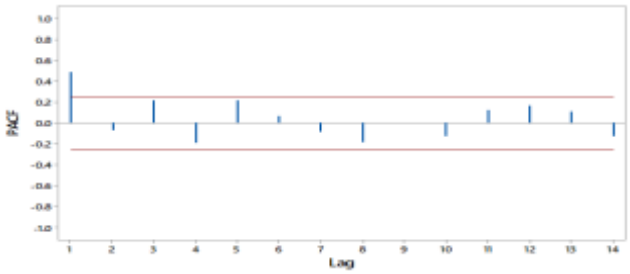


Figure: 6 PACF Plot of differencing logarithmic Population

The point at lag 1 is clearly out of the positive limit whereas spike at lag 1 is close to the negative limit; other spikes at different lags in PACFs are clearly within the 95 % limits.

Table 1: Auto correlation and PACF of 2^{nd} difference for logarithm population

LAG	ACF	ACF T-Statistics	LBQ1	PACF	PACF T-Statistics
1	0.486	3.891	15.865	0.486	3.891
2	0.185	1.222	18.208	-0.067	-0.536
3	0.226	1.455	21.740	0.213	1.707
4	0.054	0.336	21.944	-0.189	-1.515
5	0.099	0.619	22.651	0.214	1.716
6	0.221	1.366	26.195	0.070	0.563
7	0.070	0.423	26.560	-0.082	-0.654
8	-0.102	-0.612	27.345	-0.188	-1.501
9	-0.092	-0.548	27.992	-0.003	-0.027
10	-0.164	-0.975	30.101	-0.130	-1.038

Table 1 are indicated values of autocorrelation coefficients (ACF) and partial autocorrelation coefficients (PACF), Students t- statistics, Ljung Box statistics and P-values corresponding to the different lags from 1 to 10 of 2nddifference of natural logarithmic population series. It is concluded that all the autocorrelation coefficients (ACF) and partial autocorrelation coefficients (PACF) are perform significantly from zero and consequently the 2nddifference of the logarithm of the population series appears to be stationary. The correlogram is that it helps in determining the p, q values of the ARIMA model. The ACF and PACF plot of second (2nd) differenced of logarithm population series is denoted by yt for yt= 1,2……66 ,where Yt =∇zt .It is observed that ACF and PACF yt are described by correlations to turn in symbol and which tend to damp out with increasing lag. As a result, the autoregressive moving average of order (p, d, q) are proposed since both the ACF and PACF of the yt appear to be tailing off. Thus, we have selected ten ARIMA model is parameter values for different orders of autoregressive (AR), auto regressive integrated moving average (ARIMA) models e.g. AR(1), AR(2), MA(1), MA(2), ARIMA(1,2,1), ARIMA(1,2,2), ARIMA(2,2,1), ARIMA(2,2,2), ARIMA(3,2,1), ARIMA(3,2,2), ARIMA(4,2,1), ARIMA(4,2,2), ARIMA(5,2,1), and ARIMA(5,2,2) (p, d, q) models of population data by d= 2, differenced stochastic process and it to most appropriated models are preferred to non-seasonal Holt–Winters algorithm forecast the population in Karachi on the future.

According to the ARIMA identification and estimation checking, if our selected more than one model provides similar information, the insignificancy of model is shown minimum number of parameters for improve the estimation and interpretation of parameters. If (P-value ≤0.05) corresponding to an estimate in the ARIMA model, the hypothesis that the parameter equal to zero is rejected on the further proviso (P-value ≥0.05) consequent to an estimate in the model, the hypothesis that the parameter equal to zero is not rejected which suggested that the explanatory variable should not be consist of in the model. As a result, in the perspective of all diagnostic checking, Parameter estimates and forecasted error test the proposed appropriate model is ARIMA (1,2, 1) model is identified. The parameters of the fitted model are estimated using parameters descriptive values Hessian error, Asymptotic error with 95 % confidence interval of lower and upper bounds statistic in table 2

Table 2: The Parameters descriptive statistics of most adequate ARIMA (1,2,1) Model

Parameter	Value	Hessian Standard error	Lower Bound	Upper bound (95%)	Asymptotic standard error	Lower bound (95%)	Upper bound (95%)
AR(1)	-0.26	0.144	-0.55	0.02	0.13	-0.518	-0.01
MA(1)	-0.88	0.099	-1.07	-0.682	0.065	-1.004	-0.75

Also, the ACF, PACF and p-values of the Durbin Watson statistic for the residuals of the ARIMA(1,2,1) model within the confidence interval, least value of error so we have no evidence to reject the model. The residuals plot of the autocorrelations of ARIMA (1,2,1) model, we see that the autocorrelations values are statistically equal to zero. The appropriateness of each AR (*p*) models can be tested by using the diagnostic checking namely R–square, Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Final Prediction Error are shown in table 3.

Table 3: The Selected ARIMA (p, d, q) model and Diagnostic Test

Diagnostic checking	ARIMA (*p*, *d*, *q*)									
Criterion	(5,2,2)	(5,2,1)	(4,2,2)	(4,2,1)	(3,2,2)	(3,2,1)	(2,2,2)	(2,2,1)	(1,2,2)	(1,2,1)
R-Square	0.941	0.939	0.951	0.936	0.945	0.930	0.970	0.981	0.972	0.990
SSE	0.010	0.010	0.010	0.010	0.099	0.010	0.011	0.011	0.015	0.010
MSE	0.016	0.019	0. 016	0. 013	0.016	0.016	0.002	0.017	0.018	0.013
Variance	0.015	0.000	0.000	0.000	0.010	0.010	0. 019	0. 019	0. 016	0.015
MAPE	0.158	0.140	0.200	0.180	0.165	0.188	0.198	0.160	0.170	0.120
EF	0.018	0.019	0.019	0.019	0.018	0.009	0.015	0.019	0.010	0.010
Durbin Watson	2.122	1.996	1.925	1.849	1.804	1.759	1.942	1.980	2.134	1.680
Log-likelihood	-375.4	-375.4	-375.4	-375.4	-375.6	-375.2	-373.7	-373.0	-373.0	-372.9
AIC	-359.4	-361.4	-361.4	-363.4	-363.6	-365.2	-366.9	-365.0	-363.7	-365.0
BIC	-342.1	-346.3	-346.3	-350.4	-350.6	-354.4	-360.5	-356.4	-352.9	-356.3
HIC	-356.8	-359.4	-359.4	-361.9	-362.1	-364.2	-366.5	-364.3	-362.7	-364.3
AIC/BIC	1.050	1.044	1.044	1.037	1.037	1.030	1.018	1.024	1.031	1.024
HIC/BIC	1.043	1.038	1.038	1.033	1.033	1.028	1.017	1.022	1.028	1.022

The Bayesian Information Criteria (BIC) and Akaike Information Criterion (AIC) are established that the model is statistically significant, appropriate, and adequate. Moreover, the low value of MSE and RMSE indicates a good fit for the ARIMA (1,2,1) model. As well, the high value of the R-Square indicates a perfect prediction over the mean. Investigation are showed that the ARIMA (1, 2, 1) model is superior to the other selected models having the least BIC value specifically AIC, Our result represents the least value of REMSE, highest value of Log-Likelihood and smaller values for HIC: BIC ratio and greater values for the ratio AIC: BIC for ARIMA (1, 2, 1) specifically with respect to the selected ten ARIMA models depicts are declared as adequate models which is inventory at the table 3.Meanwhile, the population time series data used in this research did not demonstrate any type of seasonality, the population data seems to be non-seasonality, seasonal forecasting are not be discussed in this paper.

we also evaluated the accuracy of the population forecasting may be affected by the move from ARIMA modelling to Holt-Winters algorithm and extrapolation linear and exponential trend forecasting methods in the basis of normality check test. we considered the linear and exponential extrapolation equation (30 and 31) for forecasting of population time periods series (for 5-years, 10-year and 15-year) form as Linear F_{L05} = 6.0560 + (*0.076923*) (1.307111), F_{L10} = 6.0560 + (*0.1428571*) (1.307111) and F_{L15} = 6.0560 + (*0.2*) (1.307111) ,similarly exponential forecasting form as F_{E5}= (6.0560) exp [(ln(0.822585) / 65) *5*] ,F_{E10}= (6.0560) exp [(ln(0.822585) / *70*) *10*] and F_{E15}= (6.0560) exp [(ln(0.822585) / *75*) *15*].

Our result shown that accuracy of ARIMA (1,2,1) model and Holt-Winters is dispirited to exponential extrapolation trend forecasting methods (table 4). The forecasted Populations for the years 2020, 2025, 2030 are also depicts in table 5.

Table 4: The Most Appropriate Models forecasts by Normality test

Model	Time Interval	Forecasted Year	Geary's α statistic test	MAPE	SSE	RMSE	MPE
ARIMA (1,2,1)	1951-2015	2030	0.95	0.15	0.011	0.013	0.015
Holt-Winters			0.979	0.16	0.014	0.015	0.018
Linear			0.861	0.46	0.034	0.079	0.047
Exponential			0.861	0.56	0.043	0.095	0.055

Table 5: Forecasted Populations for the Years 2020, 2025, 2030

Model	2020 (log Population)	2025 (Log Population)	2030 (Log Population)
Linear	6.1565	6.2427	6.3174
Exponential	6.1476	6.2273	6.2972
ARIMA(1,2,1)	7.465	7.566	7.667
Holt-winter	7.47	7.577	7.684

the normality check test as Geary's α statistic test, Sample Kurtosis b_2 statistic and Mean Absolute Percentage Error values for the ARIMA and Holt-Winters models are very close to the MAPE values, Geary's α statistic values and Sample Kurtosis b_2 statistic values are shown that the ARIMA and Holt-Winters forecasted residuals values do not appear normal distribution.
In fact, the differences in MAPE values for ARIMA and Holt–Winters algorithm forecasts fixes not exceed 5, percentage values, then, as a comparison of all above methods, also ARIMA and Holt-Winters forecasted normality MAPE values are comparatively small, the ARIMA (1,2,1) forecast seems to be the most accurate for the period of 5,10 and 15 years (2020,2025 and 2030) estimated time intervals and Holt-Winters algorithm forecasted appearances to be 5 to 15 years estimated time intervals(table 6 and figure 7).

Table 6: Holt Winter forecasting for the Years 2016-2030

Time Interval	Prediction	sqrt(MSE)	Lower	Upper
2016	7.385	0.014	7.356	7.413
2017	7.406	0.021	7.364	7.448
.2018	7.427	0.029	7.370	7.485
2019	7.449	0.038	7.374	7.524
2020	7.470	0.048	7.376	7.564
2021	7.492	0.059	7.377	7.607
2022	7.513	0.070	7.376	7.650
2023	7.534	0.082	7.374	7.695

2024	7.556	0.094	7.371	7.741
2025	7.577	0.108	7.366	7.788
2026	7.599	0.121	7.361	7.836
2027	7.620	0.136	7.354	7.886
2028	7.641	0.150	7.347	7.936
2029	7.663	0.166	7.338	7.987
2030	7.684	0.181	7.329	8.039

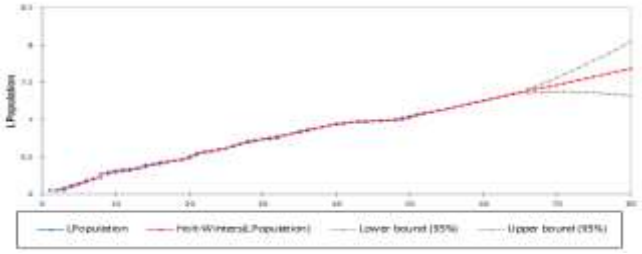


Figure: 7 ARIMA (1,2,1) and Holt-Winters forecasts population from 1951 to 2030.

These forecasting methods are rectifying change from Holt-Winters to ARIMA modelling. These forecasting methods are rectifying transformation from Holt-Winters to ARIMA modelling.

$(1-0.400)^2 Y_t = Z_t - (2-0.780-0.312)Z_{t-1} + (1-0.780)\,Z_{t-2}$

subsequently, ARIMA (1,2,1) is appear most appropriate forecasts for the population of the city of Karachi for the next 15 years. The ARIMA Forecasting based on the fitted model are calculated up to lead time of 14, and the one-step forecasting, and the 95% confidence limits are presented in table.7 and figure.8.

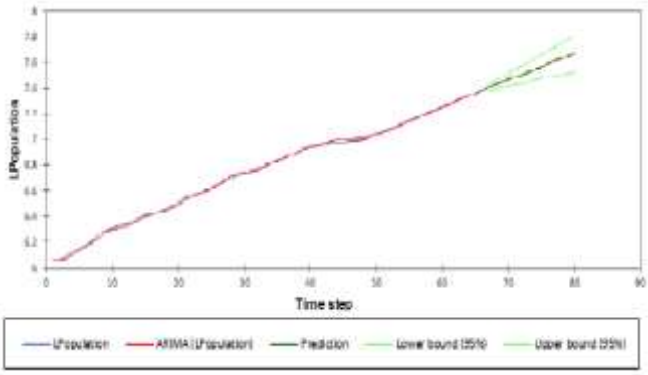


Figure 8 The population forecast from 1951 to 2030 by ARIMA (1,2,1)

Table: 7 One-step forecast for the Years 2016-2030 by ARIMA (1, 2, 1) Model

Lead time	Forecast	95% Lower Limit	95% Upper Limit
2016	7.4032	7.356	7.413
2017	7.4044	7.3791	7.4298
2018	7.4247	7.3912	7.4583
2019	7.4449	7.4022	7.4876
2020	7.4651	7.4141	7.5161
2021	7.4853	7.426	7.5447
2022	7.5056	7.4379	7.5731
2023	7.5257	7.4498	7.6017
2024	7.5459	7.4615	7.6303
2025	7.5661	7.4731	7.6591
2026	7.5863	7.4846	7.688
2027	7.6065	7.496	7.717
2028	7.6267	7.5072	7.7462
2029	7.6469	7.5182	7.7756
2030	7.667	7.5291	7.8051

The lower values and upper values stand for the lower bounds and upper bounds of the Confidence Interval (C.I). There is a 95% chance that the forecasted values will fall into this range. The actual and forecasted values are reasonably close, which confirms that our adequate ARIMA should be superior for forecasting.

IV. CONCLUSION

While ARIMA models are usually used only in time series analysis, it is concluded that the ARIMA (1, 2, 1) is a fitted and most appropriate model out of the other fitted AR (2) ARIMA models. An ARIMA (1,2,1) model had the least BIC value of −356.3269 MAPE of 0.1509, RMSE of 0.012435 and R-square of 0.9891 similarly, although this model is the most appropriate model based on the BIC is as well as the entire Diagnostic Check test**.** Among the most adequate ARIMA model, Holt-Winter algorithm and extrapolation trend Forecast methods are used to forecast population timeseries, thus normality check test are suggest to differences in MAPE values are relatively small, the Holt-Winters and ARIMA (1,2,1) model seems to be the most accurate model for the 15-year forecast period and the ARIMA

(1,2,1) is the most accurate model for the specific years i.e.2015 to 2030,2020, 2025, and 2030. The city of Karachi is being the world's eight urban populated city has a logarithm population of 7.33 in 2015. It is also testimony that population would be 7.656 by 2030 which is almost same to our forecasted population. In briefly, the estimates provided in table 1.

using ARIMA (1,2,1) are close to other researcher's finding and are equally important for Government of Pakistan, Non-Government Organizations as well as insurances companies, Health Department for future planning and projects.

Karachi is the most populous city in Pakistan, and it has a most important seaport and financial center. The population density is about 6,000 per square kilometers (15,500 per square miles) and the World's third largest city population. The city of Karachi has tackled various problems by uncontrollable human population, urban management, and planning. The Model results are helpful for future planning and judicious distribution of resources for development. The results of this study will also prove to be useful for future researchers working on epidemiology and oncology field to improve and rectify the current pandemic COVID-19 disease in Karachi Region. We will further discuss and elaborate it in our next communication.

Acknowledgement

We are thankful to the **Pakistan Bureau of Statistics** for providing population data and Higher Education Commission (HEC) Pakistan for financial support (vide NRPU grant #20-4039/R&D/HEC/14/697). The content is the part of first author's Doctoral thesis.

REFERENCES

[1] Zakria, M., Stochastic models for population of Pakistan,PhD thesis University of Karachi ,2009.

[2] DHAMO, E. and L. PUKA, *An ARIMA birth number per month model for Albanian population*. Month: , 2000.

[3] Brockwell, P.J. and R.A. Davis, *Introduction to time series and forecasting*. Springer, 2016.

[4] Akaike, H., On entropy maximization principle,*Application of statistics*, 1977.

[5] Hussain, M., S. Abbas, and M. Ansari, Arabian seawater temperature fluctuations in the twentieth century, *Journal of Basic and Applied Sciences*, 8(1): pp. 105-109, 2012.

[6] Ljung, G.M. and G.E. Box, On a measure of lack of fit in time series models, *Biometrika*,65(2): pp. 297-303, 1978.

[7] Hiscock, P. and T. Maloney, Australian lithic technology. *The Routledge Handbook of Archaeology and Globalization*: pp. 301, 2016.

[8] Ong, C.-S., J.-J. Huang, and G.-H. Tzeng, Model identification of ARIMA family using genetic algorithms. *Applied Mathematics and Computation*, 164(3): pp. 885-912, 2005.

[9] Meyler, A., G. Kenny, and T. Quinn, Forecasting Irish inflation using ARIMA models, 1998.

[10] Box, G.E., G.M. Jenkins, and G.C. Reinsel, Time series analysis: *forecasting and control*, Prentice Hall, Englewood Cliffs, NJ: *pp. 282-285* , 1994.

[11] Clement, E.P., Using normalized bayesian information criterion (BIC) to improve box-jenkins model building. *American Journal of Mathematics and Statistics*, 4(5): pp. 214-221, 2014.

[12] Burnham, K.P. and D.R. Anderson, Model selection and multimodel inference: a practical information-theoretic approach. *Springer Science & Business Media*, 2003.

[13] Dhamo, E. and L. Puka. 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*, 2010.

[14] Pankratz, A. and A. Pankratz, Forecasting with univariate Box-Jenkins models, *concepts and casses*, 198

[15] Hamilton, J.D., Time series analysis, *Princeton university press Princeton*,Vol. 2 , 1994.

[16] Tayman, J., S.K. Smith, and J. Lin, Precision, bias, and uncertainty for state population forecasts: An exploratory analysis of time series models. *Population Research and Policy Review*, 26(3): pp. 347, 2007.

[17] Spyros, M., S.C. Wheelwright, and R.J. Hyndman, *Forecasting: methods and applications*. Editorial JohnWiley & Sons, Inc. Tercera Edición. Estados Unidos, 1998.

[18] Shumway, R.H. and D.S. Stoffer, *Time series analysis and its applications: with R examples*,Springer Science & Business Media, 2006.

[19] Carriero, A., G. Kapetanios, and M. Marcellino, Forecasting government bond yields with largeBayesian vector autoregressions. *Journal of Banking & Finance*, 36(7): pp. 2026-2047, 2012.

[20] Walters, A. and Q. Cai, Investigating the use of Holt-Winters time series model for forecasting population at the State and sub-State levels. *University of Virginia*, 2008.

[21] Chatfield, C. and M. Yar, Holt-Winters forecasting: some practical issues. *The Statistician*, pp. 129-140, 1988.

[22] Ng, Enders KO, et al. "mRNA of placental origin is readily detectable in maternal plasma." *Proceedings of the National Academy of Sciences* 100.8 pp. 4748-4753, 2003.

AUTHORS

First Author (Correspondence Author)– Dr. **Muhammad Ilyas**, PhD, Department of Mathematics, Government College University Hyderabad, Sindh, Pakistan
dr.m.ilyas@gcuh.edu.pk
Second Author –Dr. **Shaheen Abbas**, PhD, Laboratory of Applied Mathematics and Data Analysis (LAMDA) Mathematical Sciences Research Centre, Federal Urdu

University of Arts, Sciences and Technology, Karachi, Pakistan
email address. shaheen.abbas@fuuast.edu.pk,

Third Author – **Dr. Syed Akhtar Raza ,PhD**
Department of Compute Science Federal Urdu University of Arts, Sciences and Technology, Karachi, Pakistan. akhtar@fuuast.edu.pk

Fourth Author Dr. Wajid Ali, PhD , Department of Mathematical Sciences Karakoram International University Gilgit. Pakistan wajid.ali@kiu.edu.pk

Fifth Author-Dr Syed Masood Raza , PhD
Department of Physics Federal Urdu University of Arts, Sciences and Technology, Karachi, Pakistan.
masood.raza@fuuast.edu.pk