Journal of Probability and Statistical Science

20(1), 98-107, August 2022

Forecasting of Immigrants in Canada Using Forecasting Models

Sharandeep S. Pandher
University of Regina

Arzu Sardarli,
First Nations University of Canada

Andrei Volodin
Sino-Canada Research Centre of Nonlinear Dynamics and Noise Control, Xiamen University of Technology and the University of Regina

ABSTRACT In Canada, the number of international students, temporary workers and refugees from every part of the world grows each year. Therefore, forecasting immigration is important for the economy of Canada and Labor Market. In this regard, four forecasting approaches have been applied to the annual data of immigrants for the period 2000-2019. The accuracy of Moving average (MA), Autoregressive (AR), Autoregressive moving average (ARMA), Autoregressive integrated moving average (ARIMA) models were checked via comparing Akaike's information criteria (AIC), Bayesian information criteria (BIC), Mean error (ME), Root mean square error (RMSE), Mean absolute error (MAE), Mean percentage error (MPE), Mean absolute percentage error (MAPE) and Mean absolute scaled error (MASE) and graphical approaches such as ACF plots of residuals. Experimental results showed that ARIMA (1,2,4) is the best-fitted model for forecasting immigrants in Canada. Selected forecasting approaches are applied to predict immigrants for five years from 2020-2024.

Keywords; Immigrants, ARIMA, AIC , BIC, ME,[1] and MASE.

1 Introduction

The first phase of immigrations to Canada was two hundred years ago when French people, American investors and military personnel from Britain came to Canada and settled in what is now the province of Quebec. The next wave of immigration came mainly from Britain and Ireland during the American invasion to counter French influence in this region. After that, the world wars were the impetus behind European people migrating to Canada. Currently, the majority of immigrants come from China, India, and other south Asian Countries [1].

The main causes of immigration to Canada are the stability of the economy, quality of education, and standard of life. Forecasting immigrants plays an important role in

□Received October 2021, revised December 2021, in final form December 2021.

□Sharandeep Singh Pandher (corresponding author) is a PhD student in the Department of Mathematics and Statistics at the University of Regina, Regina, Saskatchewan S4S 0A2, Canada; email: sssinghsharan@gmail.com. Arzu Sardarli is affiliated to the First Nations University of Canada. Andrei Volodin are affiliated with the Sino-Canada Research Centre of Nonlinear Dynamics and Noise Control, Xiamen University of Technology and the University of Regina

understanding the needs of the labor market and maintaining the economic growth of the country in the future.

Shaw [2] highlights some of the challenges to migrate in South Africa by analyzing historical data. In this connection, Canadian government must be aware of future trends of immigration to determine the required needs of the country. In this context, the availability of sufficient data and past statistical analysis of immigrants helps the government make suitable migration programs and policies that benefit the country. [3] In 2017, The Minister of Immigration, Refugees and Citizenship introduced a three-year immigration plan for the period 2019-21 that includes a target of 330800 immigrants in 2019 and 350,000 in 2021; however, only 313580 immigrants came in 2019. Various migration programs are used to achieve the desired immigration numbers: the Federal Economic, Provincial/Territorial Nominees, family reunification, Quebec selected skilled workers and Business etc.

Migration information is important to Canada and other countries, because researchers and scientists can use the data to elucidate migration patterns. Various Forecasting models, such as Holt winter, Exponential smoothing, Bass model, linear and non-linear models and regression models, are applied to predict the number of immigrants and Population [4].

Autoregressive (AR), Moving Average (MA), Autoregressive and Moving Average (ARMA) and Autoregressive Integrated Moving Average Models (ARIMA) are useful in Migration modeling and forecasting. They are able to catch the time series annual, quarterly, monthly and weekly patterns as any existing time series data can be used by them. According to Keilman [5], appropriate methods and very close attention should be used when forecasting population movement, as it is uncertain. In this regard, the ARIMA model can be a more suitable and accurate model. The goal of this research is to model the adequacy and forecast immigrants to Canada using ARIMA, ARMA, AR and MA models. R- Software packages are applied for model fitting and forecasting.

2 Literature Review

The initial motivation for forecasting modelling is to examine the approaches and measures for making inferences from historical information. A model can be used for prediction, review and explanation of the sample information. Any regular pattern of data can be controlled by forecasting models [10].

The accurate forecasting of immigrants is important for every country that helps the government to make the right policy in the favor of immigrants so that they will get a job opportunity in the job market and live a better quality of life. In this regard, many researchers worked on the data set of immigrants.

Kaushal et al. [11] worked on the employment and earnings growth of immigrants in Canada and the United States between 1996 and 2008. Results showed that the average number of immigrants has a lower growth rate compared to native-born Canadians. While reverse situation exists in the United States. Makoni et al. [12] studied the problem of immigrants in the Republic of Zimbabwe through the Seasonal ARIMA model for the monthly data 2001-2017. The adequacy of the model was checked via the MSE, RMSE and AIC. The proposed model was also used for two-year forecasts of immigrants. Results showed that the number of immigrants reached a peak in December compared to other months.

Zakria and Muhammad [13] forecasted the trend of the population of Pakistan for 20 years into the future through the ARIMA (1, 2, 0) model. They used data from 1951-2007 and found that the entire population of the Islamic Republic of Pakistan will reach 230.7 million in

2027. The efficiency of the model was verified by AIC, SC, MSE, P-values and graphical techniques such as ACF and PACF plots of residuals. This predication was nearest to the forecast of 229 million Pakistan population by various bureaus.

Jakub et al. [14] worked on the problem of misled prediction for international migration in the United Kingdom. They studied ongoing forecasting approaches of immigration processes to examine the uncertainty of forecasting models through error and scale of uncertainty. Results indicates that no single forecasting approach suited for different flows so the author suggested a tailored approach should use for forecasts .

Kumara [15] applied SARIMA, Holt winters (HW) and the Grey model (GM) on the monthly data of international tourists arriving in India for the period 2000 to 2015. Results indicated that the HW and SARIMA models are appropriate for the historical dataset as compared to the Grey model in terms of turning point analysis (TPA), U statistics and posterior variance ratio test (PVRT) and mean absolute percentage error (MAPE).

Thabani [16] used the Box-Jenkins ARIMA model for annual time series data of Mexico population for the period 1960-2017. Experimental results showed that ARIMA(3,2,1) is convenient in terms of AIC and the population will continue to grow for the next 30 years reaching 180 million.

Brain [17] introduced the relationship between statistical methodology and time series analysis by the ARIMA model. The majority of approaches that define the statistical correlation between the variables cannot provide the information of error in the next measurement on real data. There is analysis variance and regression outcomes may be faulty and misguiding. This type of data may be examined by forecasting models such ARIMA.

Raymer [18] has worked on the demand of improved forecasting models for foreign migration through the bilinear model along with Bayesian inferential structure. His team studied forecast sex and age structure of emigration and immigration in Britain and another kind of time series data related to immigration age and sex in South Korea, Sweden and Australia. The efficiency of the prediction is cross checked and measured with the observed time-series data.

Christine [19] studied stationary and non-stationary presence of unit roots and seasonal unit roots of the time series data of international tourism by the ARIMA model for the period 1975-1989 in Australia. The MAPE and RMSE were used to check the accuracy of the ARIMA model. However, the existing model predicts the sightseer arrival from Singapore for 7 years.

Chu[20] developed a hybrid approach of non-seasonal ARIMA and sine wave nonlinear regression for the prediction of foreigner visitor arrival in Singapore. The new approach provides the lowest mean absolute error as compared to other uni-variate forecasting models.

3 Methodology

Uncertainty in the demand of overseas workers in the Canadian labor market can be reduced by using forecasting models such as the Moving Average (MA), Autoregressive (AR), Autoregressive and Moving average (ARMA) and ARIMA models. The most suitable forecasting model was identified with the help of accuracy methods. In this study, eight accuracy methods were used: Akaike's information criteria(AIC), Bayesian Information Criterion(BIC), Mean error (ME), Mean Square Error(MSE), Root mean square error (RSME) and Mean Absolute percentage error (MAPE) etc. The number of trials has been applied in the forecasting methods, out of which the smallest value is adopted for eight accuracy methods.

3.1 Material and Methods:

We collected annual migration data from the Government of Canada for the years of 2000 to 2019 [21]. Four different type of forecasting models applied on this dataset that explain in the subsection.

3.1.1 Moving Average Model

The Moving Average method is widely used in time series analysis. The linear output variable of the moving average model depends on the present and previous values of a stochastic term. The MA model [6] is generally set up in the following way:

$$\begin{aligned} x_t &= z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2} + \dots + \theta_k z_{t-k} \\ E(z_t) &= 0, V(z_t) = \sigma_z^{\ 2}, E(z_t z_s) = 0, s \neq t \end{aligned} \tag{1}$$

where k is the order of the MA model, and z_t is the estimated residual at each time period. The model fitting of the MA model is complex, as terms of lagged error are not detectable. The main advantage of this model is that it works in linear and non-linear trends.

3.1.2 Autoregressive Model

The autoregressive model of order k is defined in the following way

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_k x_{t-k} + z_t \tag{2}$$

where $\{z_t\}$ is white noise series and θ_i are parameters of the model.The drawback of the AR model is that it may be non-stationary as compared to the Moving Average Model [7].

3.1.3 ARMA Model

In 1951, Whittle introduced the concept of the general ARMA model in his dissertation. After two decades, this model became very famous in the book by Pelham Box and Jenkins[8]. It is a combined approach of the AR and MA models in which the AR model works on its past values while the MA model concerns the term of error for time series analysis. It is more convenient when the process is a function of a series of invisible surprises along with a personal approach. The mathematical equation of the ARMA (k, l) model as below:

$$\begin{aligned} x_t &= \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_k x_{t-k} + z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2} + \dots + \theta_l z_{t-l} \\ E(z_t) &= 0, V(z_t) = \sigma_z^{\ 2}, E(z_t z_s) = 0, s \neq t \end{aligned} \tag{3}$$

3.1.4 ARIMA Model

The ARIMA model was introduced to solve the problem of non-stationary behaviour in time series data. In this regard, a differencing step can be used to eliminate the non-stationary trend of the dataset in the ARMA model. When this step is added, the model is called the generalization of Autoregressive and Moving average model [8]. The mathematical equation of the model is

$$\phi_k(B)(1-B)^d x_t = \theta_l(B) z_t \tag{4}$$

where ϕ_k and θ_l are polynomials of order "k" and "l" and "d" is the difference of the time series.

3.1.5 Formulation of the forecasting model:

The formulation of the forecasting model includes the following steps. [i]

1. Determination of Model structure.
2. Measuring the coefficients of formulation.
3. Fitting the model test on the estimated residual.

4. Prediction of the future trend of the dataset.

3.2 Measuring Forecasting Error

Currently, researchers and scientists do not agree on the measure used to find the best suitable forecasting model. Accuracy is widely used to determine the superior forecasting model, because it plays a vital role in assessing the quality of prediction. The aim of the forecast is to minimize errors. The difference between a true value and its forecast value is known as forecasting error [9]. Some ordinary indicators are Akaike's information criteria, Bayesian information Criterion, Root means square error and mean absolute percentage error.

3.2.1 Akaike's information criteria (AIC)

Akaike's information criteria introduced by Hirotugu Akaike who was a Japanese Statistician. It was first used in information theory, and today is applied widely by researchers to study time series models. The selection of the model is the most challenging task in statistical inference that can be solved via AIC. The mathematical representation of AIC for the model in the following way.

$$AIC = 2K - 2\ln(\hat{L})$$

where k is the number of estimated parameters and $\hat{L}$ is the greatest value of the likelihood function of the model.

3.2.2 Bayesian Information Criterion (BIC or SIC):

Schwarz introduced the Schwarz information Criteria in 1978; although it is better than the AIC, it is not appropriate in a complex set of models.

$$BIC = K.\ln(n) - 2\ln(\hat{L})$$

where n is the number of observations in the model.

3.2.3 Mean Square Error (MSE) or Mean Square Deviation(MSD)

It is widely used to evaluate the quality of forecasting models and eliminate a number of predictor variables without losing the predictive ability of forecasting models in time series analysis. The value of the MSE nearest to zero indicates the best quality forecasting model. The average squared difference between the predicted values and the true values is known as the mean square error.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i$ is the predicted value of the data. The square root of MSE is called the root mean square error (RSME)

3.2.4 Mean Absolute percentage error (MAPE or MAPD)

The accuracy of forecasting models is predicted via mean absolute percentage error in time series. It provides useful information about the trend of time series during the forecasting. It is applied as a loss function for regression problems, but the major drawback of this evaluation is that it cannot work with a zero value in time series.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{A_t - F_t}{F_t}\right|$$

where A_t is true value while F_t is a forecast value.

3.2.5 Mean Absolute error (MAE)

It is assessing the forecasting error in time series analysis as below

$$e(t) = y(t) - \hat{y}(t \mid t-1)$$

where $y(t)$ represents observation and $\hat{y}(t \mid t-1)$ shows forecast $y(t)$ based on all the previous observations. In statistics, the mean absolute error is defined in the following way

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

where y_i is the predicted value and x_i the true value.

3.2.6 Mean Absolute Scaled error (MASE):

Hyndman and Koehler designed the mean absolute scaled error in 2005. It is a scale free error grade that provides each error as a ratio compared to a baseline's average error. The benefits of it is that it never produces undefined values.

$$MASE = mean\left(\frac{|e_i|}{\frac{1}{T-1}\sum_{t=2}^{T} |Y_t - Y_{t-1}|}\right)$$

where e_i is the forecast error for the given period.

4 Result and discussion

Four forecasting models were used for model parameter estimation, five-year forecasting and diagnostic checking. Twenty years of data of Canadian immigrants were used for modelling purposes.

4.1 Model estimation:

R Package was used to determine the estimations of parameters models(Table 1). The values of AIC, and BIC were smallest for ARIMA (1,2,4) that were 177.94 and 183.28 and the MASE value was noted as 0.57, which suggested that ARIMA (1,2,4) was the most adapted model for forecasting immigrants to Canada. Overall, ARIMA (1, 2, 4) has the lowest values on 7 evaluation criteria as compared to the other three forecasting approaches.

Models	AIC	BIC	ME	RMSE	MAE	MPE	MAPE	MASE
MA(0,0,3)	195.82	200.80	0.55	24.63	18.76	0.79	7.41	0.81
AR(3,0,0)	193.54	198.52	5.47	22.93	17.83	1.27	6.97	0.77
ARMA(1,0,2)	193.06	198.04	5.10	22.33	17.50	1.15	6.86	0.75
ARIMA(1,2,4)	177.94	183.28	2.09	19.56	13.27	0.15	5.14	0.57

Table 1: Comparison of Forecasting Accuracy of the Models

4.2 Descriptive Statistics of the dataset

Descriptive statistics used to know the significance features of the dataset which represented in Table 2. It shows the minimum and maximum value of Candian immigrants is 199.17 and 323.19. The average value of migrants is 257.96. The dataset is positively skewed and kurtosis value which is greater than 3.

Min	Max.	Median	Mean	S.D	Skewness	Kurtosis
.17	323.19	255.39	257.96	30.43	0.34	3.35

Table 2: Descriptive statistics of Canadian Immigrants

4.3 Forecasting of Immigrants by Fitted four forecasting Models

Forecasting models predict the number of immigrants for the period 2020-2024 (Table 3). To evaluate the forecasting capability of MA, AR, ARMA and ARIMA Models, significant measures of sample period accuracies were computed. The MAPE and MAE of ARIMA (1,2,4) for immigrants were 5.14 and 13.27, respectively, which showed the high accuracy of model over the other three forecasting approaches. Results (Figure 1) showed increasing trends in the number of immigrants for the ARIMA (1,2,4) model.

Models	2020	2021	2022	2023	2024
MA(0,0,3)	275.135	277.797	268.152	259.655	259.655
AR(3,0,0)	284.027	300.488	295.171	287.431	292.707
ARMA(1,0,2)	300.147	310.423	307.240	304.293	301.565
ARIMA(1,2,4)	297.414	314.430	322.403	329.556	337.113

Table 3: Forecast of immigrants for the period 2020-2024

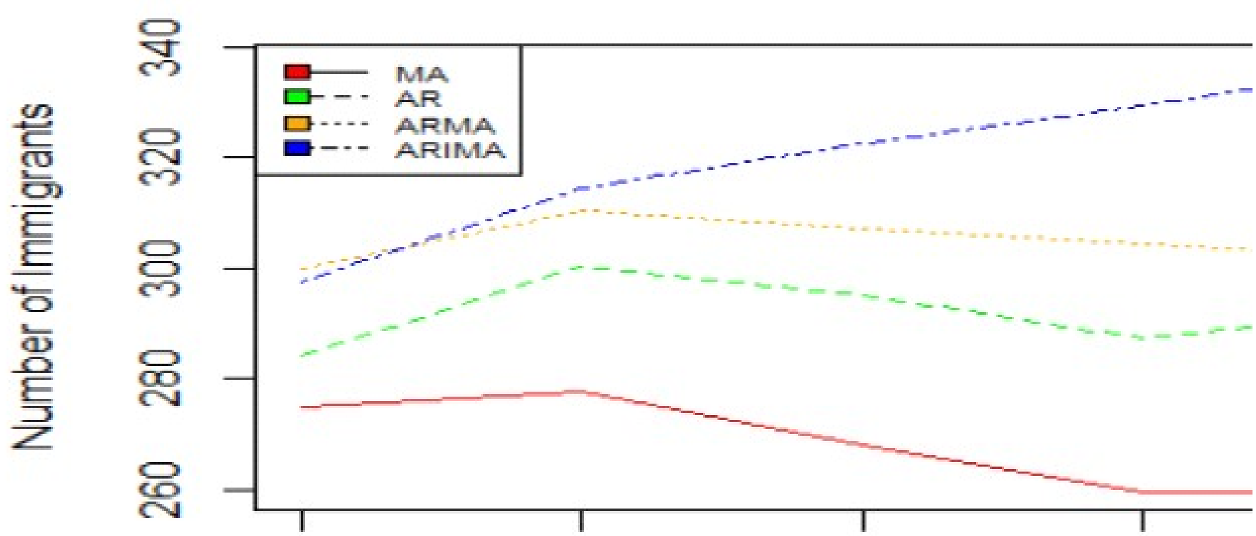

Figure 1: Comparison of the AR,MA,ARMA and ARIMA models

4.4 Diagnostic checking

The sample autocorrelation function (ACF) are useful qualitative to verify diagnostic of the model validation. The ACF of residuals (Fig 2-5) showed the goodness of fit of the models: ie. none of this autocorrelation was significantly different from zero. This demonstrated was that the chosen forecasting models were suitable for forecasting immigrants to Canada.

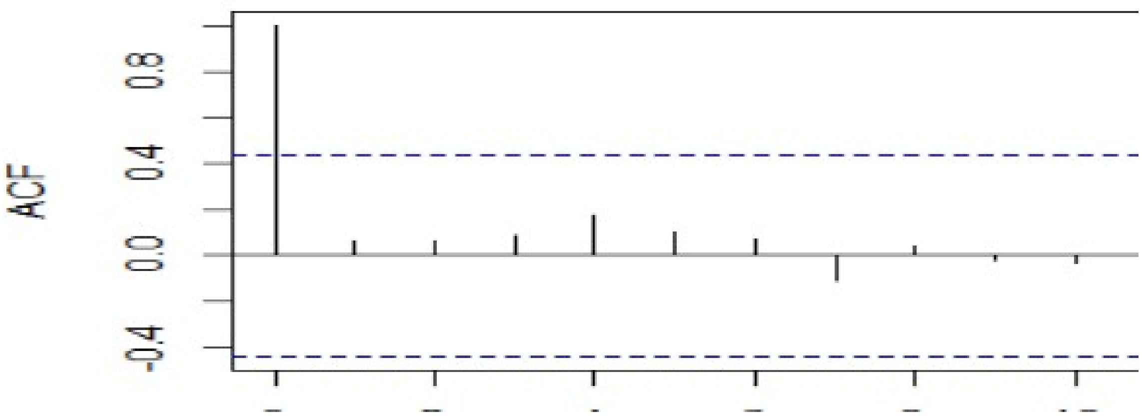

Figure 2: Auto correlation function for the MA model

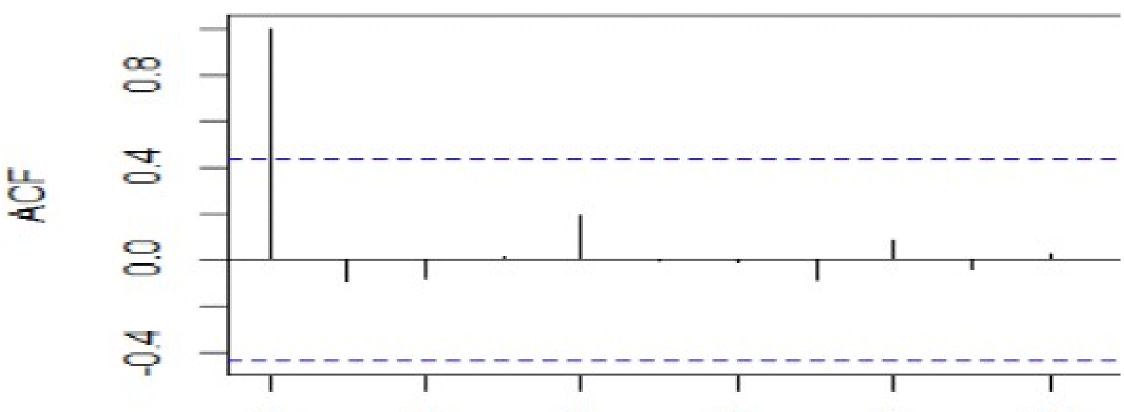

Figure 3: Auto correlation function for the AR model

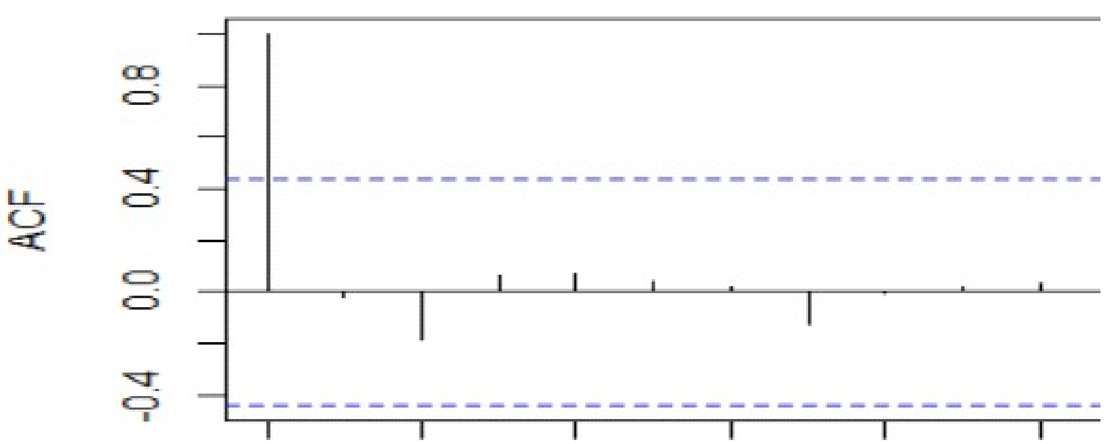

Figure 4: Auto correlation function for the ARMA model

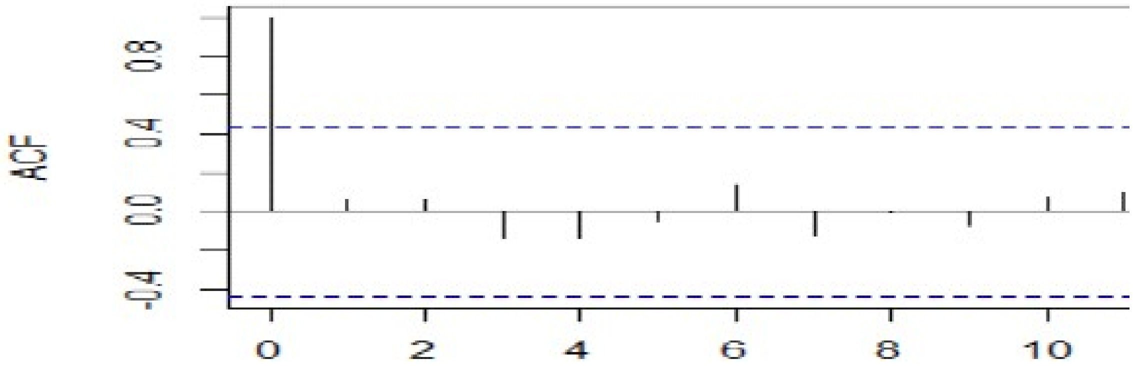

Figure 5: Auto correlation function for the ARIMA model

:

5 Conclusion

In this article, four types of forecasting models are applied to predict immigrants in Canada. The selection of an appropriate forecasting model will be identified by smallest forecast error value. The main motivation behind this research was to identify the best quality of forecasting models for the dataset, on the basis of Accuracy level and forecasting the number of immigrants for the period 2020-2024.

References

[1] Ninette Kelley and Michael Trebilcock, The Making of the Mosaic: A History of Canadian Immigration. *Policy Second Edition Ontario, University of Toronto Press*, 2010.

[2] Shaw,R.P(1975). Migration theory and fact. *Bibliography series No.5. Philadelphia Regional Science Research Institute*.

[3] Ahmed Hussen, P.C, M.C(2018). Annual Report of parliament on immigration 2018. *Minister of immigration, Refugees and Citizenship Canada*.

[4] Agrawal, U.D(2000). Population projections and their accuracy, https://www.amazon.com/Population-Projections-their-accuracy-Agrawal/dp/8176461733

[5] Keilman, N., Pham, D.Q and Hetland, A(2001). Norway's Uncertain Demographic Future, Social and Economic studies 105. *Oslo: Statistics Norway*. https://www.researchgate.net/publication/242019483 Norway's uncertain demographic future.

[6] Singh, A.P. , Gaur, M.K., Kasdekar, D.K, Agrawal, S.(2015). A study of time series Model for Forecasting of Boot in Shoe Industry. *International Journal of Hybrid information Technology,* **8.**

[7] De Beer, J(1997). The effect of uncertainty of migration on national population forecasts : the case of the Netherlands. *Journal of Official statistics*, **13**, 227-243.

[8] Box, G. E. P. and G. M. Jenkins(1976), Time series Analysis: Forecasting and Control. San Francisco: Holden-Day.

[9] J. S. Armstrong and F. Collopy, Error measures for generalizing about forecasting methods: Empirical comparisons, *International Journal of Forecasting*, **8**, 69-80.

[10] Abdullah, S. Karaman and Tayfur, Altok(2004), An Experiment study on Forecasting Using Tess Processes. Proceeding of the winter simulation conference, R. G. Ingalls, M.D. Rossetti, J. S. Smith, and B.A. Peters, eds. 437-442.
[11] Kaushal, N., Lu, Y., Denier,N., Wang, J.S.H. and Trejo, S.J.(2016), Immigrants employment and earnings growth in Canada and the USA: evidence from longitudinal data. *Springer-Verlag Berlin Heidelberg*, **29**,1249-1277.
[12] Makoni, T., Murwendo, T. D., Mawonike, R. and Chipumuro, M. (2019), Modelling and forecasting Zimbabwe's immigrants using SARIMA models. *African journal of Hospitality, Tourism and Leisure*,**8**, 1-8.
[13] Zakria, M. and Muhammad, F(2009). Forecasting Population of Pakistan using arima models. *Pakistan Journal of Agriculture Science*,**46**,214-223.
[14] Bijak, J., Disney, George, Findlay, A.M., Forster, J.J., Smith, P.W.F., Wisniowski, A. (2019) Assessing time series models for forecasting international migration, *Journal of forecasting*,**38**, 470-487.
[15] Kumari, K. (2016).Forecasting foreign tourist arrivals in India using different time series models. *International journal of Emerging Technologies in computational and Applied Sciences*,**15**,38-43.
[16] Thabani, N. (2019) . Addressing the population question in Mexico: A Box-Jenkins ARIMA approach. https://mpra.ub.uni-muenchen.de/92440/.
[17] Nelson, B.K.(1998). Statistical methodology Vs time series analysis using autoregressive integrated moving Average (ARIMA) models. *Academic emergency medicine*, **5**, 739-744.
[18] Raymer, J. and Wisniowaski, A.(2018) Applying and testing a forecasting model for age and sex patterns of immigration and emigration, *Journal of demographic, Taylor and Francis,***3**,339-355.
[19] Lim, C and McAleer, M.(2002), Time series forecast of international traveler demand for Australia. Tourism Management,**23**,389-396.
[20] Chu,F.L.(1998), Forecast tourism: a combined approach, Tourism Management, *Elsevier science,***19**,515-520.
[21] Erin Duffin(2019), Number of Immigrants in Canada, *Statcan*. ID:403063