Forecasting Canadian Immigration Patterns with Time Series Models

Thai Pham
John Joshua Bardelosa
Michael Ahana
Thompson River University, Kamloops, BC, Canada
November 24, 2024

Abstract

This project investigates Canadian immigration patterns using the CANSIM 051-0004 dataset, which includes data on immigrants by province or territory from 1971 to 2023, alongside demographic indicators such as births, deaths, emigrants, and returning emigrants. Released by Statistics Canada on September 25, 2024, the dataset provides a comprehensive foundation for analyzing and forecasting immigration trends.

The study conducts a comparative analysis of multiple forecasting models, spanning traditional time series techniques and advanced machine learning-based approaches. Model performance is rigorously evaluated using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Among the traditional methods, the Exponentially Weighted Moving Average (EWMA) model demonstrated superior performance, effectively capturing short-term variations in the data. Additionally, the hybrid model developed in this study, combining XGBoost with ARIMA, emerged as the most accurate forecasting approach overall. These results highlight the advantages of blending statistical and machine learning models, offering valuable insights into the complex dynamics of immigration trends and supporting informed decision-making for Canada's demographic planning.

Key words: Times series analysis, Dynamic regression, Hybrid method, Deep learning

I. Introduction

Forecasting immigration patterns is critical for informing policy decisions, resource allocation, and long-term planning in diverse nations like Canada. Immigration trends are influenced by a variety of demographic and socio-economic factors, including births, deaths, emigration, and the return of emigrants. Analyzing these complex relationships requires advanced forecasting models capable of capturing both linear trends and temporal dependencies within the data.

Before fitting the models, we conducted a comprehensive analysis of the data, including examinations of trends, seasonality, and other temporal patterns. Statistical analyses were also performed, such as calculating the correlation matrix of key indicators and investigating their interrelationships. These analyses provided insights into the dynamics of the variables and informed the selection of model features. To prepare the data for modeling, we applied transformations and differencing to ensure stationarity, meeting the assumptions required for time series analysis.

Building on these insights from the data analysis phase, we designed a hybrid modeling approach that combines ARIMA and XGBoost. The ARIMA component is first used to capture the linear relationships and temporal dependencies in the immigration data. Then, XGBoost is employed to model and predict the residuals from the ARIMA model, improving the overall accuracy by capturing non-linear patterns that ARIMA alone may miss. This hybrid approach leverages the strengths of both models—ARIMA's ability to model time series data and XGBoost's strength in handling complex, non-linear relationships—to address the limitations of traditional models in capturing residual dependencies or shifts in dynamics over time.

This report compares the performance of the proposed hybrid model with built-in models in R and benchmark approaches, including traditional Time Series models, dynamic Time Series models, and models such as Prophet and NNAR. The findings not only demonstrate the robustness of the hybrid method but also highlight its potential as a superior forecasting tool for complex, multi-dimensional datasets like CANSIM 051-0004.

II. Data

The dataset used in this study is from CANSIM 051-0004, covering annual immigration data for Canada from 1971 to 2023. It includes variables such as the number of immigrants, deaths, births, emigrants, and returning emigrants, sourced from Statistics Canada (released on September 25, 2024). The data can be directly accessed via the following link: CANSIM 051-0004 Dataset

Exploratory Data Analysis (EDA) was conducted to examine trends, seasonality, and correlations among variables. As shown in the correlation matrix, the number of immigrants is strongly correlated with deaths (correlation coefficient = 0.871) and moderately correlated with emigrants (correlation coefficient = 0.604) and

returning emigrants (correlation coefficient = 0.533). Births showed weak correlations with other variables, indicating limited interaction with immigration trends.

The immigration data reveals a slight upward trend over the study period, with notable fluctuations. A significant peak in immigration occurred in 2021, likely due to policy changes under the TR2PR pathway, following a slight dip in 2020, potentially caused by the impact of the COVID-19 pandemic. Despite these variations, the trend exhibits a generally mild upward movement over the decades.

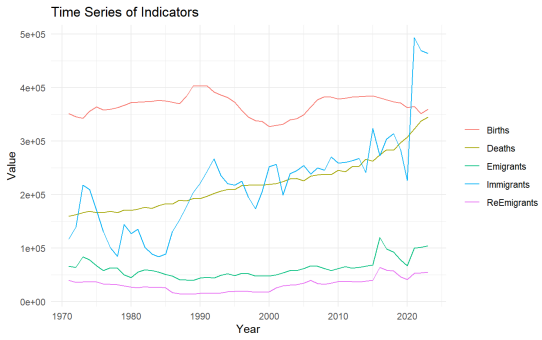

Fig. 1. Times series of indicators

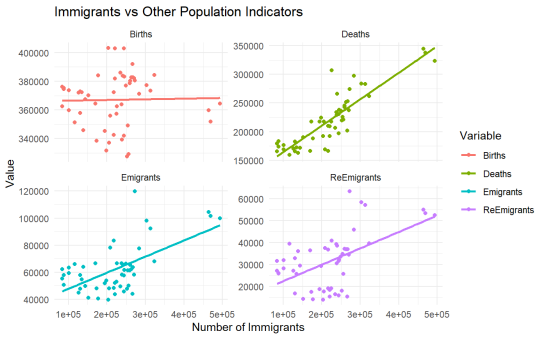

Fig. 2. Immigrants vs Other Population Indicators

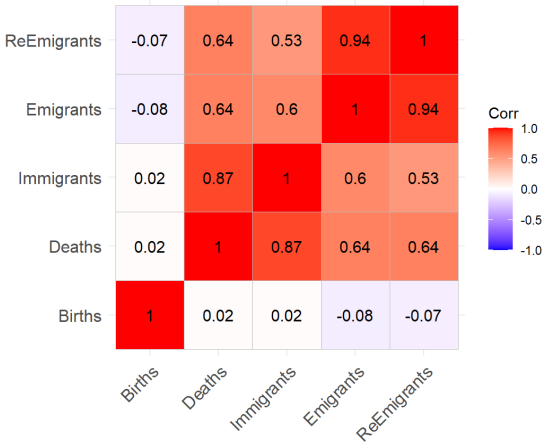

Fig. 3. Correlation matrix

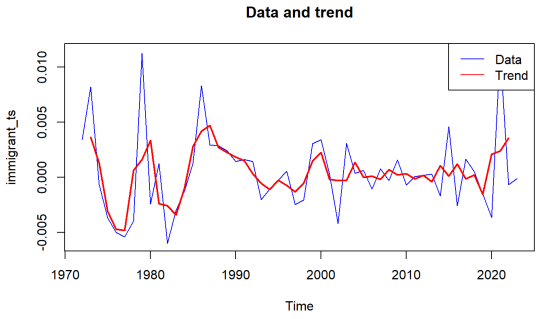

Fig. 4. Trend exploration

Fig. 5. Time Series and Trend Analysis

Time series analysis further confirmed the absence of seasonality in immigration data. The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots offered additional insights:

For the raw data (without transformation or differencing): The ACF tails off gradually up to lag 8, indicating that immigration numbers retain some level of dependency over several years. However, the gradual decline suggests that these dependencies are not strong and dissipate over time. Meanwhile, the PACF shows significant spikes at lag 1 and lag 4, meaning that immigration numbers are directly influenced by their values one year prior and, to a lesser extent, four years prior. These findings suggest that immigration numbers exhibit a moderate level of autocorrelation but without a pronounced seasonal or repetitive pattern. After applying a Box-Cox transformation followed by first differencing: Both ACF and PACF plots revealed no significant lags, indicating that the transformed data achieved stationarity. Stationarity of the dataset was also verified using the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. These transformations ensured the dataset was suitable for time series forecasting.

The final dataset, consisting of annual data from 1971 to 2023, was used to build forecasting models. Immigration numbers served as the target variable, with deaths, births, emigrants, and returning emigrants as predictors.

III. Method

In this project, we employed several built-in time series models from R to forecast immigration patterns in Canada. The first approach focuses on analyzing the time series (TS) pattern of immigration, and the

second incorporates additional explanatory variables such as Deaths, Births, and Returning Emigrants to improve forecasting accuracy. The models applied include:

- ARIMA (AutoRegressive Integrated Moving Average): ARIMA is a popular model used for forecasting time series data with trends and seasonality. It combines three main components: autoregression (AR), integration (I) to make the data stationary, and moving average (MA) to handle noise, capturing temporal dependencies in the data.
- TSLM (Time Series Linear Model): TSLM is used to model linear relationships between the target variable (Immigrants) and one or more explanatory variables (such as Deaths, Births, and Returning Emigrants). It captures both short-term fluctuations and long-term trends in the data.
- NNAR (Neural Network Autoregression): NNAR is a neural network-based autoregressive model that captures nonlinear patterns in time series data. It is particularly useful for modeling complex dynamics and non-linear dependencies in the data.
- Prophet: Prophet is a robust forecasting tool designed for time series data with multiple seasonalities and holiday effects. It captures trend and seasonal patterns effectively, making it suitable for data with irregular seasonal effects.
- EWMA (Exponentially Weighted Moving Average): The EWMA model applies exponential weighting to past observations, giving more importance to recent data while smoothing out historical trends. It is commonly used to capture short-term trends and fluctuations in time series data.

These models were first applied to the univariate time series of immigration data (Immigrants) to identify inherent patterns such as trend and seasonality. Subsequently, the models were extended to include the explanatory variables (Deaths, Births, Returning Emigrants) to enhance forecasting accuracy by capturing the effects of these variables on immigration trends.

To enhance forecasting performance, we proposed a hybrid model combining ARIMA and XGBoost. XGBoost is a powerful machine learning algorithm widely applied in time series forecasting due to its ability to model non-linear relationships and complex patterns in data. By utilizing gradient boosting on decision trees, XGBoost effectively captures intricate dependencies between lagged features and target variables, making it particularly suitable for refining residuals. In this hybrid approach, ARIMA models the linear trends and temporal dependencies in the immigration data, while XGBoost predicts the residuals, addressing non-linear patterns that ARIMA may not capture. The steps to construct this hybrid model are as follows:

- **Step 1**: Fit an ARIMA model to the immigration data, using explanatory variables such as Deaths, Births, and Returning Emigrants to capture linear relationships and temporal dependencies. This step provides an initial forecast of the immigration series.
- **Step 2**: Extract the residuals from the fitted ARIMA model. These residuals represent the portion of the data that the ARIMA model could not explain, including complex non-linear patterns.
- **Step 3**: Fit an XGBoost model to the residuals from the ARIMA model. The XGBoost model captures any remaining non-linear dependencies within the residuals, enhancing the prediction accuracy.
- **Step 4**: Combine the forecasted values from the ARIMA model with the forecasted residuals from the XGBoost model. This combined forecast improves the accuracy of the overall prediction by accounting for both linear trends and non-linear dependencies.

IV. Result and Discussion

The time series data was preprocessed to achieve stationarity prior to model fitting. This involved applying a Box-Cox transformation and first differencing to address non-stationarity. After preprocessing, most models demonstrated strong performance, as evaluated by cross-validation metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Forecast accuracy across the models was generally consistent, with only slight variations observed.

Among the baseline models, the Exponential Weighted Moving Average (EWMA) showed the best performance. With an optimized smoothing parameter ($\alpha = 0.06$), it achieved an RMSE of 0.00366 and an MAE of 0.00239. These results highlight EWMA's effectiveness in capturing short-term trends while minimizing noise in the data.

The hybrid model that combines ARIMA and XGBoost achieved the best results, with the lowest RMSE of 0.00187 and MAE of 0.00153. This performance underscores the power of combining linear and nonlinear modeling techniques. By integrating ARIMA's ability to capture linear patterns with XGBoost's strength in handling complex nonlinear relationships, the hybrid model effectively captured the dynamics of the immigration data. Additionally, its use of multivariate indicators, such as deaths, births, emigrants, and returning emigrants, significantly boosted the accuracy of immigration number forecasts.

A comparison was conducted between models trained on preprocessed data and those trained on raw, unprocessed data, which only utilized the single immigration time series without incorporating multivariate indicators. The results clearly demonstrated significant improvement in performance when data preprocessing and multivariate indicators were included. The table below presents a summary of the RMSE and MAE for models fitted on both preprocessed and raw data:

TABLE I
RMSE AND MAE FOR MODELS FITTED ON PREPROCESSED DATA

Model	RMSE	MAE
Mean	0.003673857	0.002282767
EWMA (optimal alpha = 0.06)	0.003663649	0.002387911
LM w/ ARIMA(0,0,0) errors	0.004522102	0.003010401
TSLM with knot	0.004838149	0.003541369
NNAR (1,2)	0.00770	0.004525957
Prophet	0.006672516	0.005570481
XGBoost	0.004156153	0.003105009
Hybrid ARIMA-XGBoost	**0.001874198**	**0.001530026**

TABLE II
RMSE AND MAE FOR MODELS FITTED ON RAW DATA

Model	RMSE	MAE
Mean	73742.25	67079.81
EWMA (optimal alpha = 0.3)	124727.3	145292.7
ARIMA(0,1,0)	51530.71	40800.73
TSLM	134232.6	117664.3
NNAR(1,1)	170621.6	140109.5
Prophet	134367.3	117763.6

All related project documentation and resources can be accessed via the following: GitHub link

V. FUTURE WORK

While this project demonstrates the effectiveness of the hybrid ARIMA-XGBoost model and highlights the benefits of data preprocessing, there are several avenues for further exploration.

- Incorporating Additional Features: Although this study utilized multivariate time series indicators such as births, deaths, emigrants, and returning emigrants, other socio-economic and policy-related variables could be integrated to further enhance model performance. Variables such as unemployment rates, political stability, or migration policies could potentially improve forecasting accuracy for immigration data.
- Model Optimization: While the hybrid ARIMA-XGBoost model performed well, further optimization of hyperparameters through more advanced techniques such as Bayesian optimization or genetic algorithms could be explored to fine-tune the model's performance.
- Model Comparison with Advanced Deep Learning Approaches: Although traditional machine learning models like XGBoost performed well, future studies could explore deep learning models, including LSTM (Long Short-Term Memory) networks, which are particularly suited for time series forecasting. A comparison between these models and the current ones could provide insights into their relative strengths and weaknesses.
- Incorporating External Shocks: The impact of external events such as the COVID-19 pandemic has been shown to disrupt immigration patterns. Future work could focus on modeling such shocks using intervention models or regime-switching models to capture sudden changes in trends.
- Model Deployment and Real-Time Forecasting: Future work could also focus on the deployment of the developed models for real-time forecasting. Implementing a dynamic, real-time forecasting system would allow for adaptive predictions that can account for new incoming data and sudden changes in immigration trends.
- By addressing these areas, future research could further improve the accuracy and applicability of models for immigration forecasting, providing valuable insights for policymakers and other stakeholders.