ACENET
Microcredential in Advanced Computing
ISP Report *Template*

Project title: Forecasting Family Physician Availability: Trends and Dynamics in New Brunswick and Canada

Participant name: Marie-Line Forbes

Date: July 31, 2024

Abstract:

This project forecasts the future availability of family physicians in New Brunswick using ARIMA modeling, covering 52 years of historical data. The ARIMA model demonstrated effectiveness in short-term predictions, with parameter optimization highlighting the importance of combining automated and manual approaches.

1. Introduction

The availability of family physicians is essential for maintaining effective healthcare services and public health. This project aims to look at the challenges associated with the shortage of family physicians in New Brunswick by analyzing historical data and forecasting future trends. "How can historical data on family physicians in New Brunswick be used to forecast future availability, and how do these forecasts compare with other provinces across Canada?"

2. Background

In New Brunswick, as in many other regions, there have been growing concerns about the availability of family physicians. According to a recent survey by the New Brunswick Health Council, only 79% of New Brunswickers had access to a permanent family doctor or nurse practitioner in the past year, a decline from 85% the previous year. This shortage of accessible primary care has led to increased reliance on emergency departments, walk-in clinics, and other services, contributing to fragmented care and potentially worsening health outcomes, particularly for individuals with chronic conditions. Factors such as an aging workforce, retirements, and the increased complexity of patient cases have been identified as contributing to this decline in primary care access.

This research focuses specifically on New Brunswick, utilizing 52 years of historical data on family physicians from 1971 to 2022 to understand and forecast physician availability. Although the primary emphasis is on New Brunswick, the dataset also includes information from other Canadian provinces. This comparative analysis allows for a broader context and helps to benchmark New Brunswick's trends against national patterns. Advanced time series analysis and machine learning techniques, specifically ARIMA models, are employed to generate forecasts for future physician numbers and to compare these predictions with those from other provinces.

3. Analysis

To conduct this research, I utilized the dataset titled "Supply, Distribution and Migration of Physicians in Canada, 2022 — Historical Data," sourced from the Canadian Institute for Health Information. This dataset covers 52 years and provides comprehensive information on the supply and distribution of physicians in Canada, categorized by specialty. It includes demographic, educational, and migration details.

For the purposes of this project, I created two specific dataframes from the dataset:

1. family_medicine_df_nb: This dataframe contains information solely about family physicians in New Brunswick.
2. family_medicine_df_canada: This dataframe includes aggregated information about family physicians across all provinces in Canada.

Before applying machine learning models, I conducted an exploratory analysis to identify key trends and insights from both dataframes. This initial exploration aimed to determine whether there were notable differences between New Brunswick and the national average. Specifically, I compared the gender distribution, average age, and physician-to-100,000 population ratio of family physicians in New Brunswick against the Canadian benchmark. The analysis revealed that these metrics in New Brunswick were comparable to the national average, with no significant anomalies observed.

Moving on to the machine learning model, the ARIMA (AutoRegressive Integrated Moving Average) model will be initially applied to the family_medicine_df_nb dataframe. The goal is to forecast the future number of family physicians in New Brunswick, leveraging the 52 years of annual time series data available. The ARIMA model is well-suited for this task as it captures patterns and trends in historical data to predict future values. The ARIMA model requires the definition of three parameters: (*p, d, q):*

- *p*: The order of the Autoregressive part of ARIMA.
- *d*: The degree of differencing.
- *q*: The order of the Moving Average part.

Before applying ARIMA, it is essential to ensure that the time series is stationary, meaning that the mean, variance, and autocorrelation remain constant over time. Stationarity is tested using the Augmented Dickey-Fuller Test (ADF), which helps determine the appropriate value for the d parameter if the series is found to be non-stationary (0 if stationary) . The ADF test results on the family_medicine_df_nb dataframe revealed that the time series is non-stationary. Consequently, one differencing transformation was applied to the dataset, and the ADF test was performed again. After one differencing, stationarity was achieved, indicating that *d* is equal to 1.

To determine the values for the p and q parameters, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were analyzed:

The ACF plot showed significant autocorrelation at lag 1 (x-axis 0) but no significant autocorrelation at other lags, indicating a spike that extends beyond the confidence interval only at lag 1. This suggests that q should be set to 1.

The PACF plot displayed a significant spike at lag 1 (x-axis 0), indicating direct correlation with the previous observation but no significant correlations at other lags. This suggests that p should be set to 1.

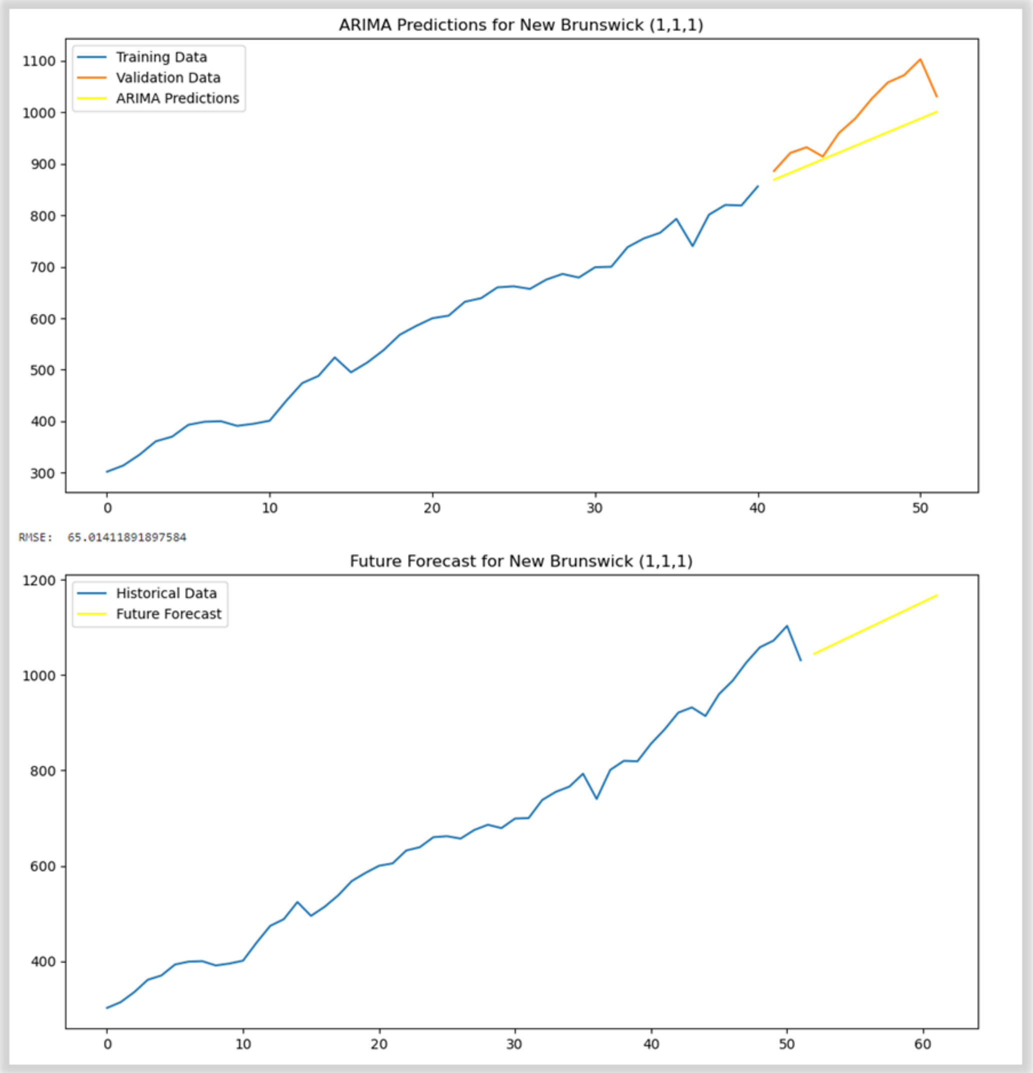
Based on these interpretations, the ARIMA model parameters are determined to be: p = 1, d = 1, and q = 1. I will share the results of theses parameters in the result section. It is also possible to automate the process of finding the optimal parameters (p, d, q) using a grid search approach. This method will be employed to explore various parameter options for the family_medicine_df_nb dataframe.

For the High-Performance Computing (HPC) component, a Python script will be used to loop through each province and territory, perform grid search for the best parameters, and store the results. The script will also generate predictions and future forecasts for each province and territory, leveraging HPC resources to handle the computational demands efficiently.

4. Results

Figure 1: displays two graphs based on the family_medicine_df_nb dataframe, modeled with ARIMA parameters (1, 1, 1):

- Graph 1: This line chart shows the ARIMA model fitted to both the training and validation data.
- Graph 2: This line chart presents the forecasted number of family physicians from 2023 to 2032. The ARIMA model predicts a steady increase in the number of family physicians in New Brunswick. The forecasted values indicate a consistent upward trend, with the number of family physicians expected to reach approximately 1873 by 2032



```
RMSE:  65.01411891897584
```

Figure 2 displays a future forecast line chart based on the family_medicine_df_nb dataframe, modeled with the ARIMA parameters (1, 2, 0) , RMSE 27.919699.

The Grid Search results indicate that the parameter *d* is set to 2, which may suggest over-differencing. This choice leads to a forecast showing a step-like declining trend. This demonstrates the importance of complementing automated parameter selection with manual testing. While Grid Search is a valuable tool, it's essential to validate and fine-tune parameters to ensure realistic and meaningful forecasts. This approach helps avoid overfitting and aligns the model more closely with expected trends.

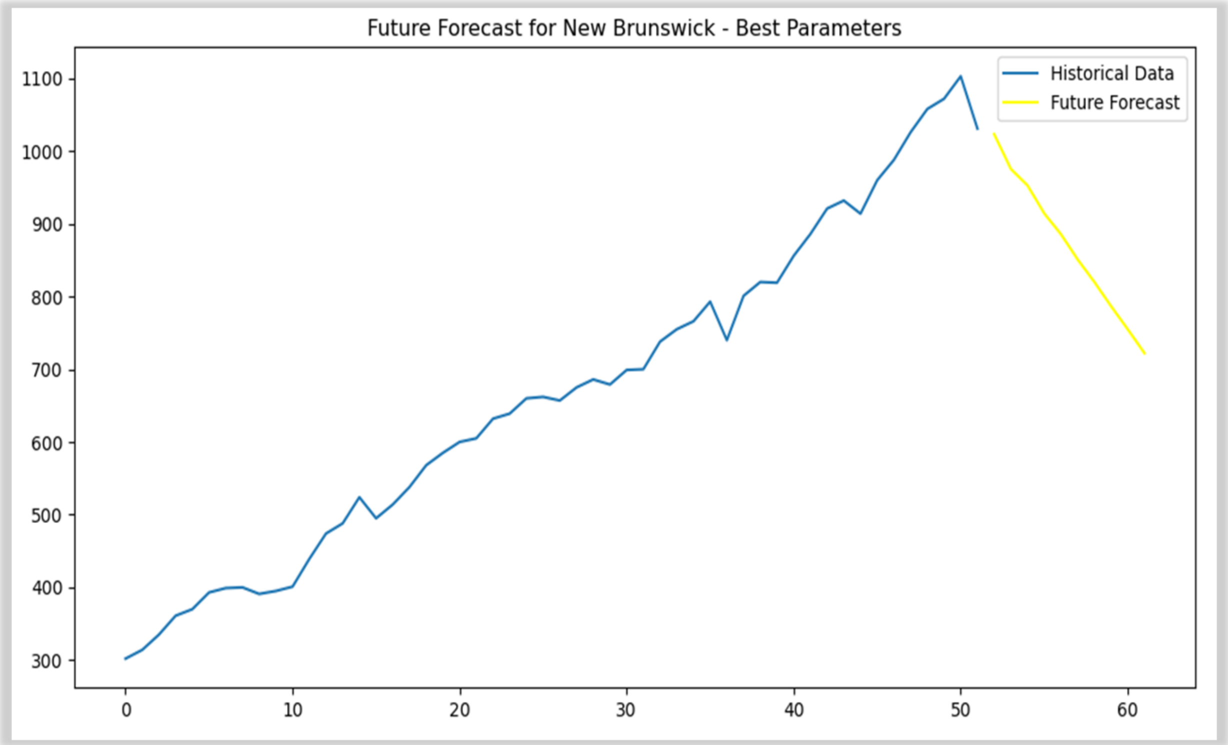
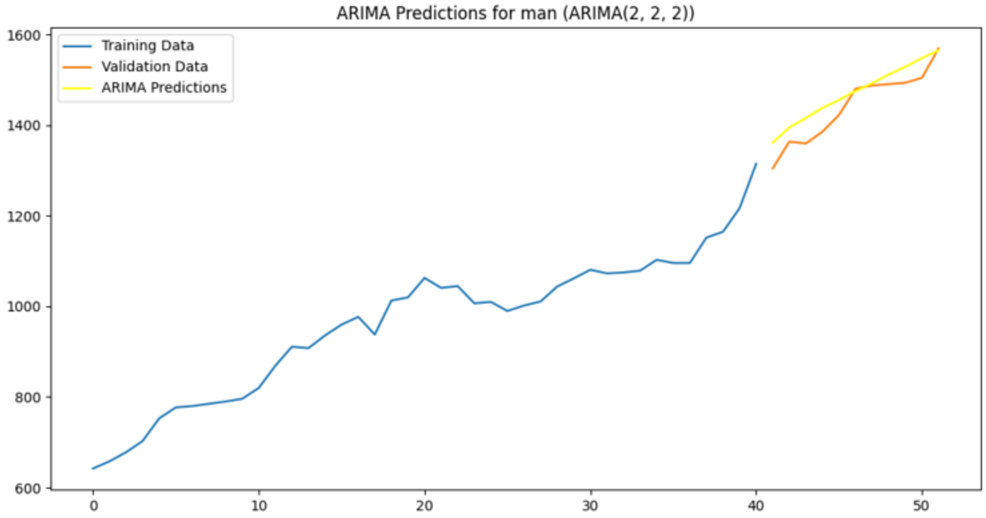
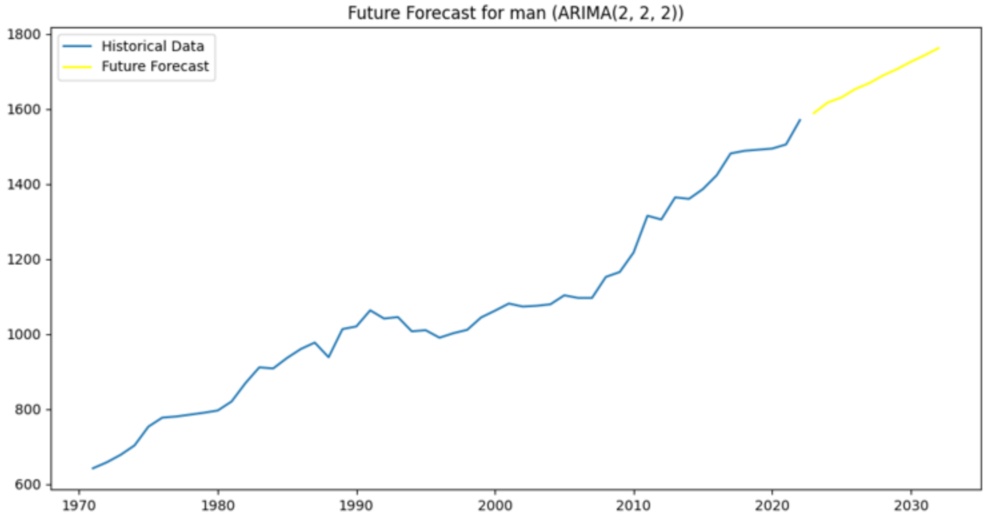


Figure 3 displays two graphs for Manitoba based on HPC results: ARIMA Predictions and Future Forecasts. This example illustrates the forecast results for Manitoba. The ARIMA model with the optimal parameters (2, 2, 2) provides a reliable forecast for the future number of family physicians in the province, demonstrating effective modeling and prediction performance.



RMSE=36.782312015979485



5. Discussion

The project aimed to forecast the future number of family physicians in New Brunswick using ARIMA modeling. While the ARIMA model with parameters (1, 1, 1) was successful in providing reasonable forecasts, a few challenges were encountered throughout the process.

- Feature Selection vs. Model Choice: Significant time was spent on selecting relevant features. In hindsight, prioritizing model choice could have streamlined the research process.
- ACF and PACF Interpretation: Determining the optimal q and p parameters from ACF and PACF plots proved difficult, affecting model accuracy.
- Parameter Optimization: Finding the best ARIMA parameters involved extensive trial and error. Automated methods can help but may not always be precise. Manual testing is crucial for improving model accuracy.
- DataFrame Size: The size of the DataFrames, especially when filtered by province, raised concerns about model robustness and forecast reliability.

Despite these challenges, the project provided valuable insights into time series forecasting and the application of ARIMA models in predicting the number of family physicians. The results demonstrate the model’s capability to offer short-term forecasts, though there is room for improvement and further exploration.

Conclusion

The ARIMA model effectively forecasted the short-term trends in family physician numbers in New Brunswick, demonstrating its utility for predicting future values based on historical data. The model's parameters (1, 1, 1) yielded reasonable forecasts, though the process of optimizing these parameters required considerable trial and error. This highlighted the need for manual validation in addition to automated methods to achieve precise results.

However, the ARIMA model relies on the assumption that future trends will mirror past patterns, which may overlook significant changes in healthcare policies or other external factors. To address these limitations, future research should consider expanding the model to include additional medical specialties within the dataset and exploring alternative forecasting methods. These steps could improve accuracy and provide a more comprehensive understanding of healthcare workforce trends.

References

[1] CBC News, “Only 79% in N.B. have access to permanent primary-care provider, survey finds” cbc.ca. https://www.cbc.ca/news/canada/new-brunswick/new-brunswick-health-council-primary-care-survey-2023-doctor-five-days-1.7238089 (Consulted July 20, 2024)

[2] “Built In,” “Time Series Forecasting with Python: A Comprehensive Guide,” builtin.com. https://builtin.com/data-science/time-series-forecasting-python (Consulted July 8, 2024)

[3] “Towards Data Science,” “Interpreting ACF and PACF Plots for Time Series Forecasting,” towardsdatascience.com. https://towardsdatascience.com/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c (Consulted July 7, 2024)

[4] “Machine Learning Mastery,” “Grid Search for ARIMA Hyperparameters with Python,” machinelearningmastery.com. https://machinelearningmastery.com/grid-search-arima-hyperparameters-with-python/ (Consulted July 7, 2024)

[5] Alkaline ML, “Tips and Tricks for pmdarima,” alkaline-ml.com. https://alkaline-ml.com/pmdarima/tips_and_tricks.html (Consulted July 17, 2024)

Supplementary Materials

https://github.com/MLForbes01/ISP-ACENET

Dataset: Canadian Institute for Health Information. Supply, Distribution and Migration of Physicians in Canada, 2022 — Historical Data. Ottawa, ON: CIHI; 2023.

https://www.cihi.ca/sites/default/files/document/supply-distribution-migration-physicians-in-canada-2022-data-tables-en.xlsx