Predicting Canadian Population Growth Through Machine Learning

Aaron Creor

Computer Science
Department of Natural Sciences
The University of Texas at Austin

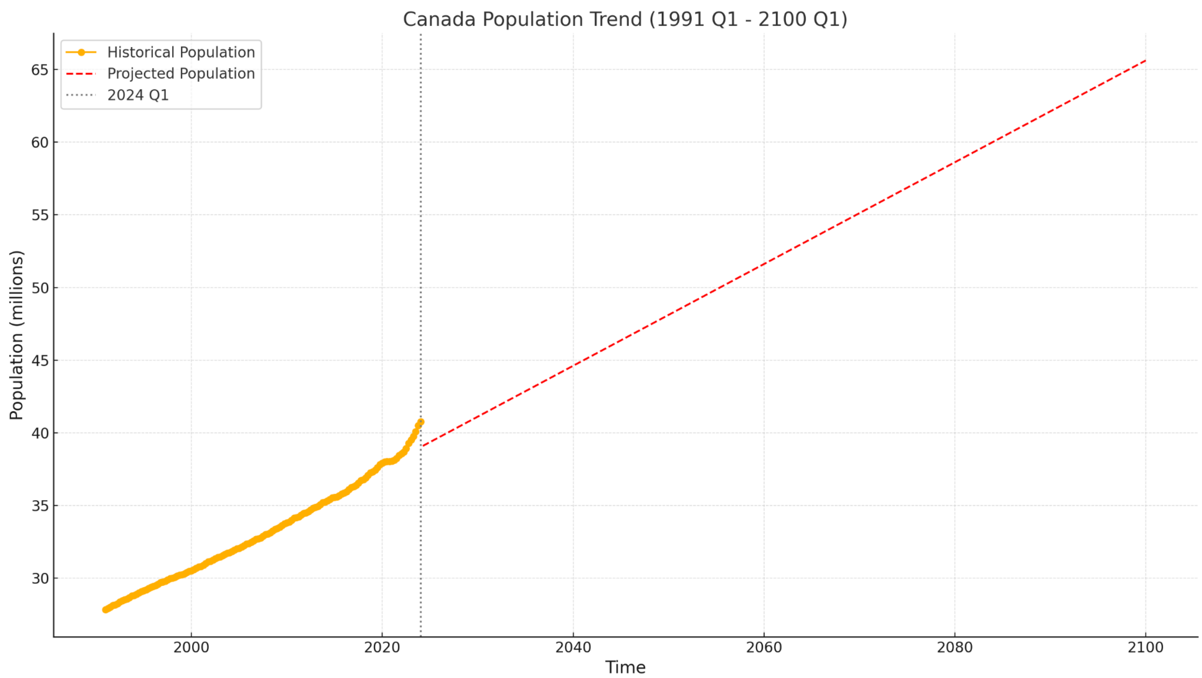
November 2024

1. Introduction

The purpose of this paper is to predict, with a high degree of confidence, Canadian population trends leading up to the year 2100. Significant literature exists on the topics of world population growth, and to a lesser extent Canadian population growth. 2100 was chosen as a date as it's common among other papers in the field, and is after the predicted world population peak in 2064 (Vollset et al., 2020). Canada's population growth is almost solely the result of immigration, being responsible for 97.6% of its growth in 2023 (Statistics Canada, 2024). With the world's predicted population decline being the result of decreasing fertility rates, a contrast will be made between Canada's immigration dependent growth, and the world's fertility driven decline (Aitken, 2022).

Predicting Canadian population growth involves many data points beyond a simple trend line produced from yearly population values. Producing an estimate based solely on quarterly population values, from Q1 1998 to Q1 2024 provides the result seen in Figure 1 (Statistics Canada, 2024). This is too simplistic and doesn't take into account trends in areas such as immigration, mortality, and fertility rates, which are critical when predicting population growth (Miladinov, 2021).

Figure 1



Simple Canadian Population Trend Line

Therefore this research is focused on predicting Canadian population growth based on a wide set of factors, with a high degree of confidence, beyond a simple trend line. We will compare individual provinces and territories to see where growth is concentrated, and analyze past anomalies in Canadian population trends. Different models will be compared to help determine what is most appropriate in predicting Canadian population growth.

2. Background

Modeling future population growth is something that has engaged scientists, statisticians, and governments for generations. A major organization pushing for these projections early on, at least globally, has been the United Nations, with their research done on long term population projections (McNicoll, 1992). Who provided predictions for the years 2050, 2100, and 2150, which are regularly revised. One of the earliest serious attempts at world population prediction, which inspired future study in the field, can be attributed to William Petty, who explored population growth in terms of "doubling time" (Petty & Morley, 1888).

Adjacent to these world population growth predictions were early models of national population growth. One example of this is Thomas Robert Malthus who predicted English population growth using geometric progression in the early 19th century (Malthus, 1826). Later examples include US census projections which use the Cohort-Component method of predicting population growth in the United States, as shown in (1) (Hollmann, 2022). Specific interest should be taken in this model as it takes into account the key factors of population change, fertility, mortality, and migration (Smith et al., 2013). Canada also uses the same system for its population predictions (George, 2001).

$$P(t + n) = P(t) + B(t) - D(t) + I(t) - E(t) \tag{1}$$

Where P is population at time t over period n
B and D are the number of births and deaths between t and n
I and E are the number of immigrants and emigrants between t and n

Looking at Canada specifically, the largest source of population data and predictions is Statistics Canada, an agency formed in 1971 with the aim of producing statistics and reports to better understand Canadian demographics, economy, resources, and culture (Government of Canada, 2024). Stats Canada currently has predictions for Canada population growth up to 2073, and provincial/territorial predictions up to 2048. There are multiple predictions based on different growth scenarios, such as slow aging, fast aging, high growth, low growth, and various levels of medium growth (Government of Canada, 2024).

3. Data & Sources

The data for this paper is sourced from Statistics Canada, who have current and historical population data at a national, provincial and municipal scale. For the purpose of this paper data from between Q1 1998 to Q1 2024 is used (Statistics Canada, 2024). Immigration data is available from 1971 to 2024, however only the data from 1998 to 2024 is used (Statistics Canada, 2024). Mortality data is available from Q1 2020 to Q1 2022, all of this data is used (Statistics Canada, 2023). Number of births monthly from 1991 to 2023 is available, all data including and after 1998 is used (Statistics Canada, 2024). We use emigration data which exists on the same Statistics Canada dataset as the immigration data, available between 1971 to 2024, but limited to 1998 to 2024 (Statistics Canada, 2024). Finally Canadian GDP growth between the years 1998 and 2023 is included, with data available as far back as 1981 (Statistics Canada, 2024). This is included as population growth could be correlated with GDP growth using an econometric model of population growth (Tung, 1984).

The dataset includes the following files:

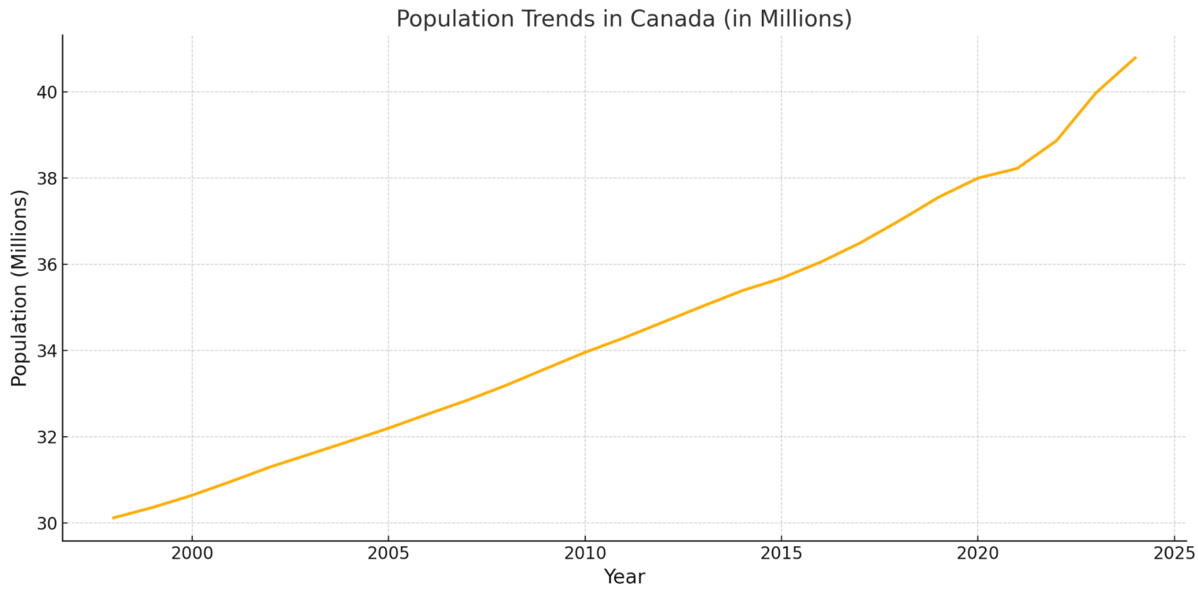
- population_history_canada.csv
- immigration_history_canada.csv
- mortality_history_canada.csv
- birth_history_canada.csv
- emigration_history_canada.csv
- gdp_history_canada_current_prices.csv

The combination of these datasets provides enough information to attempt modeling Canadian population growth using several systems, including the Cohort-Component method, microsimulation models, and econometric models.

3.1 Data Exploration

Before predictions are made we want to gain a general understanding of the data we're working with so we can best utilize it. Thus we will start with some exploration of our given data. Consider Figure 2, which shows Canadian population growth since 1998.

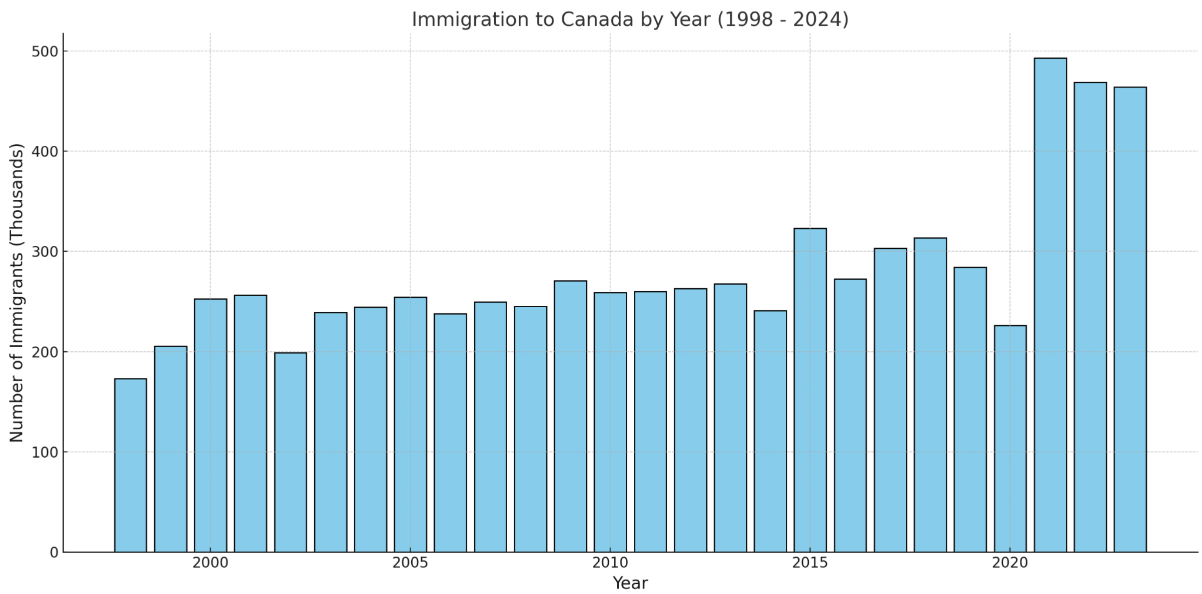
Figure 2



Historical Canadian Population

Now consider Figure 3 which shows the total number of immigrants coming into Canada each year. Notice the dramatic increase in the years following 2020.

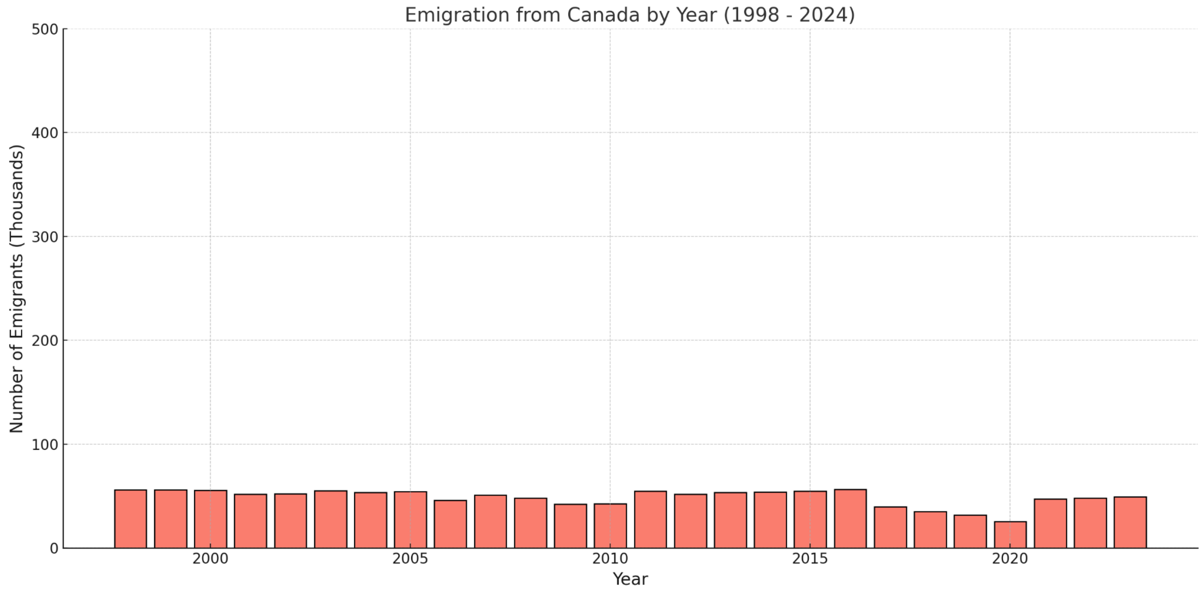
Figure 3



Historical Canada Immigration

Next, Figure 4 shows historical emigration data for Canada post 1998. When comparing with the previous graph we can see that many more people are coming to Canada than leaving.

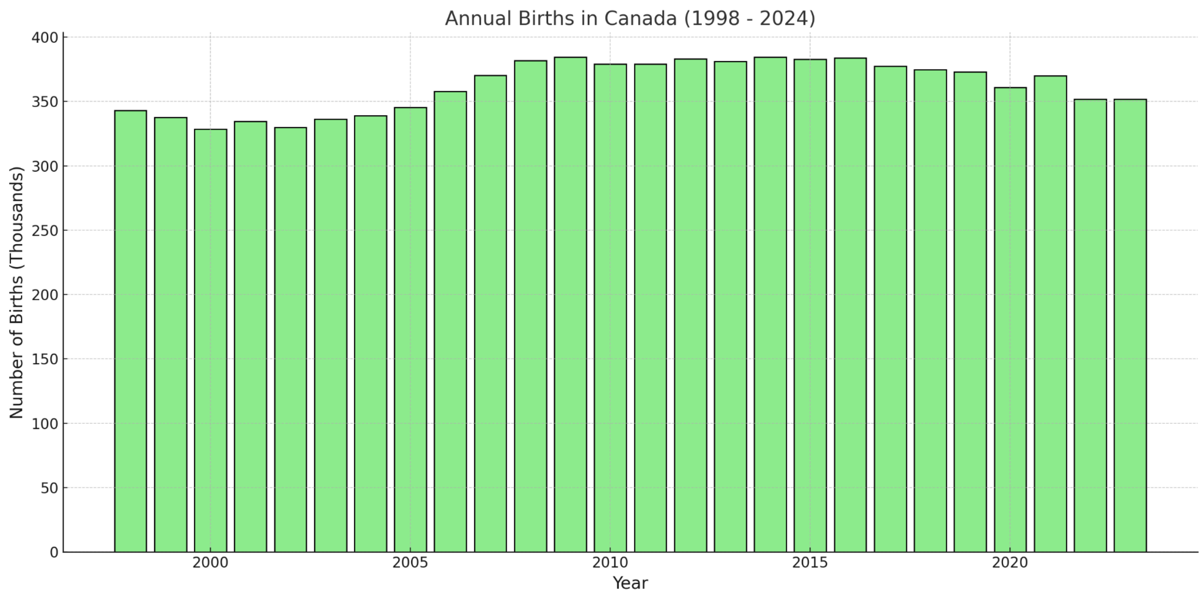
Figure 4



Historical Canada Emigration

Figure 5 demonstrates the history of Canadian birth rates since 1998. We can see that the number of births is relatively equal despite an upwards trend in population. Canada matches the trend in other developed nations in having a low fertility rate (Brauner-Otto, 2016). Canada currently receives more immigrants than newborn babies (Serebrin, 2024).

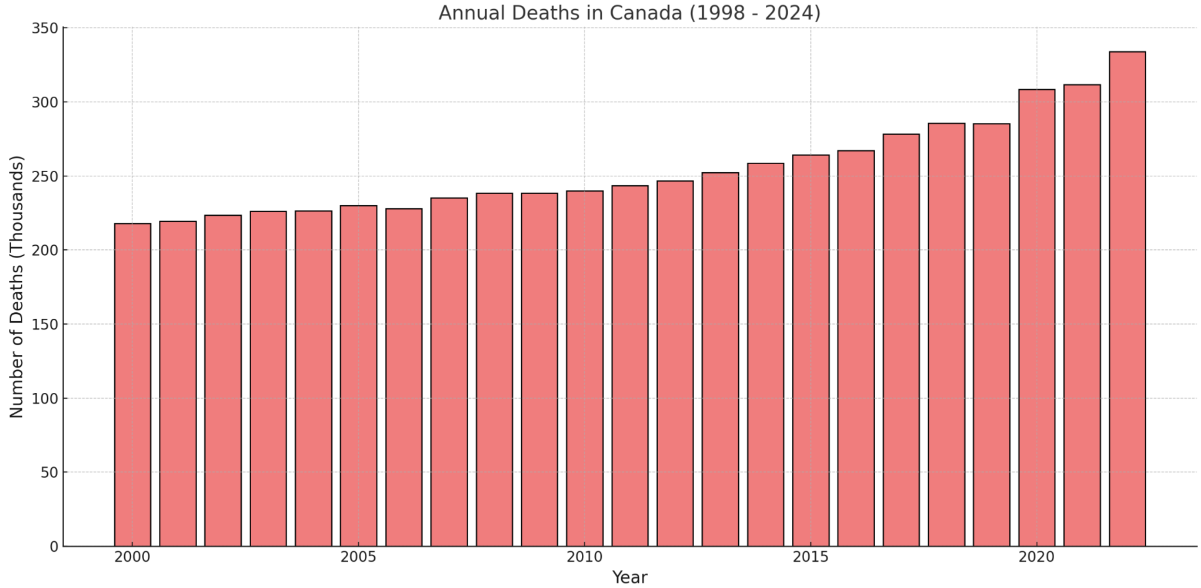
Figure 5



Historical Canadian Fertility

Figure 6, however, shows Canadian mortality rates since 1998. A clear increase can be seen during the Covid-19 Pandemic.

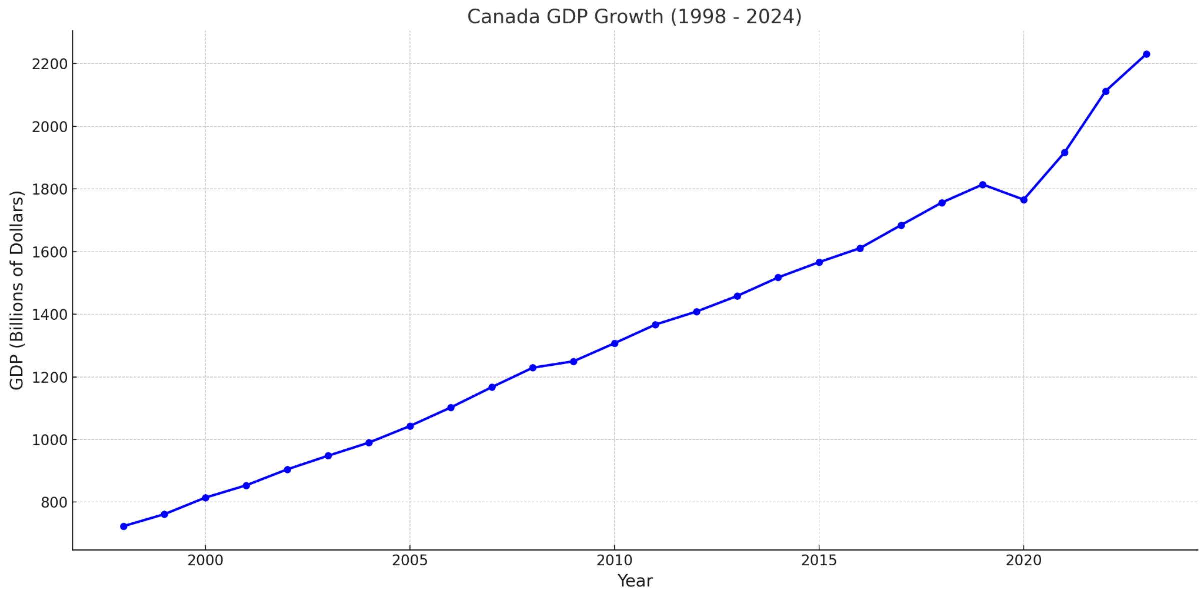
Figure 6



Historical Canadian Mortality

Finally we look at Figure 7 which shows Canadian GDP growth since 1998. Again the Covid-19 Pandemic is clearly visible on this graph. Data provided is in 2024 USD.

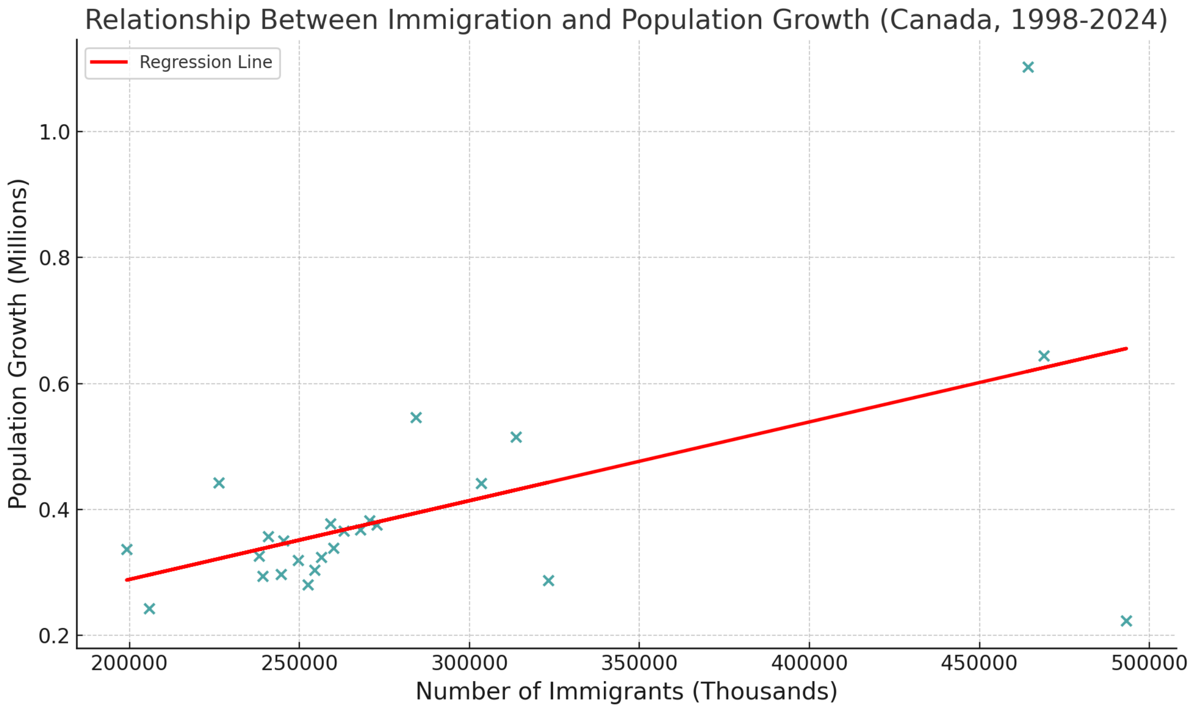
Figure 7



Historical Canadian GDP Growth

Now we can take an opportunity to find some relationships between this data. Figure 8 shows a moderate positive relationship between immigration and population growth in Canada. As the number of immigrants to Canada has increased, so has its rate of population growth. The correlation coefficient between these when using a regression model is 0.554.

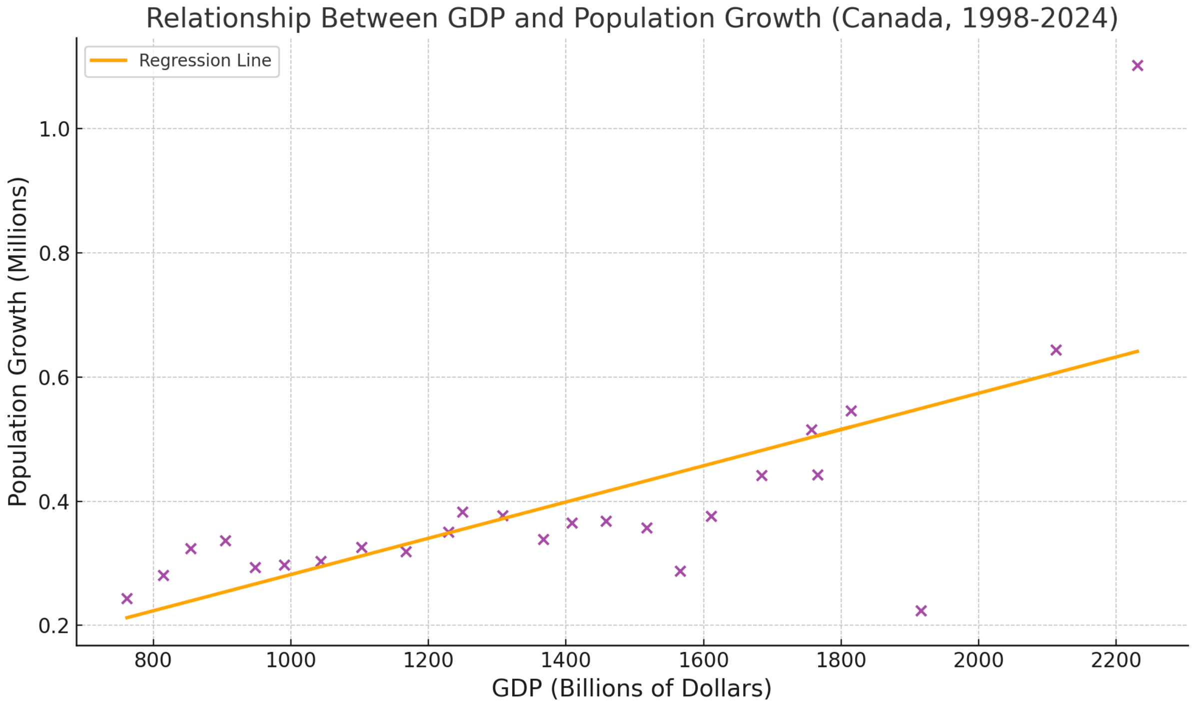
Figure 8



Immigration and Population Growth in Canada

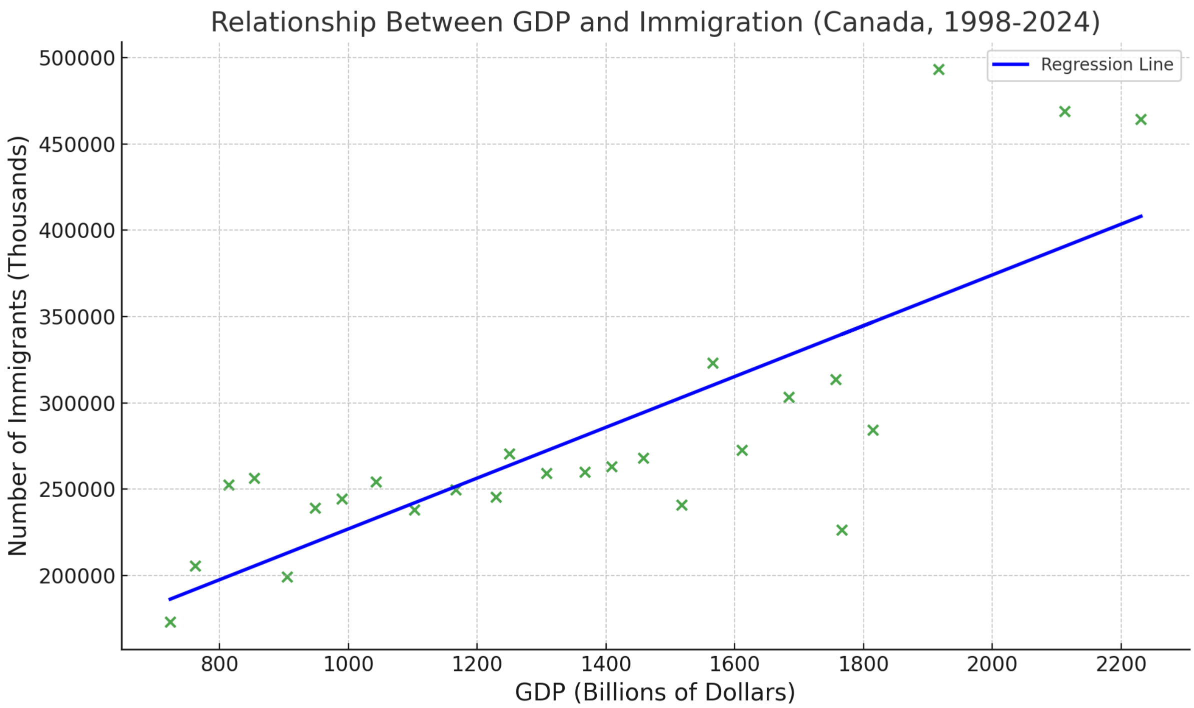
Figure 9 and 10 demonstrate the relationship between GDP growth and population growth, along with GDP growth and immigration. Showing the close relationship between the three (Previl, 2024). The correlation coefficient between GDP and population growth is 0.685, while the coefficient between GDP and immigration is 0.786, when using a regression model.

Figure 9



Canadian GDP growth and population

Figure 10

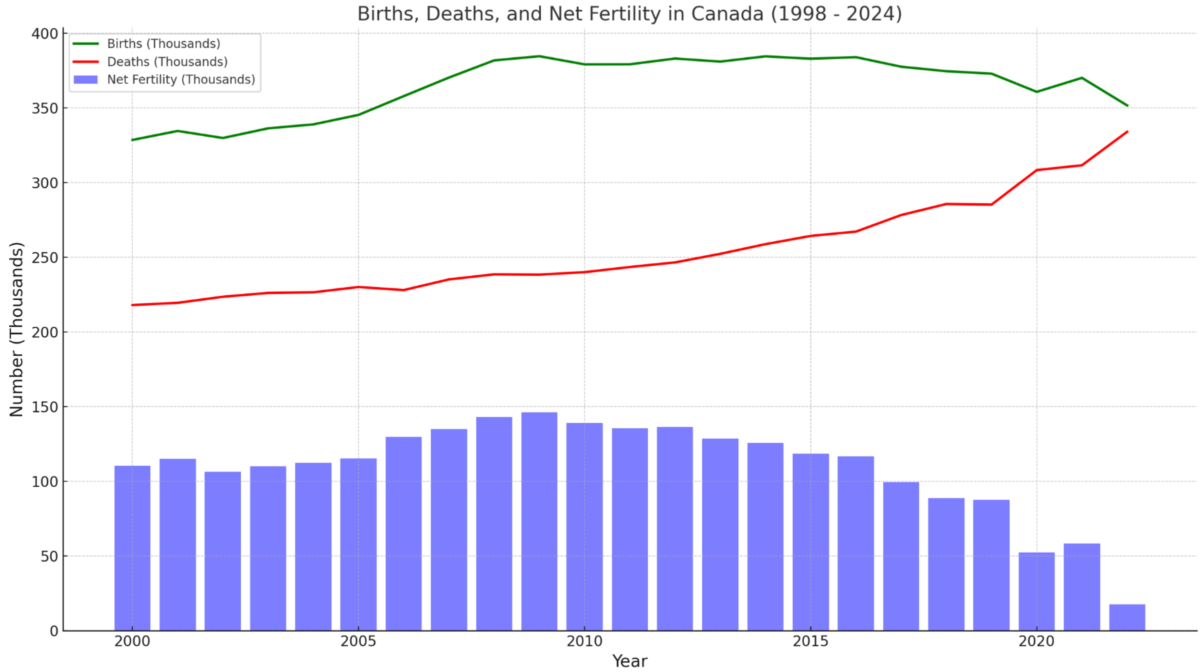


Canadian GDP growth and immigration

3.2 Takeaways

After considering the dataset for this paper, some clear trends and relationships in the data become clear. For example, GDP growth and immigration is closely related in Canada, which is backed up by existing research (Akbari & Haider, 2017). Additionally, and by extension, Canadian population growth and GDP are also related (Hosen, 2019). We see that immigration to Canada is a much more significant factor than emigration, with the rate being approximately ten times higher (Beaujot & Rappak, 1989). Finally, consider Figure 11, where we see that Canadian fertility is nearly negative, with mortality rates increasing and fertility rates decreasing. In recent years Canadian fertility has hit a record low (Aziz, 2024).

Figure 11



Historical Canadian Fertility

4. Methods

Random Forests Regression

The first model we will use to predict future Canadian population growth is the Random Forests method. The Random Forests method has been used in other research to model population data (Stevens et al., 2015). It is effective in instances where we are dealing with complex and diverse datasets, with a variety of features. Here we will expand on the features in our data set, and our results with the Random Forests method. See Table 1 for features and their importance scores.

Table 1

Feature Importance Scores for Random Forests Regression

Feature	**Importance**
Emigrants	0.322129
Deaths	0.228941
Net Fertility	0.180067
Net Migration	0.117017
Immigrants	0.103784
Births	0.048063

Using the Random Forests Regression we predict Canada will have a population of 66.25M in the year 2100. This is based on data from Q1 1998 through Q1 2024. In addition to using raw birth/death and immigration/emigration values we combine these into net fertility and net migration. The target variable is population growth, or the yearly change in population between years which is used to visualize or data. Figure 12 demonstrates the visualized results. When training the model we split our data into two parts, the training set and test set, which were used to train and test the model. Model performance was measured using the mean squared error (MSE) which shows how close our predictions were to the actual data points. The equation for MSE can be seen in (2).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2)$$

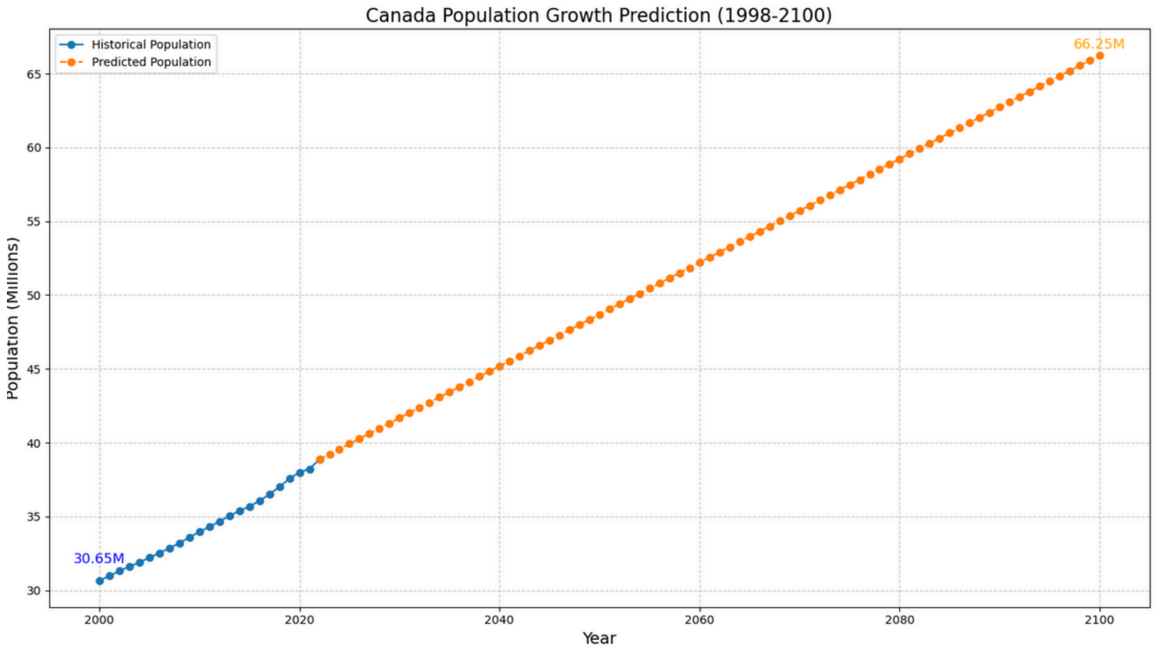
Where n is the number of data points

Where y_i is the actual value of the i-th data point

Where $\hat{y}_i$ is the predicted value of the i-th data point and

$(y_i - \hat{y}_i)^2$ is the squared difference between the actual and predicted values of each point

Figure 12

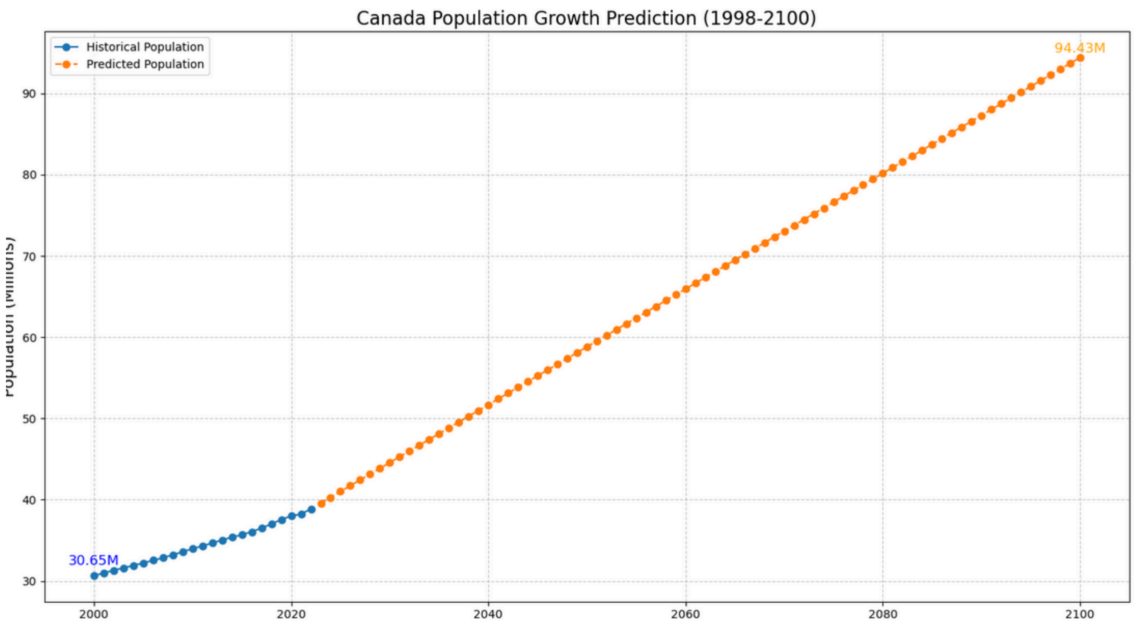


Canada Population Prediction (Random Forests Regression)

Autoregressive Integrated Moving Average

Our previous model, the Random Forests Regression was a linear regression model, which gives us a constant rate of growth; it's fitted to a straight line through our historical data. The Autoregressive Integrated Moving Average (ARIMA) model however, is able to better process trends and seasonality, giving us potentially a more accurate result, shown in Figure 13.

Figure 13



Canada Population Prediction (Autoregressive Integrated Moving Average)

The results of our ARIMA model give us an estimated 2100 population of 94.43M, in comparison to the 66.25M of the Random Forests Regression. The ARIMA equation is somewhat more complex than the MSE used in Random Forests, as seen in (3). Though official predictions for the year 2100 don't exist, the estimate of 94.43M is closer to 100M, which is a goal of some Canadian political organizations, and within reason when extrapolating existing Stats Canada predictions (Blatchford, 2016). Official estimates in a medium high growth scenario put Canada's population at 87.23M in the year 2073 (Statistics Canada, 2024).

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots \theta_q \epsilon_{t-q} + \epsilon_t \tag{3}$$

Where y_t is the target's value at time t

Where c is the constant or intercept (not applicable in our model)

Where ϕ represents the coefficients of autoregressive terms

Where θ represents the coefficients of moving average terms

Where ϵ_t is the residual error at time t

Vector Autoregression

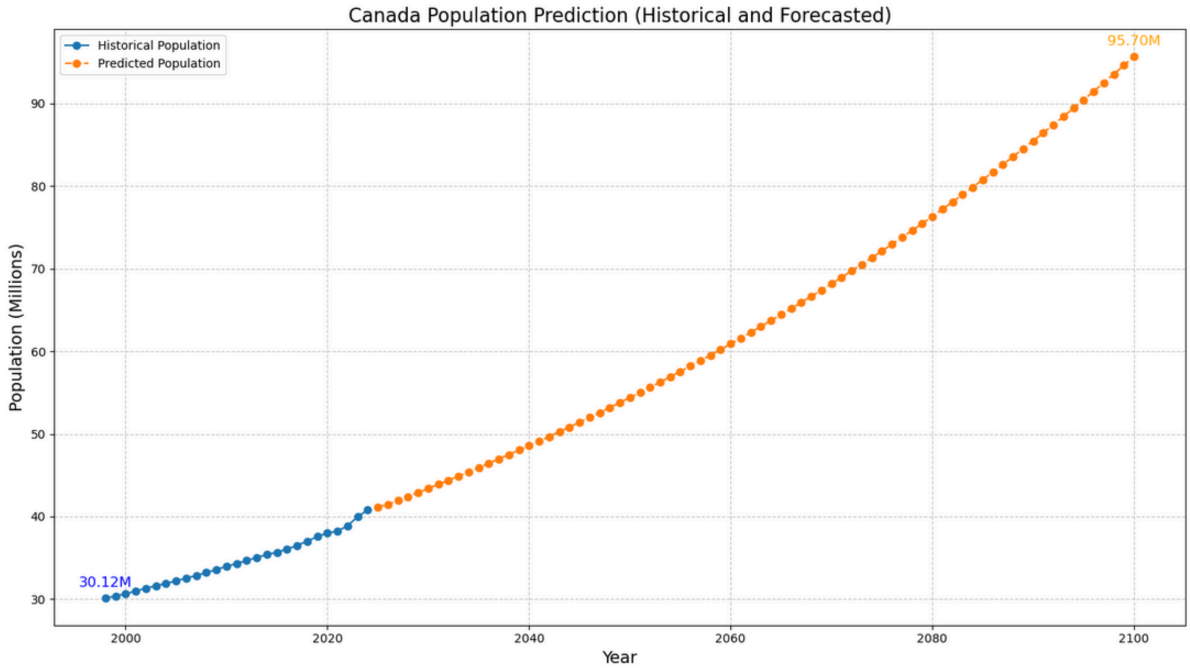
The final model we will look at is the Vector Autoregression (VAR), for which we will use to create an econometric prediction of Canadian population growth using historical GDP growth. VAR models several variables as a linear function, in our case population growth and GDP, as seen in (4). Additionally we also include the mortality and migration data used in previous models. The results of the Vector Autoregression model can be seen in Figure 14

$$\begin{aligned} y_t &= \phi_{11} y_{t-1} + \phi_{12} x_{t-1} + \epsilon_{y,t} \\ x_t &= \phi_{21} y_{t-1} + \phi_{22} x_{t-1} + \epsilon_{x,t} \\ &\dots \end{aligned} \tag{4}$$

Where y_t is population growth at time t

Where x_t is GDP growth at time t

Figure 14



Canada Population Growth (Vector Autoregression)

We see with the results of this new model that we get even closer to the population prediction of 100M in 2100. It's interesting to note that we see a higher population target when we factor in GDP growth, as it has been established that there is a relationship between the two variables (Lianos et al., 2023). When compared to ARIMA the VAR model does a much better job capturing the relationship between the many variables that have been included in this prediction. Though complexity, and therefore performance is much worse (Mahajan et al., 2023).

5. Results

Comparing the results between the three used models we find that ARIMA and VAR were very similar, with a prediction around 95M. The Random Forests model however had a much more conservative prediction of 66.25M. Other studies show that Random Forests can outperform ARIMA in predictive ability, however in our case, Random Forests was the outlier between ARIMA and VAR (Kane et al., 2014). Other studies show that VAR provides very accurate forecasts (Sener et al., 2010). Current Canadian government policy is to trend population towards 100M in 2100, so based on this metric, the recent increase in immigration targets, we can say that for our purposes, VAR and ARIMA provided the more accurate results (Woolf, 2023). Despite not including GDP data with our ARIMA model, and including it with our VAR model they still had very similar results for the 2100 Canadian population prediction. Predictions can be seen in Table 1.

Table 1. Model Population Predictions

Model	Year 2100
Random Forests	62.25M
ARIMA	94.43M
VAR	95.70M

With our results we were also able to further confirm the relationship between several variables in the data exploration section of this paper. With GDP, immigration, and population all being closely related. This is something further proven through outside research (Borjas, 2019). However it has also been shown that this relationship can be broken, or significantly lessened through outside events that are impossible to predict within our models (Kisswani & Khan, 2022). The relationship between immigration and population growth is obvious and static however, where immigration offsets losses from mortality and emigration. We've shown in our research that growth via immigration in Canada is very significant. Something which backs the statistic that growth via immigration accounts for nearly 100% of Canada's current growth due to decreasing fertility rates and increasing mortality with the aging population.

Including GDP in our predictions was of particular interest due to its existence outside the Cohort system. The Cohort system is able to measure and track population growth through simple math of key variables, so the inclusion of GDP in this measure gives an opportunity to predict population growth outside of obvious variables (Kitov, 2006). While the inclusion of GDP didn't significantly alter the results between the ARIMA model and the VAR model a small difference was found. Additionally the inclusion of this additional variable increased the accuracy of our predictions.

6. Conclusion

In conclusion we have produced results from three different models in an attempt to predict Canadian population growth using historical data, such as mortality, fertility, and migration. Our results were bimodal, with Random Forests providing a result of 62.25M, while ARIMA and VAR gave us results around 95M. In the results section we explored why the results produced by ARIMA and VAR can be considered more accurate, given the trends in Canadian policy on population growth (Corcoran, 2016). Despite Random Forests traditionally being a very strong predictor with demographics data (GU et al., 2015).

In the background we discussed the tradition of population predictions, and how there is a long history of this (Vanella et al., 2020). We looked into examples of international predictions, national predictions, and the history of Canadian population predictions. Discussed was also the importance of various factors that influence population change in a country, such as mortality, fertility, and migration. This is commonly expressed with the Cohort-Component method. Our findings backed up our initial discussion of increasing immigration rates and declining fertility that despite birth rates nearly declining below mortality, growth is still very strong, with the Canadian population likely to double by 2100 (Li, 2016).

The source for all of our raw data used to train our models has been Statistics Canada, who provide a wide variety of data on Canadian demographics. It would be possible to expand on our models, and include gender data, regional data, or even occupational data. However to limit the scope and complexity of the project only raw population numbers were considered. It would also be possible to use the same models to predict immigration growth, emigration changes, or changes in fertility. Though as discussed elsewhere in the project, fertility in Canada and most developed nations is trending downward (Sobotka et al., 2011).

In our initial data exploration in this project we were able to gain some key takeaways in showing the relationship between key variables. Such as immigration and population, GDP and population, and GDP and immigration. Having an understanding of these relationships before beginning work on our models allowed for more confident predictions and choice in which variables to choose for our models. Despite the link shown between GDP and population, it was only used with the VAR model due to a wish to not have all models be econometric, despite being a strong predictor (Cappelen et al., 2015). In the future to build on this study more additional variables could be considered in predicting population growth. Such as global conflict which could drive refugees to the country, or the political party in government at a particular time, which may influence immigration targets (Tasker, 2024). Finally one could expand on the study through accounting for historical anomalies in population data, such as the recent spike in immigration post Covid-19.

References

Aitken, R. J. (2022). The changing tide of human fertility. *Human Reproduction*, *37*(4), 629–638. https://doi.org/10.1093/humrep/deac011

Akbari, A. H., & Haider, A. (2017). Impact of immigration on economic growth in Canada and in its smaller provinces. *Journal of International Migration and Integration*, *19*(1), 129–142. https://doi.org/10.1007/s12134-017-0530-4

Aziz, S. (2024, September 26). *Canada's fertility rate has hit a record low. what's behind the drop?* Global News. https://globalnews.ca/news/10773914/canada-fertility-rate-2023/

Beaujot, R., & Rappak, J. P. (1989). The link between immigration and emigration in Canada, 1945-1986. *Canadian Studies in Population*, *16*(2), 201. https://doi.org/10.25336/p61p4q

Blatchford, A. (2016, October 23). *Influential liberal advisers want Canadian population to triple by 2100*. Global News. https://globalnews.ca/news/3020783/influential-liberal-advisers-want-canadian-population-to-triple-by-2100/

Borjas, G. (2019). Immigration and Economic Growth. *Nationa Bureau of Economic Research*. https://doi.org/10.3386/w25836

Brauner-Otto, S. R. (2016). Canadian fertility trends and policies: A story of regional variation. *Low Fertility, Institutions, and Their Policies*, 99–130. https://doi.org/10.1007/978-3-319-32997-0_5

Cappelen, Å., Skjerpen, T., & Tønnessen, M. (2015). Forecasting immigration in official population projections using an econometric model. *International Migration Review*, *49*(4), 945–980. https://doi.org/10.1111/imre.12092

Corcoran, T. (2016, October 20). *Terence Corcoran: 100 million Canadians? we could only hope*. Financial Post. https://financialpost.com/opinion/terence-corcoran-100-million-canadians-we-could-only-hope

George, M. V. (2001). Population forecasting in Canada: Conceptual and methodological developments. *Canadian Studies in Population*, *28*(1), 111. https://doi.org/10.25336/p64s4p

Government of Canada, S. C. (2024a, June 24). *Population Projections for Canada (2023 to 2073), Provinces and Territories (2023 to 2048): Technical Report on Methodology and Assumptions* . Government of Canada, Statistics Canada. https://www150.statcan.gc.ca/n1/pub/91-620-x/91-620-x2024001-eng.htm

Government of Canada, S. C. (2024b, November 26). *About Statcan*. Government of Canada, Statistics Canada. https://www.statcan.gc.ca/en/about/statcan

GU, W., VIEIRA, A. R., HOEKSTRA, R. M., GRIFFIN, P. M., & COLE, D. (2015). Use of random forest to estimate population attributable fractions from a case-control study of *salmonella enterica* serotype enteritidis infections. *Epidemiology and Infection*, *143*(13), 2786–2794. https://doi.org/10.1017/s095026881500014x

Hollmann, F. W. (2022, April 1). *Methodology and assumptions for the population projections of the United States: 1999 to 2100*. Census.gov. https://www.census.gov/library/working-papers/2000/demo/POP-twps0038.html

Hosen, A. (2019). The stability of population and GDP growth: A comparative analysis among different nations in the world. *International Journal of Business and Economics Research*, *8*(4), 180. https://doi.org/10.11648/j.ijber.20190804.13

Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison of Arima and random forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/1471-2105-15-276

Kisswani, K. M., & Khan, S. (2022). Immigration and GDP nexus: Is the association asymmetric? *Economic Change and Restructuring*, *56*(1), 215–236. https://doi.org/10.1007/s10644-022-09420-5

Kitov, I. (2006). GDP growth rate and population. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.886660

Li, N. (2016). Using the probabilistic fertility table to test the statistical significance of fertility trends. *Canadian Studies in Population*, *43*(3–4), 203. https://doi.org/10.25336/p6fp4f

Lianos, T. P., Pseiridis, A., & Tsounis, N. (2023). Declining population and GDP growth. *Humanities and Social Sciences Communications*, *10*(1). https://doi.org/10.1057/s41599-023-02223-7

Mahajan, S., Wagle, S. A., Ranjan, N., & Borde, S. (2023, March 6). *A comparative analysis of Arima and VAR algorithms for performance analysis of high-speed diesel pumps*. International Journal of Intelligent Systems and Applications in Engineering. https://ijisae.org/index.php/IJISAE/article/view/4832

Malthus, T. R. (1826). *An essay on the principle of population*. J. Murray.

McNicoll, G. (1992). The United Nations' long-range population projections. *Population and Development Review*, *18*(2), 333. https://doi.org/10.2307/1973683

Miladinov, G. (2021). The mechanism between mortality, population growth and ageing of the population in the European lower and Upper Middle Income Countries. *PLOS ONE*, *16*(10). https://doi.org/10.1371/journal.pone.0259169

Petty, W., & Morley, H. (1888). *Essays on mankind and political arithmetic*. Cassell.

Previl, S. (2024, October 25). *Immigration has kept Canada's economy afloat. what will coming cuts mean?* Global News. https://globalnews.ca/news/10830645/canada-immigration-cuts-economy-recession-chances/

Sener, E., Baronyan, S., & Mengütürk, L. (2010). Ranking the predictive performances of value-at-risk estimation methods. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1573684

Serebrin, J. (2024, March 27). *Statistics Canada says population growth rate in 2023 was highest since 1957*. CBC News. https://www.cbc.ca/news/politics/population-growth-canada-2023-1.7157233

Smith, S. K., Tayman, J., & Swanson, D. A. (2013). Overview of the cohort-component method. *The Springer Series on Demographic Methods and Population Analysis*, 45–50. https://doi.org/10.1007/978-94-007-7551-0_3

Sobotka, T., Skirbekk, V., & Philipov, D. (2011). Economic recession and fertility in the developed world. *Population and Development Review*, *37*(2), 267–306. https://doi.org/10.1111/j.1728-4457.2011.00411.x

Statistics Canada. (2023). *Table 13-10-0392-01 Deaths and Age-Specific Mortality Rates, by Selected Grouped Causes*. https://doi.org/https://doi.org/10.25318/1310039201-eng

Statistics Canada. (2024). *Table 13-10-0415-01 Live Births, by Month*. https://doi.org/https://doi.org/10.25318/1310041501-eng

Statistics Canada. (2024). *Table 17-10-0009-01 Population Estimates, Quarterly*. https://doi.org/https://doi.org/10.25318/1710000901-eng

Statistics Canada. (2024). *Table 17-10-0014-01 Estimates of the Components of International Migration, by Age and Gender, Annual*. https://doi.org/ https://doi.org/10.25318/1710001401-eng

Statistics Canada. (2024). *Table 17-10-0057-01 Projected Population, by Projection Scenario, Age and Gender, as of July 1 (x 1,000)*. https://doi.org/https://doi.org/10.25318/1710005701-eng

Statistics Canada. (2024). *Table 36-10-0222-01 Gross Domestic Product, Expenditure-Based, Provincial and Territorial, Annual (x 1,000,000)*. https://doi.org/https://doi.org/10.25318/3610022201-eng

Statistics Canada. (2024, March 27). Canada’s population estimates: Strong population growth in 2023. https://www150.statcan.gc.ca/n1/daily-quotidien/240327/dq240327c-eng.htm

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLOS ONE*, *10*(2). https://doi.org/10.1371/journal.pone.0107042

Tasker, J. P. (2024, August 29). *Poilievre says he would cut population growth after liberals signal immigration changes coming*. CBCnews. https://www.cbc.ca/news/politics/poilievre-immigration-cut-population-growth-1.7308184

Tung, S. L. (1984). An econometric analysis of the effects of population change on economic growth: A study of Taiwan. *Applied Economics*, *16*(4), 523–538. https://doi.org/10.1080/00036848400000067

Vanella, P., Deschermeier, P., & Wilke, C. B. (2020). An overview of population projections—methodological concepts, international data availability, and use cases. *Forecasting*, *2*(3), 346–363. https://doi.org/10.3390/forecast2030019

Vollset, S. E., Goren, E., Yuan, C.-W., Cao, J., Smith, A. E., Hsiao, T., Bisignano, C., Azhar, G. S., Castro, E., Chalek, J., Dolgert, A. J., Frank, T., Fukutaki, K., Hay, S. I., Lozano, R., Mokdad, A. H., Nandakumar, V., Pierce, M., Pletcher, M., … Murray, C. J. (2020). Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: A forecasting analysis for the global burden of disease study. *The Lancet*, *396*(10258), 1285–1306. https://doi.org/10.1016/s0140-6736(20)30677-2

Woolf, M. (2023, May 31). *Canada on track for 100 million population but public support can’t be taken for granted: Century initiative CEO*. The Globe and Mail. https://www.theglobeandmail.com/politics/article-canada-immigration-public-opinion/