Ontario Population Modelling

Is constant and unconstrained growth a reasonable assumption for general human population growth models?

Introduction

Population tracking has been ongoing for many centuries, with governments investing significantly into censuses that help to determine the total population of a respective area[8]. Censuses are typically completed annually to determine the qualities of a population in a town, state or country. The results are recorded internationally to monitor population changes due to changing birth, death, and immigration rates. The results are analyzed to project the likely population changes to a over time. Development to account for a growing population needs to be considered decades in advance making this information crucial. This includes but is not limited to ensuring sufficient access to healthcare, schools, transit, and food.

Further, population projection is extremely beneficial to a country because it provides insight into strategies that can help boost economy through new jobs as well as gauge political or social climates in the population. As all governments look to flourish in their economy, there is a strong motivation to continually track and predict their country's population to influence important decisions on different laws such as immigration, housing, farming and trade agreements.

Population is determined by three factors: births, deaths, and net migration. An exponential growth model that only considers these three factors predicts that the population will grow exponentially over time. However, infinite exponential growth is not a reasonable model for population growth tracking because every area is limited by a carrying capacity, which is the number of people that can be supported in a given area. In testing human populations the carrying capacity is particularly complex because it is dependent on societal and natural factors. The political, economic and cultural factors as well as climate, water and agricultural factors all contribute to the carrying capacity. The majority of carrying capacity models assume that a constant carrying capacity exists that does not change over time. For most models this is a sufficient assumption that will be used for the purposes of this project[1]. This project will investigate several carrying capacities and compare the projected population ratios between California and the USA, and extend the subsequent models to an Ontario and Canada context.

Model

Base Model

In order to compare these populations we constructed a base model described below:

- The base model(s) is discrete, linear, first-order, autonomous, and is represented as systems of linear equations.
- The discrete time unit t is measured for every five years.
- There are two state variables in the model(s): population of California (C_t), and population of rest of the US (RU_t). Both of these state variables are units of people (in thousands).
- There are three parameters in this model; (1) birth rates (b_i) (i.e California birth rates are b_c and rest of the US will be denoted as b_u), (2) death rates d_i, and (3) rate of net migration rn_i. These parameters are per capita rates and will remain constant in the model. The units of the parameters are measured as people per people per five years.
- The model(s) will be assumed to be 'closed.' That is, the model(s) only takes into account the state variables within the given country, and there is no state variable representing outside immigration. It will then be assumed that immigration to country/state/province is captured in the parameter for birth rates, while emigration is captured in the parameter for death rates.
- Homogeneity in the population (everyone in a given region is identical) is assumed in the model(s).
- The model(s) are deterministic and assumes no 'shocks' to the population. That is, we are excluding the affect of major societal, political, economic, and environmental factors that could impact population growth.

The parameters for the base model(s) are calculated as the following:

	b_i	d_i	rn_i	Growth Rate[1]
California	0.1315060	0.0472744	0.0865414	1.170773
Rest of US	0.1282137	0.0487697	-0.0073907	1.072053

The base model is constructed as the following system of linear equations (given the information above)[2]:

$$P_t = (I + B - D + M)P_{t-1} = GP_{t-1} \qquad \text{(General Equation)} \qquad (1)$$

[1] Growth rate for a given region is calculated as $1 + rn_i + b_i - d_i$

[2] Note that the identity matrix I must be introduced as it adds the population at P_{t-1} + (inflow of new population)P_{t-1}. Otherwise, from $t-1$ to t, the population at t would just be a fraction of the old population rather than the growth of the population.

The base model will be further extended towards a Ontario vs. the rest of Canada context for population projections past 2020. However, the base model for California must be adjusted as there is an absence of data recording internal migration until 1972 [9]. Similarly, Stats Canada earliest public figures for births and deaths only begin in 1991 [8]. To accommodate for the issue, our project will simplify the growth rate calculation as $1 + (P_{1960} - P_{1955})/P_{1955}$[3][4]:

Region	Pop at 1955	Pop at 1960	Growth Rate
California	12988	15206	1.170773
Rest of US	152082	163040	1.072053
Ontario	5208	6054	1.162442
Rest of Canada	10327	11656	1.128692

Table 1: All population figures are in thousands

We then have the following systems of equations[5]:

$$\begin{bmatrix} C_t \\ RU_t \end{bmatrix} = \begin{bmatrix} 1.170773 & 0 \\ 0 & 1.072053 \end{bmatrix} \begin{bmatrix} C_{t-1} \\ RU_{t-1} \end{bmatrix} \qquad \text{(Cal Model)} \tag{2}$$

$$\begin{bmatrix} O_t \\ RC_t \end{bmatrix} = \begin{bmatrix} 1.162442 & 0 \\ 0 & 1.128692 \end{bmatrix} \begin{bmatrix} O_{t-1} \\ RC_{t-1} \end{bmatrix} \qquad \text{(Ont Model)} \tag{3}$$

Lastly, a secondary growth model which examines changes in inter-state/provincial migration (denoted as T for migration transfers). The following model will be of the form:

$$P_t = TP_{t-1} \qquad \text{(General Equation)} \tag{4}$$

As mentioned earlier, since there was no recorded accounts on detailed inter-provincial migration from Stats Canada until 1972, we will compensate by assuming that the rate for 1955-1960 migration was similar to the one observed between 1972 and 1977:

Region	To Region (1)	To Region (2)	Total Pop at t_0
California (1)	12174	814	12988
Rest of US (2)	1938	150144	152082
Ontario (1)	7404	560	7964
Rest of Canada (2)	604	13651	14255

Table 2: All population figures are in thousands

[3]Note that since the result will also return the same Growth rate for the California model.

[4]Population for Canada is taken from quarterly estimates published by Stats Canada [8]

[5]Where C is California population, RU is rest of US population, O is Ontario population, and RC is rest of Canada population.

Where we then have the following systems of equations:

$$\begin{bmatrix} C_t \\ RU_t \end{bmatrix} = \begin{bmatrix} 0.93732676 & 0.01274313 \\ 0.06267324 & 0.98725687 \end{bmatrix} \begin{bmatrix} C_{t-1} \\ RU_{t-1} \end{bmatrix} \quad \text{(Cal Model)} \quad (5)$$

$$\begin{bmatrix} O_t \\ RC_t \end{bmatrix} = \begin{bmatrix} 0.92968358 & 0.0423711 \\ 0.07031642 & 0.9576289 \end{bmatrix} \begin{bmatrix} O_{t-1} \\ RC_{t-1} \end{bmatrix} \quad \text{(Ont Model)} \quad (6)$$

Graphical Results

After running the appropriate R code, we generate the following plots:

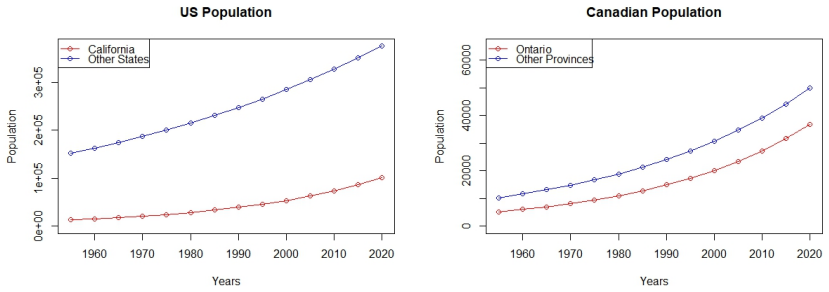

Figure 1: Modelling results are based off of Matrix (2) and (3) respectively

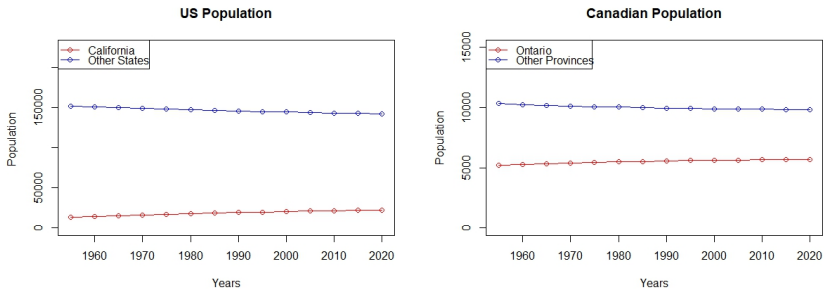

Figure 2: Modelling results are based off of Matrix (5) and (6) respectively

However, a key issue with the predicted models is that they do not accurately predict population growths. Below is the actual data on US and Canadian population growth models (compare these models with figures 1 and 2):

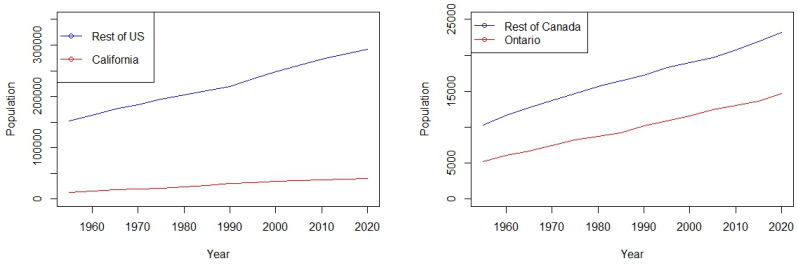

Figure 3: Actual population growth for US and Canada

In general, unlike the simulated models with base models which grow exponentially, actual trend appears more linear for both countries (and also California and Ontario). However, when considering the models as proportional growth models, the results for Ontario vs. Rest of Canada is similar in the simulated model compared to the actual (the same can not be said for the California vs. Rest of US case)[6]:

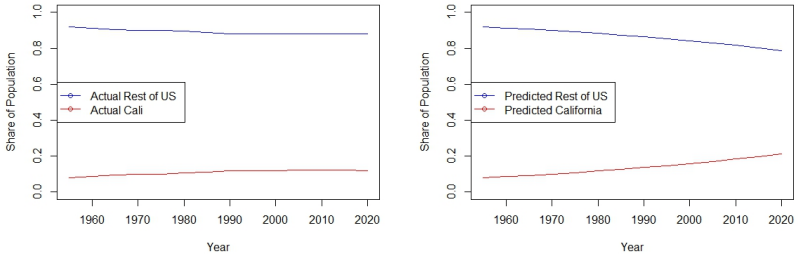

Figure 4: Shares for Actual vs Predicted model for US

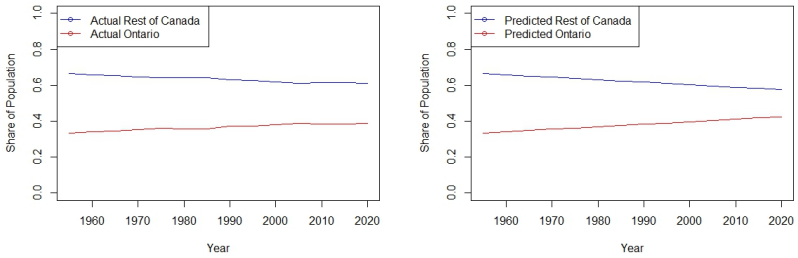

Figure 5: Shares for Actual vs Predicted model for Canada

[6]California and US population is taken from the St. Louis Federal Reserve[5].

Analysis of Base Model

With the given information, we set up the model as:

$$P_{t+1} = (I + B - D + M)P_t$$

Where B represents the birth rate, D represents the death rate, and M represents the net migration rate. Now let $G = (B + D + M)$, our model can be interpreted instead as:

$$P_{t+1} = GP_t$$

Thus, for there to exist a stable equilibrium, we would require for the condition:

$$P = GP \rightarrow G = 1$$

However, for both the base model of California and the applied base model to Ontario, we find that both models lack stable equilibrium points. That is, for the two cases:

$$\begin{bmatrix} 1.170773 - \lambda & 0 \\ 0 & 1.072053 - \lambda \end{bmatrix} \quad \text{(Cali Model)}$$

$$\begin{bmatrix} 1.162442 - \lambda & 0 \\ 0 & 1.128692 - \lambda \end{bmatrix} \quad \text{(Ont Model)}$$

The dominant eigenvalues for California and Ontario models are found to be $\lambda = 1.170773$ and $\lambda = 1.162442$ (respectively). Since for both $\lambda > 1$, we then conclude that the base models of (2) and (3) do not have stable (non zero) equilibriums, and thus the models will grow without bounds as $t \rightarrow \infty$. In fact, the non-dominant eigenvalues are also $\lambda > 1$, and thus populations (including rest of US and rest of Canada) for both models will grow without bounds.

Now working with the migration models of (5) and (6), suppose that an equilibrium exists, that is that both populations approach some steady state. For the California model (5), suppose that both equations are at equilibrium, that is to say where:

$$\begin{bmatrix} C \\ RU \end{bmatrix} \begin{bmatrix} 0.93732676 & 0.01274313 \\ 0.06267324 & 0.98725687 \end{bmatrix} \begin{bmatrix} C \\ RU \end{bmatrix} = \begin{bmatrix} a & b \\ A & B \end{bmatrix} \begin{bmatrix} C \\ RU \end{bmatrix}$$

where $a + A = b + B = 1$. From here, note that:

$$C = aC + bRU \rightarrow C = \frac{bRU}{1 - a}$$

$$RU = AC + BRU \rightarrow C = \frac{(1 - B)RU}{A}$$

Solving the above equations, we have that:

$$\frac{bRU}{1 - a} = \frac{(1 - B)RU}{A}$$

Because the total population does not grow in this model (this model is build from only net migration transfers) and is constant at the level of population at 1955, then we have that:

$$P_{1955} = C + RU \rightarrow 165070 = \frac{bRU}{1-a} + RU \rightarrow 165070 = RU[1 + \frac{b}{1-a}]$$
$$RU = \frac{165070}{1 + \frac{b}{A}} \approx 137178.07$$

Then for California, we find the equilibrium to be:

$$C = P_{1955} - RU = 165070 - 137178.07 = 27891.94$$

Checking for stability, we can see that from the Jacobian, that all eigenvalues are strictly less than 1:

$$0 = det(J - \lambda) = det[\begin{bmatrix} 0.93732676 - \lambda & 0 \\ 0 & 0.98725687 - \lambda \end{bmatrix}$$

Meaning, that $(C, RU) = (27891.94, 137178.07)$ is a stable point, and that in the long-run, we expect California to account for approximately 17% of the US population, while all other states are roughly 83% of the US population.
Repeating the exact same steps for the Ontario model (6), where:

$$\begin{bmatrix} O \\ RC \end{bmatrix} \begin{bmatrix} 0.92968358 & 0.0423711 \\ 0.07031642 & 0.9576289 \end{bmatrix} \begin{bmatrix} O \\ RC \end{bmatrix} = \begin{bmatrix} a & b \\ A & B \end{bmatrix} \begin{bmatrix} O \\ RC \end{bmatrix}$$

From the condition:

$$P_{1955} = O + RC \rightarrow 15535 = O + RC$$

Once again, repeating the same procedure as in California case, we find that $(O.RC) = (5841.24, 9693.76)$. And once again, checking for stability of equilibrium, we can see that from the Jacobian, all eigenvalues are strictly less than one:

$$0 = det(J - \lambda) = det[\begin{bmatrix} 0.92968358 - \lambda & 0 \\ 0 & 0.9576289 - \lambda \end{bmatrix}$$

Which once again implies that the result of $(O.RC) = (5841.24, 9693.76)$ is a stable solution.

Model Extension

Based on the results of the proposed analyses above and comparison among simulated results vs actual data, we can see significant discrepancies between our current model and the real-life data. Our model is predicting that the population would grow to infinity, but we know that this is unrealistic in real world circumstances.

We see that the base model does not take into account that the birth rate is used as a constant variable derived from 1960. We can see that this doesn't properly reflect the trend of birth rates as time progressed in society. As the 1960s were a period during the baby boom, the birth rate of that time would not be comparable to the birth rates that are experienced today. We've seen with the development of technology through birth control and socio-economic factors like the lack of affordable housing, and the idea of independence where having children that could drain emotional and financial resources creates a society that finds having children undesirable [7]. Similar comments could be made for both the death rate and net migration. This creates a strong motivation to create a carrying capacity variable that will maintain the unconstrained growth/decline of our variables to a reasonable standard that follows that of the real world.

There are many factors that influence a population, prominently natural, economic, political, and social elements. Specifically, the amount of land, living resources, and food make it impossible for a population to grow infinitely. However, creating variables and creating an equation with these elements would cause the model to become too complicated, therefore, we propose to extend the base population models to include a theoretical value for carrying capacity K. Carrying capacity provides a boundary to mathematical models that results in more accurate real-world data. Proposing a carrying capacity. The provided sketch is a theoretical graph of what our model may be:

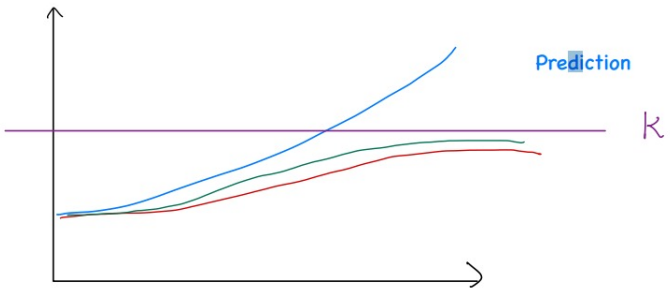

Figure 6: Where the blue line is base model predictions, the green line is actual observations, and the red line is one possible theoretical modelling result.

Using a simple logistic growth model of the form:

$$P_t = [\frac{(1-G)P_{t-1}}{K} + G]P_{t-1}$$

We can then introduce carrying capacity (K) into the population growth model, assuming that the population growth rate (G) is constant as the population approaches the carrying capacity K. This reflects the increasing scarcity of resources as the population grows, leading to increased competition and social costs associated with lower growth rates for developed countries in the 21^{st} century. When the population approaches carrying capacity, the per capita growth rate will near zero. By introducing carrying capacity into the model, we can better model the dynamics of population growth and the interaction between population and the environment. This allows for more accurate projections of future population size and a better understanding of the factors affecting population growth and stability. In this section, we will now introduce three extensions, all of which are improvements on subsequent extensions.

Extension 1: Proportion of K

In this extension we will explore a possible extension of the California model to Ontario. The motivation behind this extension is that Ontario and California are in similar positions relative to the rest of the country, they are among the top most populous and geographically largest provinces and states. Furthermore, both areas contain large cities such as Toronto and Los Angeles which stand as the most populous cities in Canada and second most populous city in the USA. Additionally, Canada and the USA are relatively similar countries with comparable, cultures, climates and geographic sizes. Both countries are composed of large immigrant populations from all over the world. Although the population of the USA is ten times that of Canada, the similarities in culture and economic goals of each nation indicate that these are comparable populations to analyze[2, 3].

The first model extension is done through guessing some general K values for California and the rest of US for 'best fit' [7] with the actual data. From or own analysis, we guessed K values to be 52000 for California and 760000 for the rest of the US. Figure 7 (in Results section) showcases the simulated results K.

From here, we investigated the relationship between our guessed capacity $\hat{K}$ for each given population in 1955 (P_0) for the following proportion:

$$\alpha = \frac{P_0}{\hat{K}}$$

In our investigation of a relationship between the population number and K for the year 1955, we find that the ratio of the population number to K was approximately 0.25 for California and approximately 0.20 for the rest of the US for the year 1955:

$$\alpha = \frac{12988}{52000} = 0.24976923 \qquad \text{(California)}$$
$$\alpha = \frac{152082}{760000} = 0.20010789 \qquad \text{(Rest of US)}$$

[7] This measure was fairly crude as it was purely done through the eye test.

From above results, we supposed that proportion of population to carrying capacity may be a helpful predictor of population growth dynamics across different regions and time periods. In particular, for major economic areas, this ratio may be around 0.25 at around 1955, and lower for all other areas. We, therefore, venture to speculate that the concept of carrying capacity is relevant to understanding population growth in cities in developed countries, where resource constraints and other environmental factors may play an essential role in determining population dynamics.That is to say, for our own application to the Ontario model, using data corresponding to the same period of 1955, we estimate the K values to be:

$$K = \frac{P_0}{\alpha} = \frac{5266}{0.25} = 21064 \qquad \text{(Ontario)}$$
$$K = \frac{P_0}{\alpha} = \frac{10327}{0.2} = 51635 \qquad \text{(Rest of Canada)}$$

However, to apply the proportions method, we first need to make a series of assumptions that allow us to extend this method to Ontario. We then investigate a series of similarities for population growth patterns, and find the following:

- 1: Both Ontario and California experienced significant population growth in the mid-20^{th} and early 21^{st} century.
- 2: Both Ontario and California are highly developed regions, and they both are the most populated province/state of their countries, and share in-common relatively stable political and economic systems.
- 3: Both regions have been their respective countries' largest economic areas.
- 4: Both regions face environmental sustainability and resource management challenges that affect population growth and development, which will directly result in affecting the growth rate. for example, both Ontario and California face issues related to air pollution and water scarcity.

Although we do need to apply similar assumptions to the rest of US and Canada, instead of doing so, lets first simply model the derived K values. As a result we find that the simulation is a poor fit when compared to the actual data (see Figure 8 under Results¿Extension Results).

For the rest of Canada, this ratio should be 0.2, however, when we try to apply this ratio on the rest of Canada, the predicted growth growth path is not consistent with real world data. This suggest that the assumptions of the US model are not suitable enough to be applied to the rest of Canada. Possible reasons to the limitations of Extension 1 can be attributed to the following:

- Geography: The rest of Canada and the rest of the US have different geographical locations, for example, the rest of Canada includes many remote and sparsely populated areas, while the rest of the US includes

many densely populated urban areas. The United States also has a large number of plains, which are ideal for human habitation and growing food. In contrast, most of Canada's landscape is frozen tundra.

- Climate: The northern part of Canada is a glacial region close to the Arctic, which is very cold and unsuitable for human habitation. The United States, on the other hand, has a suitable subtropical monsoon climate, etc. Both climate and geography interact with each other.

- Difference in Growth: In Ontario, the Greater Toronto Area (GTA) has been a major driver of population growth, while in the US, in addition to California, cities such as Los Angeles, San Francisco and San Diego have seen significant population growth. As we can see from Fig. 4, 5, Ontario's share of the population is higher than 50 % so Ontario's population growth is the main population growth for the country as a whole. Other cities such as Los Angeles, San Francisco, New York, San Diego, and Washington, D.C., have experienced significant population growth. Therefore, it can be inferred that the rest of Canada is less competitive than the rest of the United States, as evidenced by the different rates of population growth. The reasons for this are varied and can be influenced by policies, institutions, economic growth and development.

Therefore, due to limitations and the over reliance of assumptions, we conclude that the proportions method described in Extension 1 is likely not a solid method for calculating a K value.

Extension 2: Optimization of K

When constructing a model, a key component of a good model is for it to be accurately constructed out of all available information, while also being consistent in its assumptions. A major issue of Extension 1 was that we were required to carry over a number of assumptions from California and the US to Ontario and Canada in order for proportions to remain true. Another issue was that simply 'guessing' a trend is not reliable for future forecasting. Instead, suppose that we can develop some method to estimate a K value in the California case, and then test the extension to Ontario's case. Extension 2 will instead focus on developing such a method, however, a few brief (yet key) assumptions:

- 1. The logistic growth model described earlier is true. That is, for some time t, future populations is calculated as $P_t = [\frac{(1-G)P_{t-1}}{K} + G]P_{t-1}$

- 2. However, carrying capacity K, which is loosely defined as a variables for limitations on population growth, and is comprised of countless factors, is itself NOT constant. That is to say, their are fluctuations on potential growth.

- 3. All other assumptions from equation (1) are carried over.

Now, recalling the growth model again, suppose we define it as a function of P_{t-1} and K, that is:

$$P_t = f(P_{t-1}, K) = (\frac{(1-G)}{K} P_{t-1} + G) P_{t-1}$$

However, since P_{t-1} is itself a functions of P_{t-2} and so on until we have some $P_1 = f(P_0, K)$, we then find that all calculations of P_t are based on some constant initial measure of P_0. Now, from here, we have the general form of our growth function to be:

$$P_t = f(K) = (\frac{(1-G)}{K} P_0 + G)^t P_0$$

Thus our only concern now is finding some constant for $\hat{K}$ which best fits with all observed data. Now suppose that there exists some arbitrary function $V(\hat{K})$, which solves the following problem[8]:

$$V(\hat{K}) = min \sum_{t=1}^{T} \left| P_t - f(\hat{K}) \right| \quad \text{for } t = 1, 2, .., T$$

Solving this problem manually is tedious, so we write the following code (see Extension 2 Code in Appendix: R Code). Running the code, we find the optimal $\hat{K}$ for California to be $\hat{K} = 46803$ and $\hat{K} = 890053$. Graphically, this result is seen in Figure 9 (see Results > Extension Results). Focusing just on the results for California, we find that this new K to better fit with the data than our original guess of 5200.

Repeating the code once again but for Ontario and the rest of Canada, we find the optimal $\hat{K} = 18804$ for Ontario and $\hat{K} = 31488.000$ for the rest of Canada. The modelling results do appear to better fit the data (see Figure 10). And likewise, it appears that the data is a better fit for just the Ontario case as well (see Figure 11).

Extension 3: Optimizing K with Time Value

One core issue with Extension 2 is in forecasting future results. The arbitrary function $V(\hat{K})$ that we described is minimizing the total distance between actual observations with modelled observations, and it is assuming that the error in say, 1960 is as equally important to us as the error in 2020. However, if we were to model the world of 2023, wouldn't the information of 2020 be of greater importance to us than the information from 1960? Intuitively, it makes sense since the more recent the information, the more accurate it may be in describing the world of tomorrow since the parameters of 10 years ago are more like our current parameters than from 60 years ago. Extension 3 incorporates this assumption by building on the results of Extension 2.

[8]The inspiration of this method came from minimization of sum of squares for regression coefficients, as well as the "value function" method used in economic optimization problems.

Firstly, model assumptions as described in Extension 2 are carried over to this extension. Then, suppose we define y_t as the following:

$$y_t = \frac{|P_t - f(\hat{K})|}{P_t}$$

Now, suppose that we introduce some constant $\beta \geq 1$. From here, suppose again that we have the arbitrary value function $V(\hat{K})$ which solves the following problem:

$$V(\hat{K}) = min \sum_{t=1}^{T} \beta^{t-1} y_t \quad \text{for } t = 1, 2, .., T$$

The purpose of writing y_t is to 'standardize' the distance/errors relative to the accuracy of the results. That is, the error is bounded at $0 \leq y_t \leq 1$, and the more accurate the results, y_t will approach 0. While β is some 'time-value' of information that we choose.

From our experience, choosing $\beta = 1$ will result in a $\hat{K}$ that is the same as in Extension 2. As β increases, results of $\hat{K}$ will become increasingly skewed in preference towards the last entry in the data, such that the solution for $\hat{K}$ will be closer to solving the growth rate of the most recent results. However, there also exists some β values (from our experience usually some value in the range of 1.4 to 1.7) that provides a capacity in-between the two estimates. Running this code for California and rest of the US, we generate capacities of 45662 ($\beta = 2$), 46785 ($\beta = 1.5$), and 46803 ($\beta = 1$) for California, and 794039 ($\beta = 2$), 832430 ($\beta = 1.5$) 890053 for the rest of the US. For modeling results, see Figures 12 and 13 (showed separately for better detail).

Repeating the same procedure for Ontario and the rest of Canada, we find capacities of 19585 ($\beta = 2$), 19036 ($\beta = 1.5$) and 18804 ($\beta = 1$) for Ontario, and 32598 ($\beta = 2$), 32202 ($\beta = 1.8$) and 31488 ($\beta = 1$) (see Figures 14 and 15)

Results

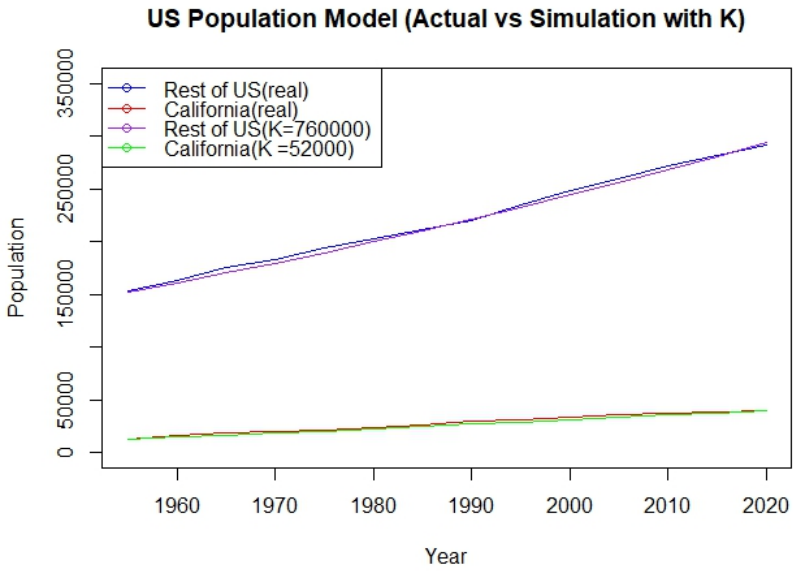

Figure 7: Population is measured in 1000s

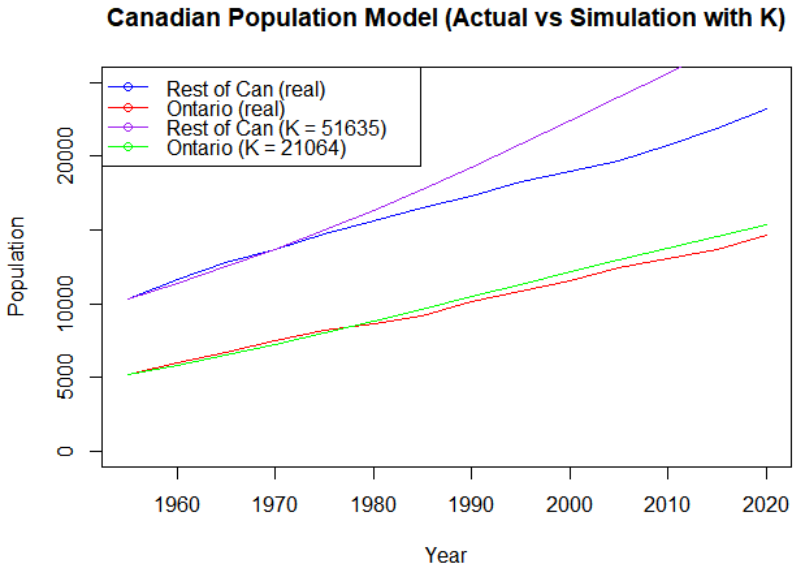

Figure 8: Population is measured in 1000s

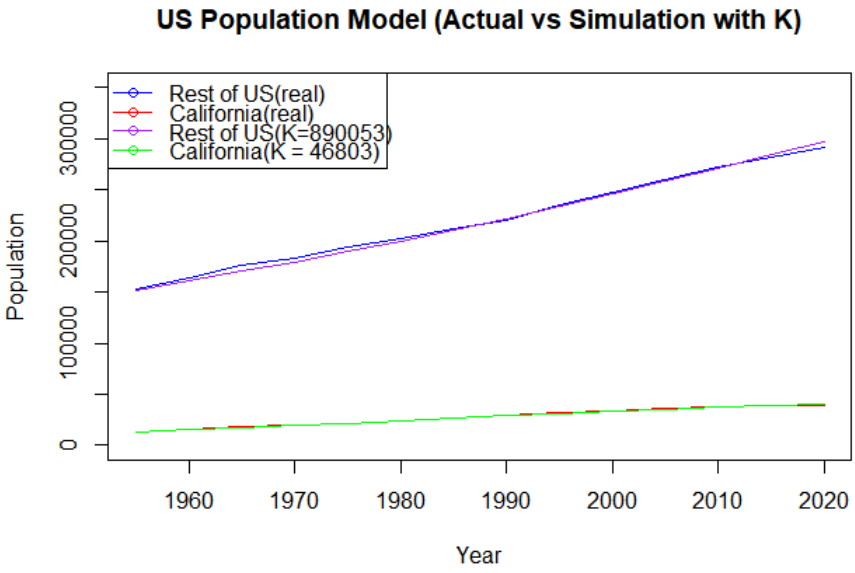

Figure 9: Population is measured in 1000s

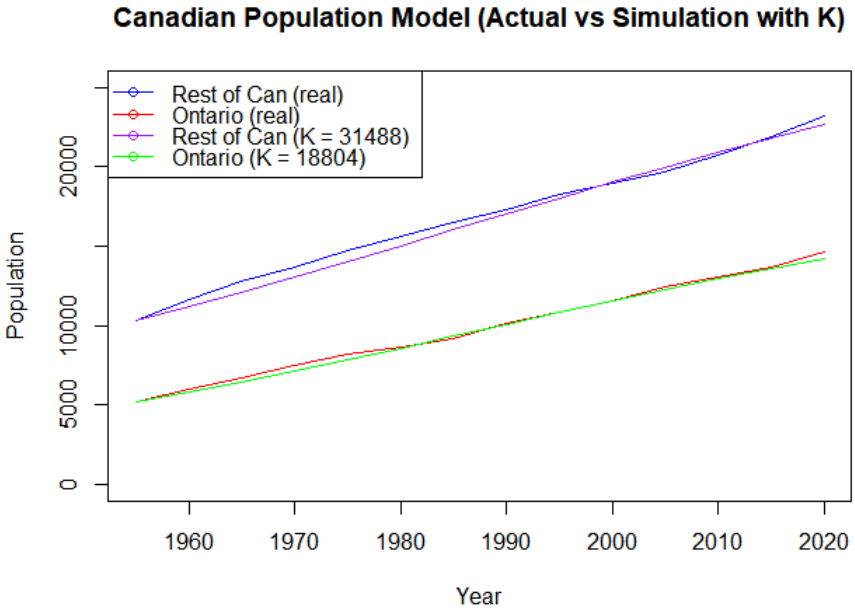

Figure 10: Population is measured in 1000s

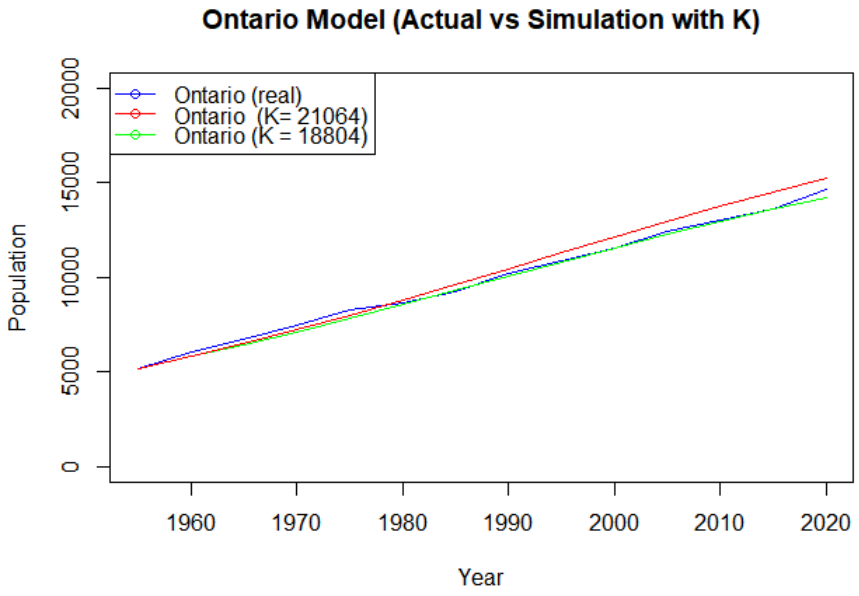

Figure 11: Population is measured in 1000s

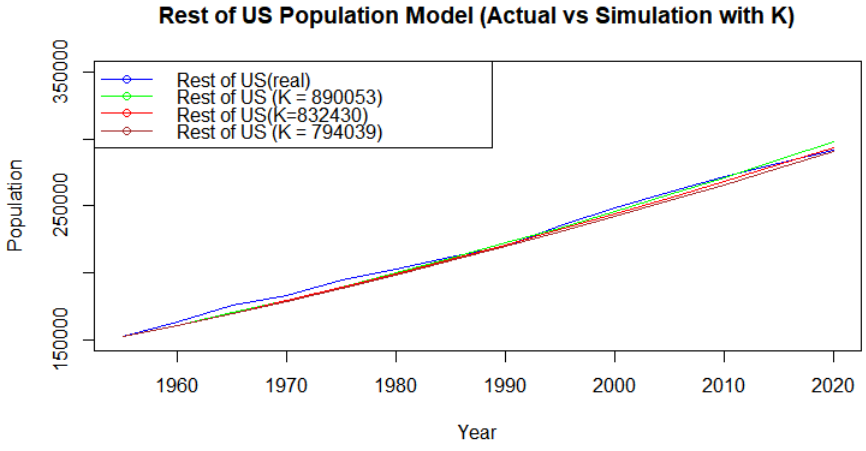

Figure 12: Population is measured in 1000s

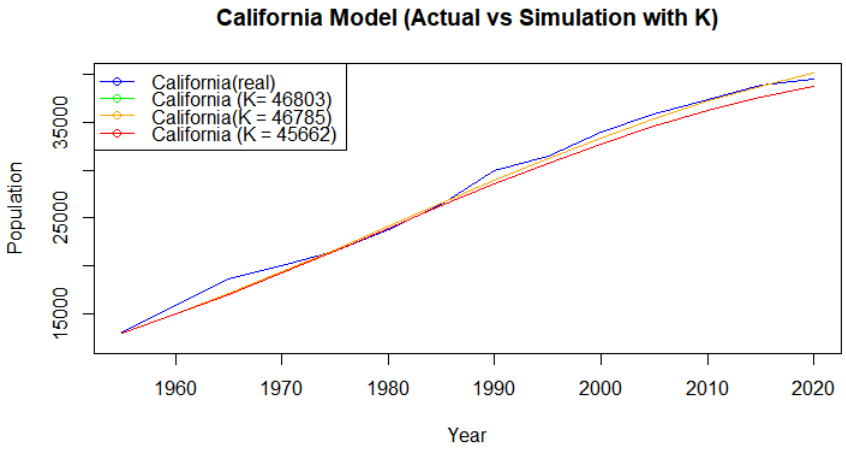

Figure 13: Population is measured in 1000s

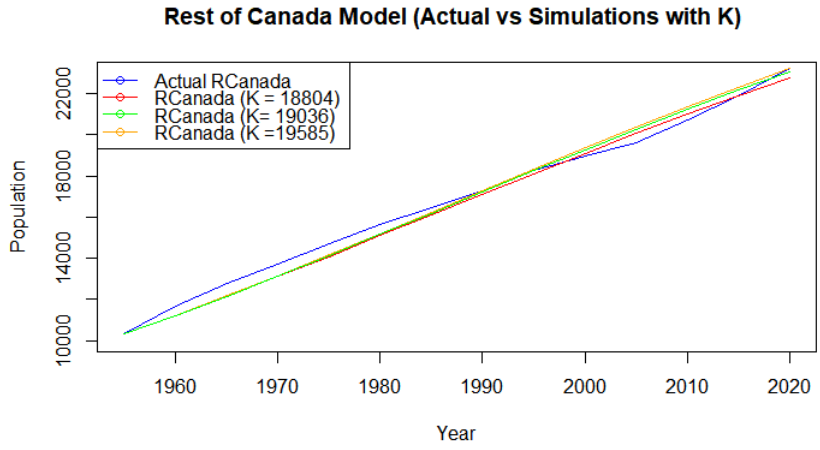

Figure 14: Population is measured in 1000s

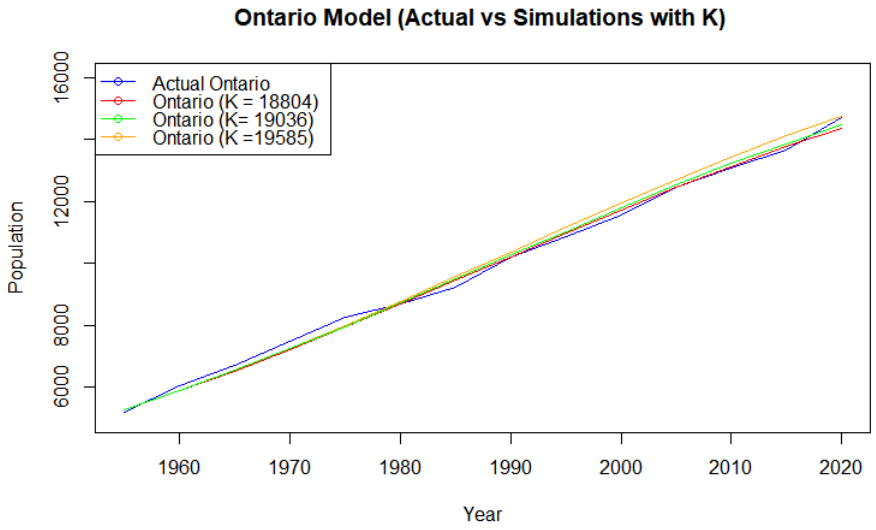

Figure 15: Population is measured in 1000s

Projection Comparison

For "Projection Comparison", we will compare our derived K values with Stats Canada's projections for Ontario's population until the year 2040 (as the projections are limited until the year 2043 for Ontario)[10, 11]. The results will focus with Stat Canada's 'low-growth' and "medium-growth' scenarios, which will be compared with K levels of 21064 (Extension 1), 18804 (Extension 2), 19036 (Extension 3) and 19585 (Extension 3). We thus have the following results:

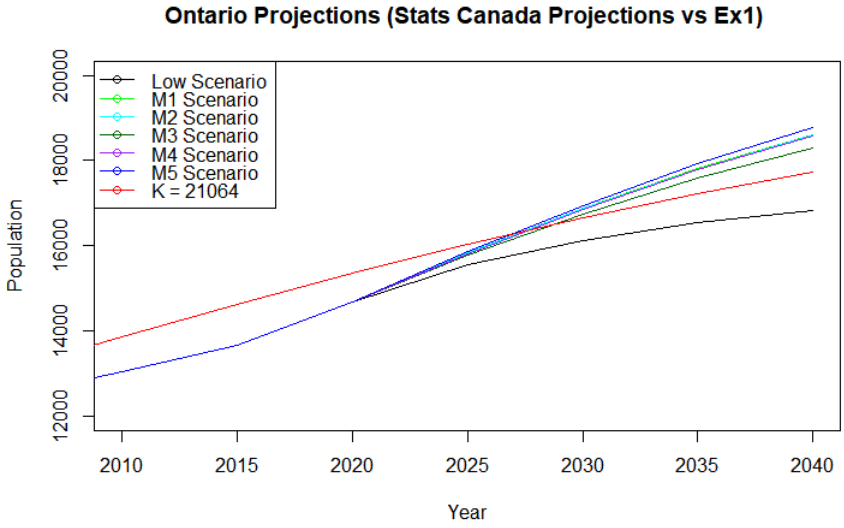

Figure 16: Population is measured in 1000s

Extension 1 results appear to be in-between both medium and low growth scenarios for Ontario projections. However, since there is a lot of room for potential error in derivation of K, the validity of this result may be more the result of pure chance.

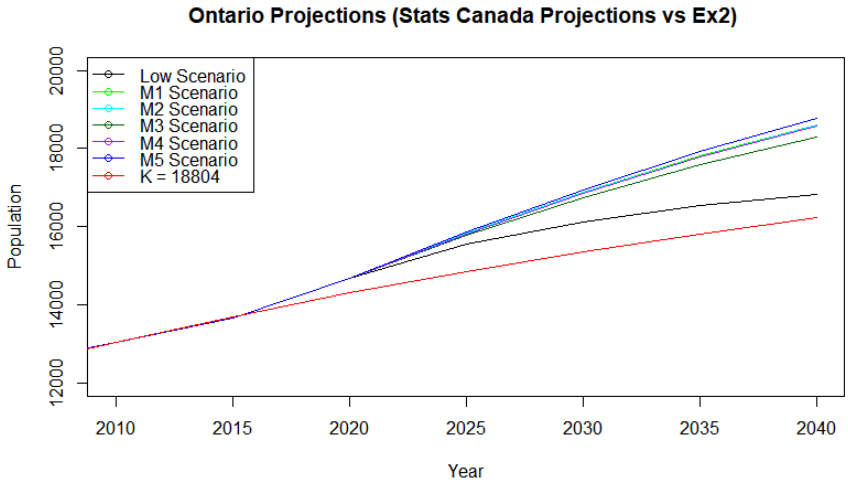

Figure 17: Population is measured in 1000s

Extension 2 results, while although inline with the data until 2020, appears to underestimate the growth for Ontario even more so than Stats Canada's lowest

scenario.

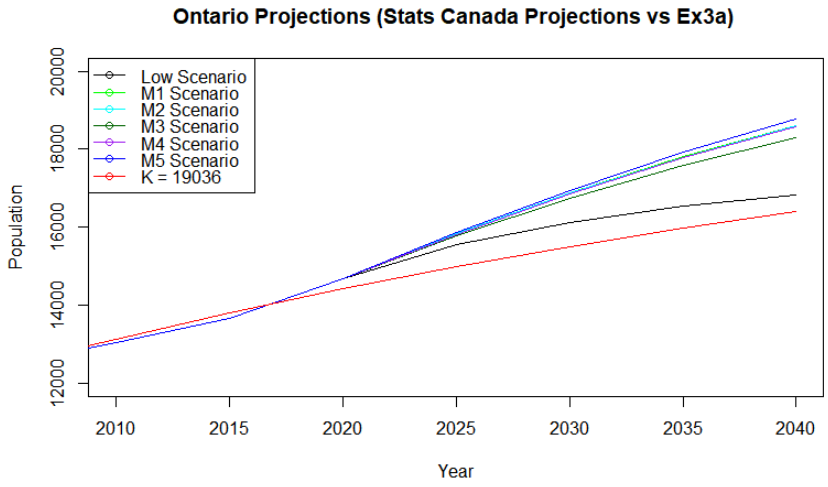

Figure 18: Population is measured in 1000s

Extension 3a, like extension 2, also appears to underestimate growth, however, the results to appear to converge towards Stats Canada's lowest scenario estimate.

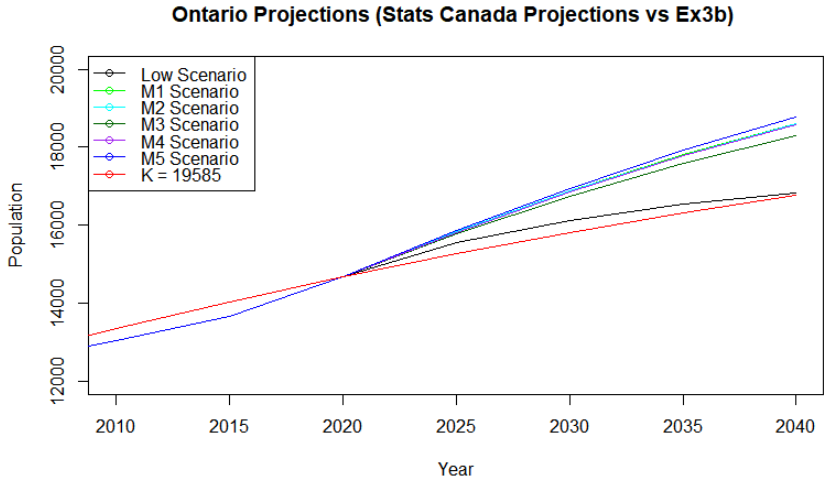

Figure 19: Population is measured in 1000s

Extension 3b appears to be the most 'accurate' estimation as it close to being inline with Stats Canada's estimate for the lowest growth scenario.

Overall, our models appear to underestimate Ontario population growth when compared to Stats Canada's own projections. This could be attributed to the following reasons:

- 1. Unlike our models which, Stats Canada builds their models with the most up to date variables contributing to growth rates (including inter-provincial migration, deaths, birth. and migration from outside of Canada).
- 2. Stats Canada models incorporate immigration from outside of Canada as the main source of population growth in their models, while our models are built on "closed" assumptions (growth through only births and inter-state/provincial migration).
- 3. Stats Canada models also incorporate difference amongst age groups and ratio of male/female population in their growth models
- 4. From the graph we can see that the curve of population growth changes periodically. Even when actual data on current population size and demographic trends are available, unexpected events such as disease outbreaks (covid-19), political unrest (war) or sudden changes in environmental conditions (climate change, earthquakes, etc.) can lead to rapid and unpredictable changes in population dynamics.
- 5. Policy changes and government goals strongly affect immigration rates, which led to a higher growth rate in 2015. Our model is based on actual data from 1955 to 2020, due to policy changes and government goals, population's growth rate in the last five years has been significantly higher than in the 20th century

Discussion

In our findings, we conclude that unconstrained growth to not be a reasonable assumption in modeling our data. However, before arriving to this conclusion, we had first constructed our base models (2) and (3) as systems of linear equations. These linear system of equations measured in discrete time intervals were shown to model exponential growth. When modelled against real-world results, we found these models to not sufficiently capture real trend as demonstrated in Figure 1 vs Figure 3. A further mathematical analysis was performed on these models and formally showed that these models exhibit unconstrained growth. In addition, we also constructed two models (5) and (6) which only account for inter-provincial/ state migration and performed a similar analysis. However, the focus for this project was growth, and thus we were focused on improving the base models of (2) and (3). However, models (4) and (5), based off of analysis, had demonstrated a slow convergence of the populations while actual data finds the proportions to not change significantly over time. This is likely due to factors that are difficult to measure, as governments tend to incentivize immigration to ensure a more even population growth across the country so legislation may be playing a role in the variations seen in the figure 4 and 5 models.

Since it is well established that a population cannot grow without bounds forever, nor that our data exhibited a traditional exponential growth, this necessitated the development and introduction of a carrying capacity (K) to our models. In our extensions, we illustrated three approaches for K value determination. In our first approach, we used an estimated K value. In our second approach, we optimized a K value from a logistic growth equation. Then in our final approach, we built on the method of our second approach but introduced a method for. Our first extension estimated K as 52 and 760 million for California and the rest of the US, respectively. This model with a the estimated K values had appeared to fit well with actual data and more appropriately than the exponential growth approach (Figure 7). In predicting the Ontario population's carrying capacity, a ratio calculated from the population in 1955 to the estimated K value used for California was found and then used to calculate a potential K value for Ontario (the same was done for rest of US to rest of Canada). This model deviated significantly from the actual data as seen in figure 8. This deviation occurs around 1970 marking the end of the baby boom that lasted from 1941 to 1970. The birth rates in this model are from the baby boom era (1941-1970) and are not similar to the birth rates as seen today. Meanwhile, the trajectory for the US was sustained since around 1975 the US government enacted legislation to raise immigration from Asia, Central America, and Mexico. In contrast, Canada did not implement similar liberal immigration policies as the US during this period, which is likely a crucial reason for our modeling discrepancies in figure 8 [6], [12].

Extension 2 used a mathematical optimization approach to calculating a K value for the logistic growth model. As figures 9 through 11 demonstrate, this extension was a much better fit than the method suggested in Extension 1 when compared with real world data. However, when considering the model alongside projected Ontario models, our own model proved to underestimate growth. Since it appears that in the last decade population growth has been increasing, our own model which simply fitted data based on the general trend since 1955, had underestimated projected growth. To compensate for recent trend, we suggested a third and final extension that build off of Extension 2. Extension 3 introduced the concept of 'time-values' that, essentially, modifies the optimization technique in Extension 2 to better capture recent trend. This model takes into account more recent population data to predict the K value and was found through figures 13 through 16 to fit accurately with real data.

Finally, the models with every determined K values were then projected to 2040 and then compared to the projections provided by Stats Canada in figures 17 through 21. These models however differ based on the expected interprovincial migration to predict a range of medium growth [4]. The first extension using a calculated K value from an estimated proportion of 0.25 using California data was modelled in Figure 17. Initially, Figure 17 largely overestimated the population until 2026 but after 2030 estimates a population only above the lowest projections. While this model is in-between the Stats Canada's own medium and low growth scenarios, the derivation for this K value was done through a number of assumptions and guess work that failed to reproduce a valid K value for rest of Canada, and thus we consider that this was a good fit by chance. Figure 18 used the K value found from extension 2, a derived K value from the population growth model, consistently underestimates the population when compared to the Stats Canada projections. Finally, figures 19 and 20 initially overestimated and then underestimated below the lowest population projection. Figure 20 best follows the low scenario described and projected by Stats Canada. Overall, this approach of considering current data to predict the K value is much more effective as the factors that affect population growth change with time. However, this is unlikely to be an accurate K value over time as the K value can change with external factors and the model here is limited only measuring the growth over 60 years, to truly assess the accuracy of this value more growth data is required.

References

[1] Cohen, J. E. Population Growth and Earth's Human Carrying Capacity. https://www.science.org/doi/10.1126/science.7618100.

[2] Data Commons. Canada. https://datacommons.org/place/country/CAN?utm_medium=explore&mprop=count&popt=Person&hl=en.

[3] Data Commons. United states of america. https://datacommons.org/place/country/USA.

[4] Dionne-Simard, D. This data visualization product provides interactive insights on the most recent population projections for canada, provinces and territories. *Government of Canada, Statistics Canada* (Aug 2022).

[5] Federal Reserve Bank of St. Louis. Resident Population in California and US. https://fred.stlouisfed.org/series/CAPOP#0. Since we were unable to find a convenient table which tracked US population data by state since 1955 from the Census Bureau, we instead used data collected by the St. Louis Federal Reserve (which itself collected the population estimates from the Census Bureau).

[6] Greene, J. R. Gerald Ford: The American franchise. *Miller Center*.

[7] Nargund, G. Declining birth rate in Developed Countries: A radical policy re-think is required. *Facts, views & vision in ObGyn 1*, 3 (2009), 191.

[8] Stats Canada. 2011 Census - About the History of the Census of Canada. www12.statcan.gc.ca/census-recensement/2011/ref/about-apropos/history-histoire-eng.cfm#shr-pg0.

[9] Stats Canada. Estimates of interprovincial migrants by province or territory of origin and destination, annual. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710002201.

[10] Stats Canada. Population Projections for Canada, Provinces and Territories: Interactive Dashboard. https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2022015-eng.htm.

[11] Stats Canada. Projected population, by projection scenario, age and sex, as of July 1 (x 1,000). https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=1710005701.

[12] Stats Canada. Population growth in Canada: From 1851 to 2061, Jul 2018.

Appendix: R Code

Base Model Code

Model 1

For base corresponding to Table 1 values and the form:

$$P_t = GP_{t-1}$$

We used the following code (using California and rest of the US inputs as an example):

```
#First Model descrbed in the project. Just adjust the inputs manually
C_in = 12988  #Inital California Population (in 1955)
U_in = 152082 #Inital Rest of US Population (in 1955)
C_1 = 15206 #Second Population for Cali (in 1960)
U_1 = 163040 #Second Population for rest of US (in 1960)

a = 1 + (C_1-C_in)/C_in #rate for Cali
b = 1 + (U_1-U_in)/U_in #rate for RUS

A <- matrix(c(a, 0, 0, b), nrow = 2, ncol = 2)
t_vals <- seq(1955, 2020, by = 5)
popn <- matrix(rep(NA, times = 2*length(t_vals)), nrow = 2,
               ncol = length(t_vals))
popn[,1] <- c(C_in, U_in)
for (t in 2:length(t_vals)){popn[,t] <- A%*%popn[,t-1]}
#Graphing Results
plot(t_vals, popn[1,], type = "o", col = "red", xlab = "Years",
     ylab = "Population", main = "US Population", ylim = c(0,375000))
points(t_vals, popn[2,], type = "o", col = "blue")
legend("topleft", legend = c("California", "Other States"),
       col = c("red", "blue"), lty = 1, pch = 1)
```

Model 2

For the model corresponding to Table 2 and the form:

$$P_t = TP_{t-1}$$

We used the following code (using California and rest of the US inputs as an example):

```
#Second Model. Once again, plug in the appropraite data
stay_in_cal =  12174 #Cali ppl  staying in Cali
stay_in_ous=  150144 #US ppl staying in US
```

```
m_from_cal = 814 #Cali ppl moving to US
m_from_ous = 1938 #US ppl moving to Cali
rto_cal =  (stay_in_cal)/(stay_in_cal+m_from_cal)
rto_ous = (stay_in_ous)/(stay_in_ous+m_from_ous)
rout_cal = m_from_cal/(stay_in_cal+m_from_cal)
rout_ous = m_from_ous/(stay_in_ous+m_from_ous)

A <- matrix(c(rto_cal, rout_cal, rout_ous, rto_ous), nrow = 2, ncol = 2)
t_vals <- seq(1955, 2020, by = 5)
popn <- matrix(rep(NA, times = 2*length(t_vals)), nrow = 2,
               ncol = length(t_vals))
popn[,1] <- c(C_in, U_in)
for (t in 2:length(t_vals)){popn[,t] <- A%*%popn[,t-1]}

plot(t_vals, popn[1,], type = "o", col = "red", xlab = "Years",
     ylab = "Population ", main = "US Population", ylim = c(0,225000))
points(t_vals, popn[2,], type = "o", col = "blue")
legend("topleft", legend = c("California", "Other States"),
       col = c("red", "blue"), lty = 1, pch = 1)
```

Carrying Capacity Models

For all carrying capacity models of the form:

$$P_t = [\frac{(1-G)P_{t-1}}{K} + G]P_{t-1}$$

We used the following code (using California and rest of the US inputs as an example):

```
t_vals <- seq(1955, 2020, by = 5)
p_initial <- 12988
p1_initial <- 152082
r_vals <- 1.170773
r1_vals <- 1.072053
K_vals <- 52000
K1_vals <- 760000
p_vals <- rep(NA, times = length(t_vals))
p1_vals <- rep(NA, times = length(t_vals))
p_vals[1] <- p_initial
p1_vals[1] <- p1_initial
for (t in 2:length(t_vals))
  #p_vals[t] <- r_vals*p_vals[t-1]
  p_vals[t] <- ((1-r_vals)/K_vals*p_vals[t-1] + r_vals)*p_vals[t-1]
for (t in 2:length(t_vals))
  p1_vals[t] <- ((1-r1_vals)/K1_vals*p1_vals[t-1] + r1_vals)*p1_vals[t-1]
```

Extension 2 Code

Extension 2 Code has two components. The first component defines a function that calculates the distance by taking some value of $\hat{K}$ and a given data set:

$$d_t = |P_t - f(\hat{K})|$$

```
SQE <- function(data, k) {
  r <- 1 + (data[2] - data[1])/data[1]
  test_vals <- c(data[1])
  counts_est <- c()
  for (i in 2:(length(data) + 1)) {
    newval <- (test_vals[i - 1]) * ((1 - r) * (test_vals[i - 1] / k) + r)
    dif_est <- abs(data[i - 1] - test_vals[i - 1])
    test_vals <- c(test_vals, newval)
    counts_est <- c(counts_est, dif_est)
  }
  return(sum(counts_est))
}
```

Next, we write the code that minimizes the arbitrary $V(\hat{K})$ function subject to $\hat{K}$[9]:

```
K_find <- function(data) {
  Opt <- c(0, 0)
  K <- data[1]
  K1 <- data[1]+1
  while (SQE(data, K) > SQE(data, K1)) {
    K <- K + 1
    K1 <- K1 + 1
    next
  }
  Opt[1] <- K
  Opt[2] <- SQE(data, K)
  return(Opt)
}
```

Extension 3 Code

Extension 3 code is fairly similar to the two codes in Extension 2, however, the code now takes an input of the β value (represented as just B) and also is calculating the distance as a proportion of actual observed data (the y_t value):

[9] Note that a step-size of 1 was used in the code in order to run the code efficiently.

```
PSQE <- function(data, k, B){
  r1 <- 1 + (data[2] - data[1])/data[1]
  test_vals1 <- c(data[1])
  counts_est1 <- c()
  for(i in 2:length(data)){
    newval1 <- test_vals1[i - 1]*((1 - r1)*(test_vals1[i - 1]/k) + r1)
    test_vals1 <- c(test_vals1, newval1)
    ratio <- (B^(i-1))*(abs(data[i] - test_vals1[i]))/(data[i])
    counts_est1 <- c(counts_est1, ratio)
  }
  return(sum(counts_est1))
}
```

Then as before, we are now optimizing for $V(\hat{K})$ subject to $\hat{K}$ and given some user specified value for β:

```
PK_find <- function(data, b) {
  Opt <- c(0, 0)
  K <- data[1]
  K1 <- data[1]+1
  while (PSQE(data, K, b) > PSQE(data, K1, b)) {
    K <- K + 1
    K1 <- K1 + 1
    next
  }
  Opt[1] <- K
  Opt[2] <- PSQE(data, K, b)
  return(Opt)
}
```

Projections Code

For this section, separately compiled a CSV file with population projection from Stats Canada and ran it alongside our modelled results. The following is an example of the good used generating for generating Figure 15:

```
plot(OntarioProject$Time,
     OntarioProject$L1,
     type = "l",
     col = "black",
     ylim = c(12000, 20000),
     xlim = c(2010, 2040),
     xlab = "Year",
     ylab = "Population",
```

```
     main = "Ontario Projections (Stats Canada Projections vs Ex3b) ")
lines(OntarioProject$Time,
      OntarioProject$L1,
      type = "l",
      col = "black")
lines(OntarioProject$Time,
      OntarioProject$M1,
      type = "l",
      col = "green")
lines(OntarioProject$Time,
      OntarioProject$M2,
      type = "l",
      col = "cyan")
lines(OntarioProject$Time,
      OntarioProject$M3,
      type = "l",
      col = "darkgreen")
lines(OntarioProject$Time,
      OntarioProject$M4,
      type = "l",
      col = "purple")
lines(OntarioProject$Time,
      OntarioProject$M5,
      type = "l",
      col = "blue")
lines(t_vals, p3_vals, col = "red")
legend("topleft",
       c("Low Scenario", "M1 Scenario",
         "M2 Scenario", "M3 Scenario", "M4 Scenario", "M5 Scenario",
         "K = 19585"),
       lty = 1,pch = 1,
       col = c("black", "green","cyan", "darkgreen", "purple", "blue", "red" ))
```

Individual Member Contributions

Note: Each group member was responsible for writing their own individual contributions.

- **Lingyun Huang:** reformatted Python code into R code for extension 2, wrote and proposed extension 1, co-wrote results, reviewed and tested code, wrote code for extension 1, provided code for carrying capacity models, offered an extensive initial rough draft, assisted in model construction, provided code and outputs for capacity models (figures 8 and 9), proof read
- **Wenxi Zhao:** Use the parameters of the basic model to calculate analytical tools, and mathematically calculate the basic model, such as stability, equilibrium, eigenvalue, eigenvector, etc., and perform analysis and prediction on it.
- **Margaret Hong:**
- **Anja Schouten:** Edited the introduction, wrote up background information for the base model, wrote the discussion.
- **Roman Gorev:** Created a shared Overleaf document for the group, proof read Analysis of Base Model (analysis of models (2) and (3)), reformatted Parker's analysis notes into the LaTeX doc, wrote the second half of base model analysis (models (5) and (6)), proof read Extension 1, wrote and proposed Extension 2 and Extension 3, provided graphs 8 through 20, wrote original code for Extension 2 (in Python) and Extension 3 (in R), provided code for Model 1, Model 2, and Projections Code, and provided Stats Canad projection data.