www.simiode.org
SIMIODE
A SYSTEMIC INITIATIVE FOR MODELING INVESTIGATIONS
& OPPORTUNITIES WITH DIFFERENTIAL EQUATIONS

STUDENT VERSION

Exploring Modeling Assumptions with Census Data

Jean Marie Linhart
Gary William Epp
Department of Mathematics
Central Washington University
Ellensburg, WA USA
JeanMarie.Linhart@cwu.edu

1 INTRODUCTION AND SCENARIO DESCRIPTION

The United States conducts a census every 10 years, which gives data on the population of the USA from 1790 to the present. This data is frequently discussed along with the exponential and logistic population models [2, 5, 8, 14, 17]

In contrast, the Republic of Guatemala has taken 16 censuses; its first census was taken in 1778 [12]. The Guatemala census has not been taken at regular intervals, and it is considerably less "smooth" than the data from the United States. There are concerns that some of this census data may not be accurate. For example, according to Squier [12, p. 45] the 1837 census was discredited at the time.[1] Records from the 1940 census were burned; this data is unavailable. Records from 1778, 1880, 1893, and 1921 were used as scrap paper and no longer exist, although the statistical information from these censuses was preserved [11]. To the best of the authors' knowledge, the data for the censuses were not compiled in one easily accessed table until author Gary Epp, then an undergraduate at Central Washington University, put this dataset together after searching out and looking through the references cited and many more as part of a class project in Mathematical Modeling. We also added it to the Guatemala Wikipedia page [18].

While the United States data is considered more reliable and certainly appears to be more smooth, there are also concerns about the accuracy of the United States census data. Few native Americans were included in the census until 1900. They are not identified in the 1790-1840 censuses.

[1] Statistician Don José de la Valle calculated Guatemala had 600,000 inhabitants in 1837.

In 1860, native Americans living in the general population were identified for the first time. Most of the 1890 census was destroyed by fire [15], although its statistical information is preserved. There is always error in census data. For example, consider the 2010 Census. The United States Census Bureau estimates a net overcount of 0.01 percent or about 36,000 people, which is low. However, not every group is counted that precisely. The non-Hispanic white population was overcounted by an estimated 0.8 percent, whereas the native american and Alaska native population on reservations were undercounted by 4.9 percent, and the black population was undercounted by 2.1 percent [16]. While one can argue that these are all small amounts, notice that 0.8 percent is 80 times 0.01 percent, and 4.9 percent is 490 times 0.01 percent. The Urban Institute estimates an overall 0.5 percent net undercount for the 2020 census, roughly 50 times the estimated 0.01 percent error of the 2010 census. They note that, as with the 2010 census, there is considerable variation in who is undercounted and who is overcounted overall. The populations of Mississippi and Texas are estimated to be undercounted by 1.3 and 1.28 percent respectively, whereas Minnesota's population was estimated to be overcounted by 0.76 percent. The Urban Institute notes that if these counts were accurate, Texas would receive over $247 million more and Minnesota would receive $156 million less in 2021 Federal Medicaid reimbursements alone. As is typical, the groups hardest to count in recent censuses were again likely undercounted in the 2020 census. Black and Hispanic/Latinx people have a net undercount of more than 2.45 and 2.17 percent respectively, and households with a noncitizen present were likely undercounted by 3.36 percent overall [1].

Even with the known problems, this data is the best we have to estimate the total population of the United States and Guatemala over the past $\sim$200 years.

We will investigate applying two common population growth models to these two datasets. The models are the exponential model and the logistic model.

Census Year	United States Population
1790	3,929,214
1800	5,308,483
1810	7,239,861
1820	9,638,453
1830	12,866,020
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	39,818,449
1880	50,155,783
1890	62,947,714
1900	75,994,575
1910	91,972,266
1920	105,710,620
1930	122,775,046
1940	131,669,275
1950	151,325,798
1960	179,323,175
1970	203,302,031
1980	226,545,805
1990	248,718,302
2000	281,424,603
2010	308,745,538
2020	331,449,281

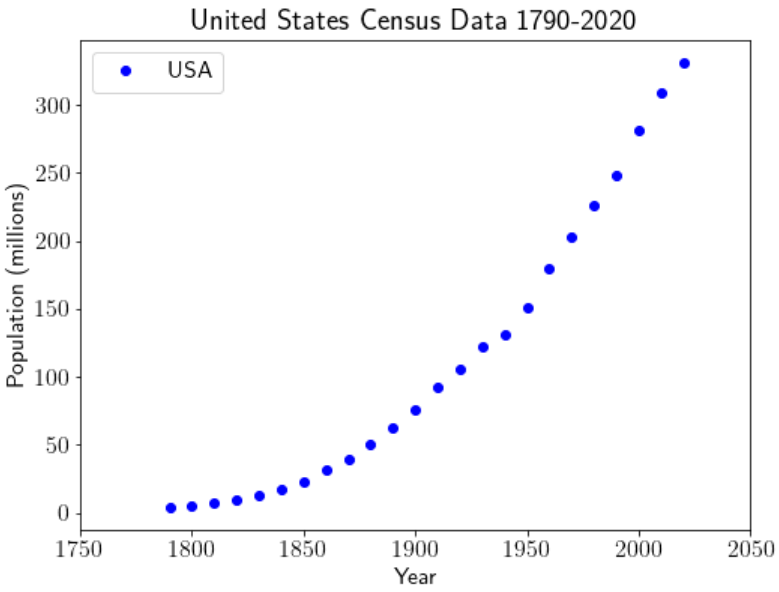

Table 1 & Figure 1: USA Population Data

Census Year	Guatemala Population
1778	430,859 [12]
1825	507,126 [12]
1837	490,787 [12]
1852	787,000[12]
1880	1,224,602 [13]
1893	1,364,678 [6]
1914	2,183,166 [7]
1921	2,004,900 [7]
1950	2,870,272 [7]
1964	4,287,997 [9]
1973	5,160,221 [9]
1981	6,054,227 [9]
1994	8,321,067 [9]
2002	11,183,388 [3]
2018	14,901,286 [10]

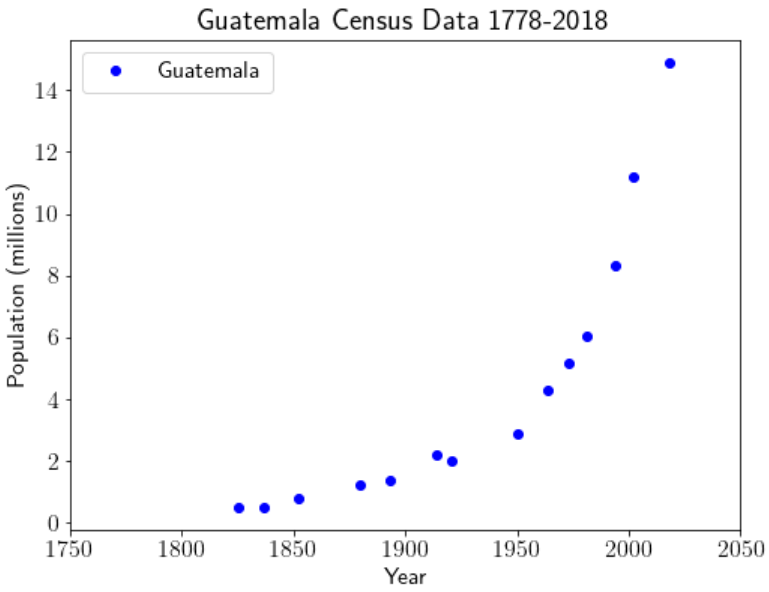

Table 2 & Figure 2: Guatemala Population Data

2 EXPONENTIAL AND LOGISTIC MODELS

The two models for population we are discussing are the exponential and the logistic model. Note that the two models are very similar at the beginning, but one obvious difference between these two models is that the logistic model "levels off" to a maximum population called the carrying capacity, whereas the exponential model grows without bound. This is pictured in Figure 3, below.

One place to start to determine whether one or both models are appropriate for modeling our two sample populations is to look at the graphs of the United States and Guatemala census data in Figures 1 and 2 and the graphs of the exponential and logistic curves in Figure 3, and answer the following questions

1) Does the shape of the exponential curve approximate the shape of the United States census data? Does it approximate the shape of the Guatemala census data? If it does not, describe how the shape of the data differs from the shape of the exponential curve.

2) Does the shape of the logistic curve approximate the shape of the United States census data? Does the shape of the logistic curve approximate the shape of the Guatemala census data? If it does not, describe how the shape of the data differs from the shape of the logistic curve.

Discuss your response with your classmates before continuing.

Note: having a curve or function resemble the data is not enough to conclude that a model is a good one in a given situation. What we really want is for the model to capture important mechanics of the physical situation that gave rise to the data. Next we will look at the equations for these two models and discuss how they came about.

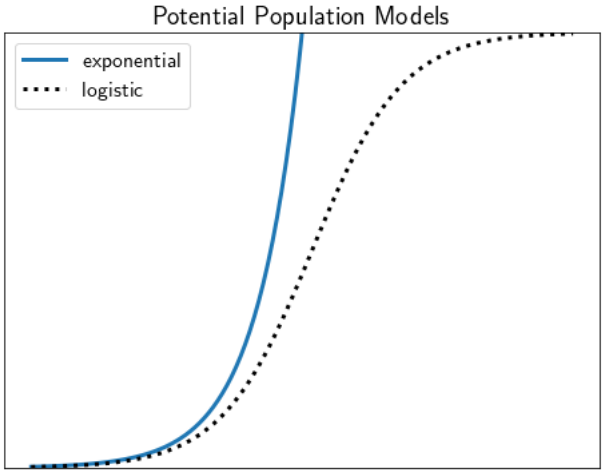

Figure 3: A graphical comparison of the exponential and the logistic population models

2.1 THE EXPONENTIAL MODEL

The exponential model is defined by the differential equation and initial value

$$\frac{dP}{dt} = rP, \qquad P(0) = P_0. \tag{1}$$

It has solution

$$P(t) = P_0 e^{rt}. \tag{2}$$

Note that $\frac{dP}{dt}$ is the change in the population with respect to time, and we have this equal to rP where r is a constant. So the main operational assumption of the exponential model is that the change in the population with respect to time is directly proportional to the population. We assume the number of births is proportional to the population, as is the number of deaths. If a population of 100 people had 2 births and 1 death in a year, then we would expect a population of 1000 to have 20 births and 10 deaths per year, and we would expect a population of 600 to have 12 births and 6 deaths per year. The change in population with respect to time is the births minus the deaths, so for a population of 100 the change is 1 per year, for a population of 600 the change is 6 per year,

and for a population of 1000 the change is 10 per year. The only thing that affects the change is the size of the population.

This is a simple and sensible assumption for population change. But is reality really this simple? Consider the following question.

3) Do you expect the same proportionality constant to work on populations throughout every historical era? Do you expect the same proportionality constant to work for every human population across the globe? Why or why not?

Discuss your response with your classmates before moving on.

When we first create a mathematical model, we try to start with the *simplest model*, and so we make simplifying assumptions. We change assumptions and complicate our model only when we need to. We might complicate our model in order to better match our data or to better answer a question. The exponential model is the simplest population model that is based on a mechanism of population change.

2.2 THE LOGISTIC MODEL

The logistic model is defined by the differential equation and initial value

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{L}\right), \qquad P(0) = P_0. \tag{3}$$

It has solution

$$P(t) = \frac{L}{1 + \dfrac{(L - P_0)}{P_0}e^{-rt}} \tag{4}$$

The logistic differential equation comes about by starting with the modeling assumption that births are proportional to the size of the population, as are some deaths. This is the same starting assumption as we had for the exponential model. Now we consider another cause of death: deaths that are due to two-person interactions. Two people can, of course, have a fight that ends in a death, and one person can murder another. Wars occur. We also consider interactions where people compete for the same resources, which might eventually lead to the death of some of the competitors. Starting from an exponential model (using a for the proportionality constant)

$$\frac{dP}{dt} = aP$$

where births and deaths are proportional to the size of the population, we include a component of deaths that are proportional to the number of possible two person interactions. If there are P people in the population, each one can interact with $P - 1$ other people, thus there are $P(P - 1)$ possible two person interactions. So to include deaths proportional to this we subtract $kP(P - 1)$,

where k is the proportionality constant. The equation becomes

$$\frac{dP}{dt} = aP - kP(P-1) \tag{5}$$

$$= aP - kP^2 + kP \tag{6}$$

$$= (a+k)P - kP^2 \tag{7}$$

$$= (a+k)P\left(1 - \frac{k}{a+k}P\right) \tag{8}$$

We can identify this with our logistic differential equation (3), by letting $a+k=r$ and $\frac{k}{a+k} = \frac{1}{L}$. For an accessible history of the logistic model check out [4].

These assumptions for the logistic model should also make sense. While it is more complicated than the exponential model, this model is also a vast simplification of population change. Consider the following question.

4) What other factors influence the size of human populations like the population of the United States or the population of Guatemala that we have omitted from both the exponential and the logistic model?

Discuss your response with your classmates and make a list of other factors that influence populations that have been omitted from this model.

We see that the exponential model and the logistic model are simplifications of a more complicated reality, but that both make sense and capture some important mechanics of population change. This is why these two models are frequently used to model populations.

3 TROUBLE

As a class we used nonlinear least squares estimation to find parameter values for the United States census data with both the exponential and logistic model as described in [5]. When Gary tried to repeat these same steps with the Guatemala data, he ran into trouble. Visually, the exponential model and the logistic model were identical when graphed with the data, see Figure 4. Upon further reflection, we discovered that the carrying capacity for Guatemala estimated for the logistic model was 2.6×10^8 million people, or 260 trillion people, compared to an estimate of 501 million people for the United States. The high carrying capacity for Guatemala seems preposterous. We also noticed a big discrepancy in the covariance matrices that measure goodness of fit. We noted the variance for the carrying capacity United States estimation was 8.15×10^2 giving us a carrying capacity of 501 ± 29 million where we take the estimate plus or minus 1 standard deviation of error. In contrast the variance for the carrying capacity of Guatemala was 1.00×10^{29}, so for Guatemala we estimate the carrying capacity as $2.6 \times 10^8 \pm 3.2 \times 10^{14}$ million. In other words, the error in the parameter estimation with the Guatemala data for the carrying capacity was larger than the carrying capacity itself.

We know that the logistic model starts out looking like an exponential model when $P << L$, and with the high estimate for L, this is precisely what we saw. In contrast, the parameter estimation and the resulting logistic model for the United States data was good, and its results were an improvement on the corresponding exponential model. See Figure 4.

We wondered, why was this? To uncover the reason, we needed to investigate the per-capita population growth rate.

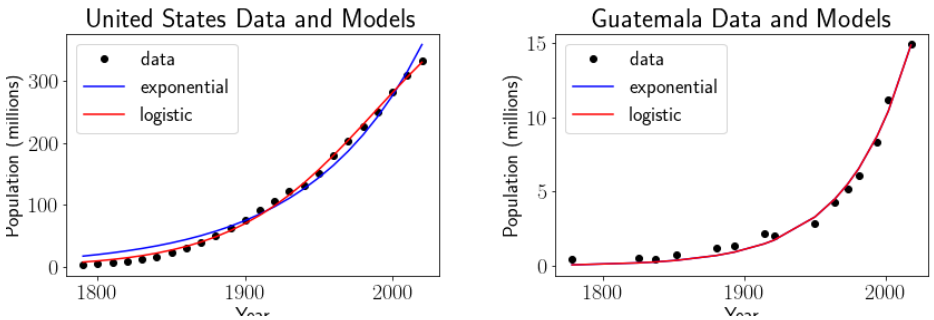

Figure 4: United States and Guatemala with exponential and logistic model curve fits. Note that the Guatemala logistic curve fit exactly overlays the exponential curve fit.

4 PER-CAPITA POPULATION GROWTH RATE

A commonly used concept in discussing population change is the *Per-capita Population Growth Rate* which we will refer to as *PPGR*. Given a population as a function of time $P(t)$, we already know that

$$\frac{dP}{dt}$$

is the change in population with respect to time. Assuming the population is growing, this is what is meant by the *population growth rate*. To get the *per-capita* population growth rate, all we have to do is to divide by the total population:

$$PPGR = \frac{1}{P}\frac{dP}{dt} \tag{9}$$

This is the formal mathematical definition of PPGR. It measures the average change in the population due to one person. For example, let's say that in a year, we have a population of 1000 adult heterosexual cisgender people made up of 500 men and 500 women, and each one of the women has one baby, and no one dies. With 500 people added to an initial population of 1000, we see on average, over the past year, one person contributed 1/2 a person to the population growth. From another angle, notice we have $P(t) = 1000$, $P(t+1) = 1500$. We approximate the derivative using the forward difference formula

$$\frac{dP}{dt} \approx \frac{P(t+h) - P(t)}{h}, \tag{10}$$

and we see this agrees with our calculation of PPGR in this scenario:

$$PPGR = \frac{1}{P(t)}\frac{dP}{dt} \approx \frac{1}{1000}\frac{(1500-1000)}{1} = \frac{500}{1000} = 1/2. \tag{11}$$

In both the exponential and the logistic differential equations, (1) and (3) respectively, the left hand side is solved for $\frac{dP}{dt}$. Thus, we can divide through by population, P, to find the assumptions about PPGR as a function of population. The exponential equation becomes

$$\frac{1}{P}\frac{dP}{dt} = r. \tag{12}$$

The logistic equation becomes

$$\frac{1}{P}\frac{dP}{dt} = r\left(1 - \frac{P}{L}\right). \tag{13}$$

The exponential model assumes a constant PPGR (the simplest possible assumption), and the logistic model assumes a linearly decreasing PPGR with increasing population (the next simplest possible assumption), as seen in Figure 5.

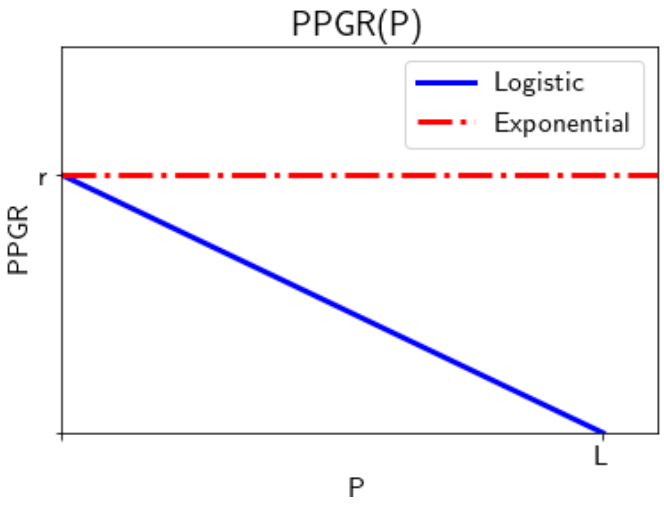

Figure 5: A graphical comparison of the exponential and the logistic PPGR models

4.1 CALCULATING PPGR FROM THE CENSUS DATA

We can use the population data given in Table 1 and Table 2 to estimate the PPGR. Take a look at the data. It is in the form $\{(t_i, P_i)\}$ where we assume $1 \leq i \leq N$, and the data is in chronological order, so $t_1 < t_2 < ... < t_N$. Recall (9) gives the mathematical definition of PPGR.

5) Assuming that population is in units of the number of people, and time is measured in years, what are the units on PPGR?

6) Give a formula to approximate $\frac{dP}{dt}$ from the population data (t_i, P_i). Will your formula calculate $\frac{dP}{dt}$ at every data point? If not, which ones does it miss?

7) Is there more than one possible formula you can use to approximate $\frac{dP}{dt}$? If so, what alternatives are available to you? Is there an advantage to using one over another?

8) Give a formula to approximate PPGR from the population data. Will this formula calculate PPGR at every data point? If not, which ones does it miss?

9) Is there more than one possible formula you can use to approximate the PPGR from the data? If so, what alternatives are available to you? Is there any advantage to using one PPGR formula over another?

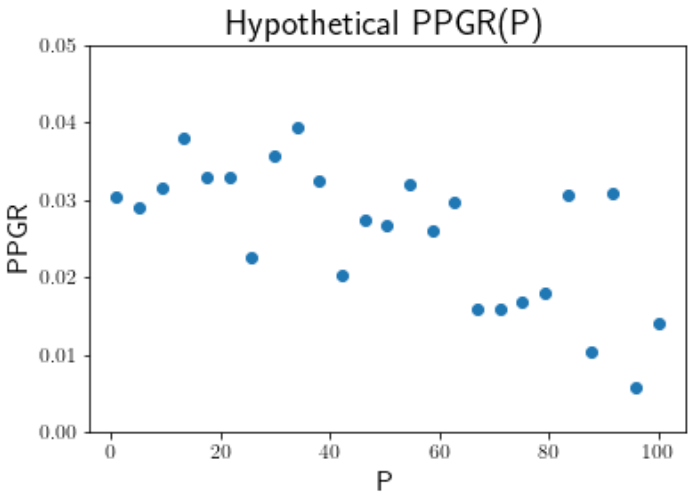

Figure 6: Pretend PPGR(P) data. How would you find the Exponential (horizontal) PPGR line in Figure 5, or the logistic line in Figure 5 from this data?

Now let us assume we have estimated PPGR from the data. In Figure 6, you see a hypothetical plot of PPGR vs. P. Discuss the following questions with your group:

10) How would you find the (horizontal) Exponential PPGR line (see Figure 5) from the data in Figure 6?

11) How would you find the logistic PPGR line (see Figure 5) from the data in Figure 6?

4.2 USING A SPREADSHEET TO MAKE PPGR CALCULATIONS

We can use a spreadsheet and simple spreadsheet calculations to perform these calculations for PPGR, and make a graph of PPGR(P). The data is included with this modeling scenario in two

formats: `.csv` format is a text-based file with "Comma-Separated Values", and `.xls` is an Excel spreadsheet format. The `.csv` version can be opened and understood by any spreadsheet (including the free OpenOffice Calc) and a variety of other calculational software. We also provide the data from this modeling scenario in Google Sheets:

- United States Census Data: `https://tinyurl.com/UnitedStatesCensusData`
- Guatemala Census Data: `https://tinyurl.com/GuatemalaCensusData`

The instructions below are written assuming students are using Google Sheets, which was the spreadsheet the authors found easiest to use. To use the read only data provided with the links above, go to the File menu and "Make a copy." Students should work in small groups of two or three, and start with the United States census data.

12) Create a new column with an estimate of dP/dt calculated from the census data.

13) Create a new column with the estimated PPGR from the census data.

14) Calculate the average PPGR, and fill a new column with this value

15) Insert a chart with a scatter plot of the calculated PPGR and average PPGR estimates on the y-axis vs. P on the x-axis. Compare your graph with the graphs of other groups to make sure you've got the correct plot. If they aren't exactly the same, figure out why. Some common issues are

    - Someone graphed PPGR(P) on the y-axis, but t on the x-axis instead of P.
    - The groups used different formulas to calculate PPGR.
    - Specific to the Guatemala PPGR, note that the data points are not evenly spaced, so in calculating a forward difference approximation to the derivative, the denominator is not a constant $h = 10$ as it would be with the United States data. You have to calculate h from the census year data.

16) Use the Chart editor to add a trendline to your chart for the PPGR data. Label it with the equation of the line.

17) Add a trendline to your chart for the average of the PPGR data. Label it with the average value of the PPGR data

18) Add descriptive axes labels and a title for your plot.

Using forward differences to approximate the derivatives, your charts from Google Sheets should look like those in Figure 7.

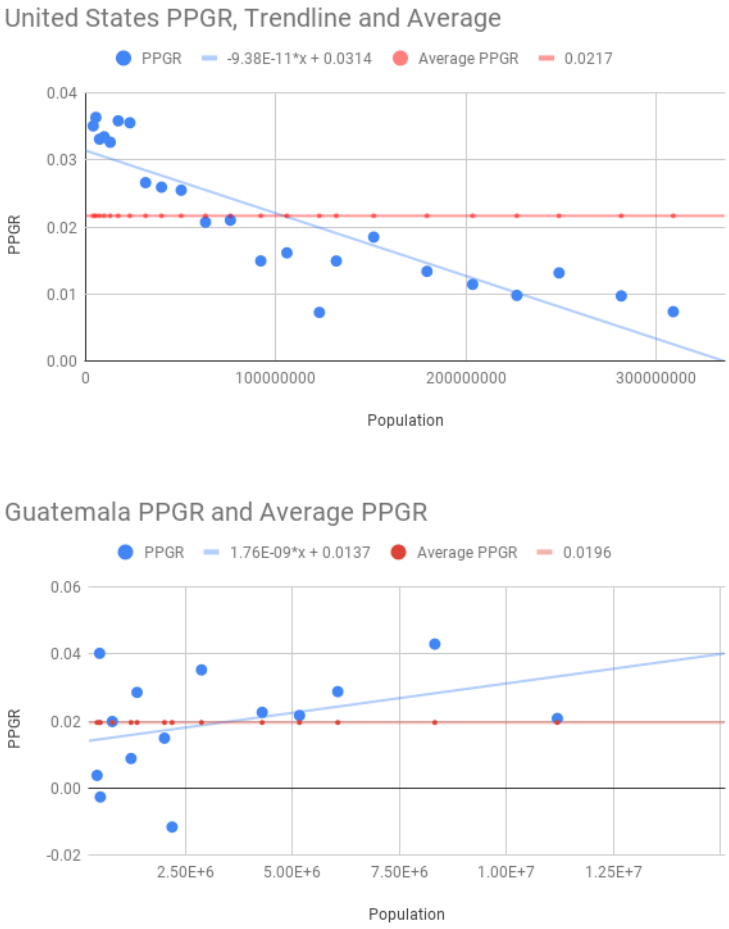

Figure 7: United States and Guatemala PPGR, Trendline and Average

4.3 CONNECTING THE PPGR GRAPHS AND THE POPULATION MODELS

Compare and contrast the results in the graphs from Figure 7. With those graphs in hand,

19) Discuss what you see. Are the pictures in Figure 7 consistent with Figure 5 for both the United States and the Guatemala census data?

20) Use your results to discuss the appropriateness of both the exponential model and the logistic model in modeling the census data for the United States and for Guatemala.

REFERENCES

[1] Elliott, Diana et al. 2021. Simulating the 2020 Census: Miscounts and the Fairness of Outcomes. https://www.urban.org/research/publication/

simulating-2020-census-miscounts-and-fairness-outcomes. Accessed 22 November 2021

[2] Gibson, Campbell and Jung, Kay. 2002. *Population Division: Historical Census Statistics on Population Totals by Race 1790-1990....* U.S. Census Bureau, Washington DC 20233. https://www.census.gov/content/dam/Census/library/working-papers/2002/demo/POP-twps0056.pdf. Accessed 10 December 2021.

[3] Holbrock, M.J. 2016. *Mayan Literacy Reinvention in Guatemala.* Albuquerque, NM: University of New Mexico Press. https://muse.jhu.edu/book/48245. Accessed 10 December 2021.

[4] Kingsland, Sharon. 1982. The Refractory Model: The Logistic Curve and the History of Population Ecology. *The Quarterly Review of Biology.* 57:1, Mar 1982. pp. 29-52.

[5] Linhart, Jean Marie. 2017. *1-066-T-USCensusModeling* and *1-066-S-USCensusModeling*, https://simiode.org/resources/4211 and https://simiode.org/resources/4210. Accessed 29 September 2021.

[6] Loveman, M. 2014. *National Colors: Racial Classification and the State in Latin America.* Oxford, UK: Oxford University Press.

[7] McCreery, D. 1994. *Rural Guatemala, 1760-1940.* Stanford, CA: Stanford University Press.

[8] Mesterton-Gibbons, Mike. 2007. *A Concrete Approach to Mathematical Modeling.* Hoboken, NJ: John Wiley & Sons, Inc.

[9] Monaghan, J.D. and Edmonson, B.W. 2000. *Supplement to the Handbook of the Middle American Indians, Volume 6: Ethnology. Handbook of the Middle American Indians.* Austin, TX: University of Texas Press.

[10] National Institute of Statistics (Guatemala), United Nations Population Fund (UNFPA). *Guatemala Population and Housing Census 2018.* http://ghdx.healthdata.org/record/guatemala-population-and-housing-census-2018. Accessed 30 September 2021.

[11] Platt, L. D. 1998. *Census Records for Latin America and the Hispanic United States.* Baltimore, MD: Genealogical Publishing Company.

[12] Squier, E. G. 1858. *The States of Central America: Compromising Chapters on Honduras, San Salvador, Nicaragua, Costa Rica, Guatemala, Belize, the Bay Islands, the Mosquito Shore, and the Honduras Inter-Oceanic Railway.* New York, NY: Harper & Brothers.

[13] Taeuber, Irene B. 1943. *General Censuses and Vital Statistics in the Americas: An Annotated Bibliography of the Historical Censuses and Current Vital Statistics of the 21 American Republics, the American Sections of the British Commonwealth of Nations, the American Colonies of Denmark, France, and the Netherlands, and the American Territories and Possessions of the United States.* Washington, DC: United States Government Printing Office.

[14] U. S. Census Bureau. 2002. *Measuring America: The Decennial Censuses 1790 to 2000*. U. S. Census Bureau, Washington DC 20233. https://www.census.gov/prod/2002pubs/pol02marv.pdf. Accessed 29 September 2021.

[15] U.S. Census Bureau. Censuses of American Indians. https://www.census.gov/history/www/genealogy/decennial_census_records/censuses_of_american_indians.html. Accessed 22 November 2021

[16] U.S. Census Bureau. 2012. Census Bureau Releases Estiimates of Undercount and Overcount in the 2010 Census. https://www.census.gov/newsroom/releases/archives/2010_census/cb12-95.html. Accessed 22 November 2021.

[17] U. S. Census Bureau. 2021. *Decennial Census of Housing and Population: By Decade*. https://www.census.gov/programs-surveys/decennial-census/decade.html. Accessed 29 September 2021.

[18] Wikipedia Contributors. 2021. *Guatemala*. Wikipedia, The Free Encyclopedia, 29 September 2021. https://en.wikipedia.org/wiki/Guatemala#Demographics. Accessed 29 September 2021.

[19] Wikipedia Contributors. 2021. *Demography of the United States*. Wikipedia, The Free Encyclopedia, 21 December 2021. https://en.wikipedia.org/wiki/Demography_of_the_United_States. Accessed 21 December 2021.