World Populations Throughout History

As represented by the Seshat Global History Databank

by Vitaly Vigasin

Introduction

This research employed data from the Seshat Databank (seshatdatabank.info) under Creative Commons Attribution Non-Commercial (CC By-NC SA) licensing.

The Seshat Databank is an instrument of a new and exciting field of scientific research called *cliodynamics* that applies mathematical methods to the research of the history of human societies. More on cliodynamics and the Seshat Databank can be found in the publications cited below:

Turchin, P., R. Brennan, T. E. Currie, K. Feeney, P. François, […] H. Whitehouse. 2015. “Seshat: The Global History Databank.” Cliodynamics 6(1): 77-107. https://doi.org/10.21237/C7clio6127917.

Turchin, P., T. E. Currie, H. Whitehouse, P. François, K. Feeney, […] C. Spencer. 2017. “Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization.” PNAS. http://www.pnas.org/content/early/2017/12/20/1708800115.full.

Objective

The subject of this analysis is assessing and comparing levels of world populations in different geographical areas throughout human history.

We will test **3 hypotheses**:

A) The rate of a population growth in different geographical regions along the whole time interval of observations/estimates are relatively universal (null hypothesis)
B) The rate of a population growth in different geographical regions along the whole time interval of observations/estimates vary significantly (alternative hypothesis)
C) The higher rate of a population growth will be found in the areas that are currently known for belonging to the most populated areas in the world such as China, India and Pakistan (an extension of the hypothesis B)

Note that we will not attempt to answer any questions with regard to what social or other factors lead to similar or different rates of population growth. Nor will we look for insights for the causes of the observed dramatic differences in the world's populations in the current era. We will only concentrate on 1) a possibility of drawing any conclusions regarding the world's populations based on the data available from the Seshat project as well as 2) testing how the intuitive statements that can be made prior to this research will eventually compare to the reality and whether there will be any counter-intuitive surprises. (For example, one can assume that we will find the rate of the population expansion in the central area of the Roman Empire being one of the highest, if not the highest, in the world considering its power and the place of that particular society in the history of mankind.)

Available Data

I used the Equinox-2020 dataset available for download from the Seshat Databank website.

The dataset contains data on variable features of so-called NGA's - "national geographic areas" - that represent complex societies established in those areas at different times of world's history.

The parameters of those societies listed in the dataset belong to a very wide range of traits and facts such as religion and ideology, territory, military, law, and many others.

Data Analysis and Cleanup

Currently, the dataset contains close to **47,500 rows** and **12 columns**.

Each row contains a record containing a specific piece of information about an NGA and a society (polity) found in the NGA in a particular interval in history, or ever, as well as references, dates, if available, numbers, if available, and many other types of complementary information.

Of all this tremendous amount of detail the one that was of interest for the purpose of this analysis was data category denoted as **“Polity Population”**, with numbers and dates where available.

As only those records were selected from the original file, the resulting dataset contained **395 rows** listing population numbers of each of the **35 NGA**’s shown below (source: http://seshatdatabank.info/)

The cleanup steps consisted of the following:

- Dropping rows that actually contained no population estimates as some of the rows contained ‘unknown’ or ‘suspected unknown’ variables instead of population data (more on the meaning of these variable codes can be found in the Seshat Code Book)
- Replacing missing values with zeros
- Dealing with values (2 rows) that had a range of population estimates as opposite to a single number
- Correcting a likely error for a particular value present in the original dataset (an email report about the finding will be sent to Seshat)
- Removing string values ‘BCE’ and ‘CE’ denoting historical eras from the dates
- Converting dates string values to integers
- Assigning negative values to all BCE dates

Additionally, the following steps were performed to make data analysis more convenient and meaningful:

- Renaming the columns
- Grouping the dataset by polity with columns containing population numbers with corresponding dates where available (see below):

Polity	NGA	Pop_Min	Pop_Max	From_Year	To_Year
AfGrBct	Sogdiana	1500000.0	2000000	-200	0
AfHepht	Sogdiana	26500000.0	0	500	0
AfKidar	Sogdiana	1000000.0	1500000	0	0
AfKushn	Sogdiana	12500000.0	13500000	100	0
AfKushn	Sogdiana	14000000.0	15000000	200	0
...	...	...	...	...	...

_

The cleaned dataframe depicted above resulted in 348 rows. The meaning of ‘Pop_Min’ and ‘Pop_Max’ values is boundaries of the range of the population estimate for a historical interval. This data structure was used for building a time-series plot depicting the growth of populations throughout history.

Yet another structure was created to analyze the *rate* of the population changes for each of the national geographic area (NGA) as shown below:

NGA	Population Min	Population Max	Rate(%)
Deccan	500.0	400000000.0	7.999990e+07
Paris Basin	100.0	28500000.0	2.849990e+07
Kachi Plain	450.0	110000000.0	2.444434e+07
Kansai	400.0	32000000.0	7.999900e+06
Susiana	800.0	30000000.0	3.749900e+06
Middle Yellow River Valley	10000.0	334000000.0	3.339900e+06
Niger Inland Delta	200.0	4000000.0	1.999900e+06

…

In this new dataframe max and min values mean minimum and maximum values per each NGA throughout the whole original dataset (i.e. all historical timeline). These values were used to calculate the rate of population growth that was in turn used as a basis of comparison of various National Geographic Areas.

Findings

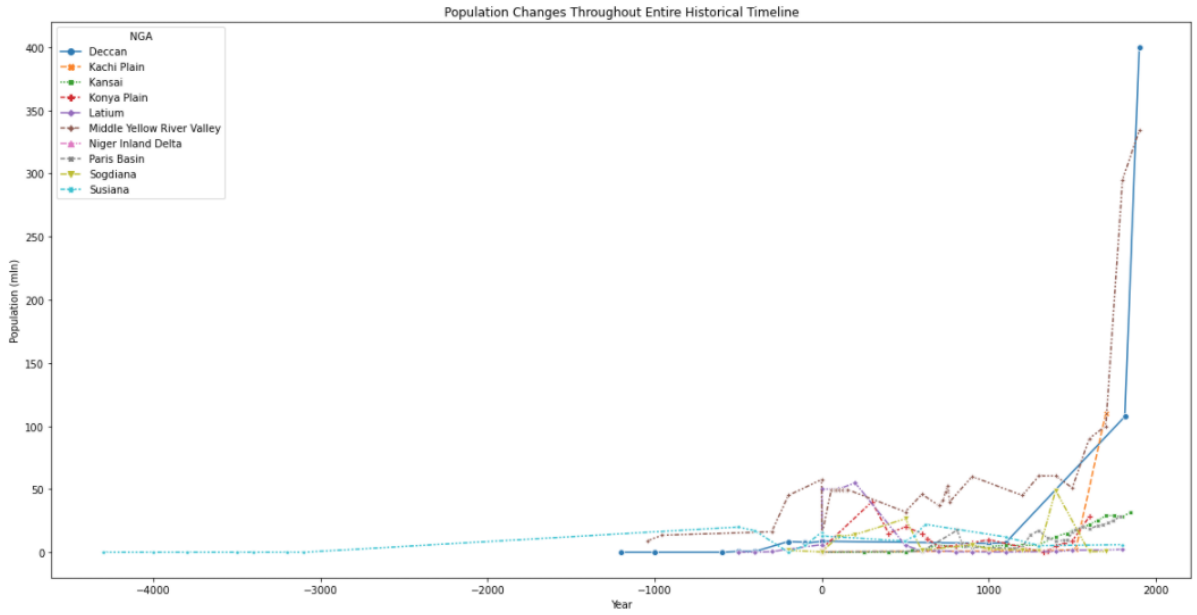
For simplicity, let’s take a look at only top “ten performers” in total population growth throughout entire history:

NGA	Population Min	Population Max	Rate(%)
Deccan	500.0	400000000.0	7.999990e+07
Paris Basin	100.0	28500000.0	2.849990e+07
Kachi Plain	450.0	110000000.0	2.444434e+07
Kansai	400.0	32000000.0	7.999900e+06
Susiana	800.0	30000000.0	3.749900e+06
Middle Yellow River Valley	10000.0	334000000.0	3.339900e+06
Niger Inland Delta	200.0	4000000.0	1.999900e+06
Konya Plain	2500.0	40000000.0	1.599900e+06
Sogdiana	15000.0	49000000.0	3.265667e+05
Latium	20000.0	55000000.0	2.749000e+05

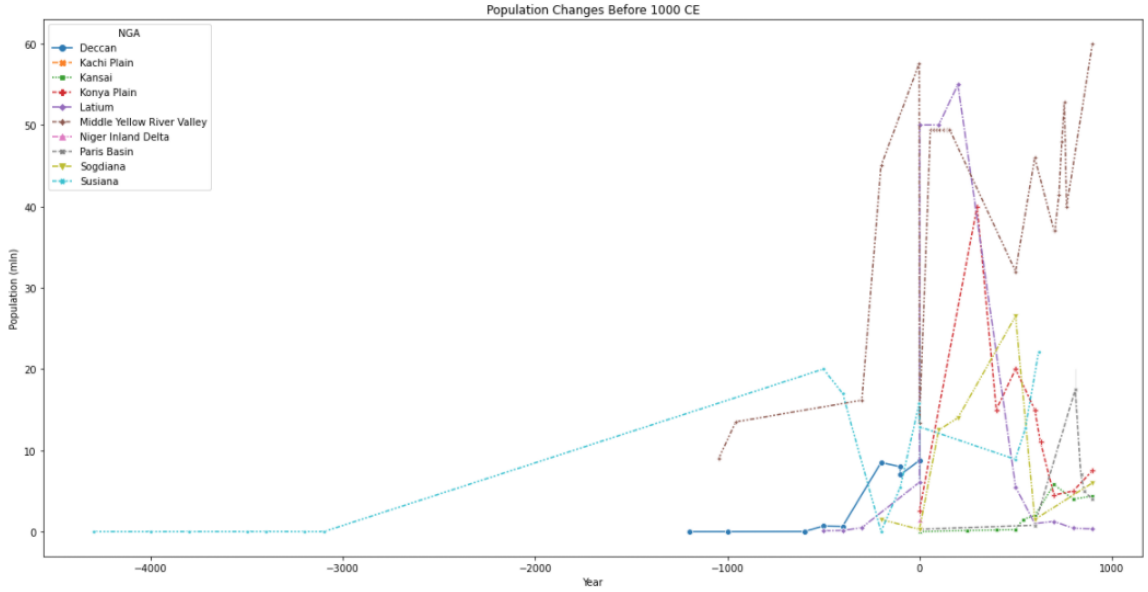
As demonstrated by the following plot, population changes have clearly happened not at the same rate for all of the geographic regions in question.

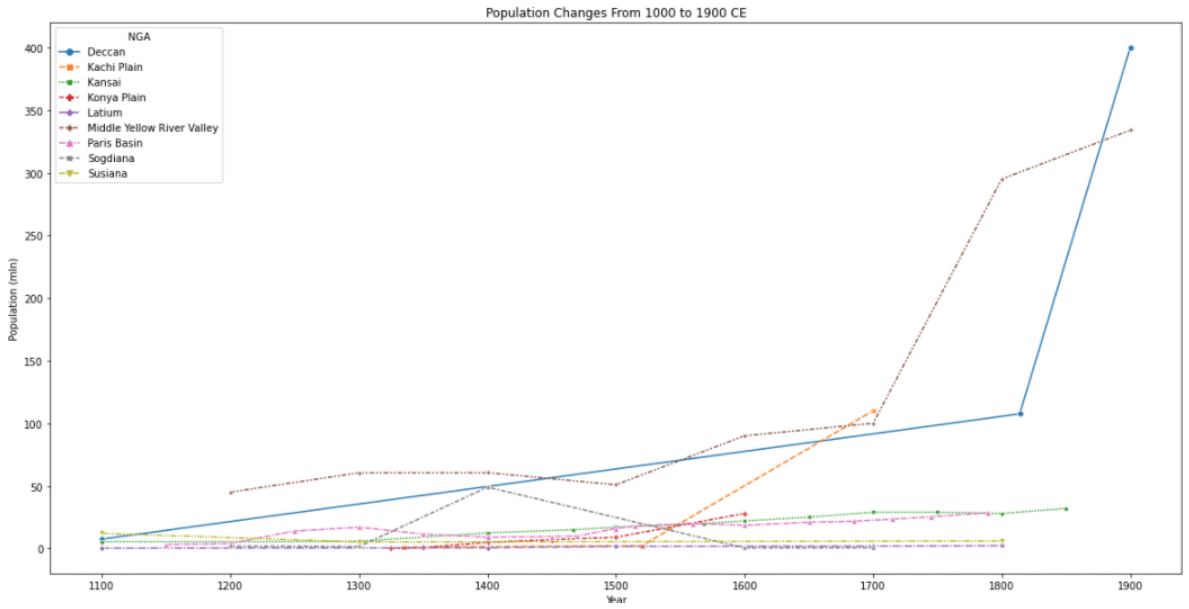
Remarkably and but perhaps not surprisingly, two regions located in modern China and India have experienced an explosive growths of

population counts which have been times higher than those of the rest of the regions.



Some interesting observations can be made if we zoom in on two distinct eras before and after 1000 CE:

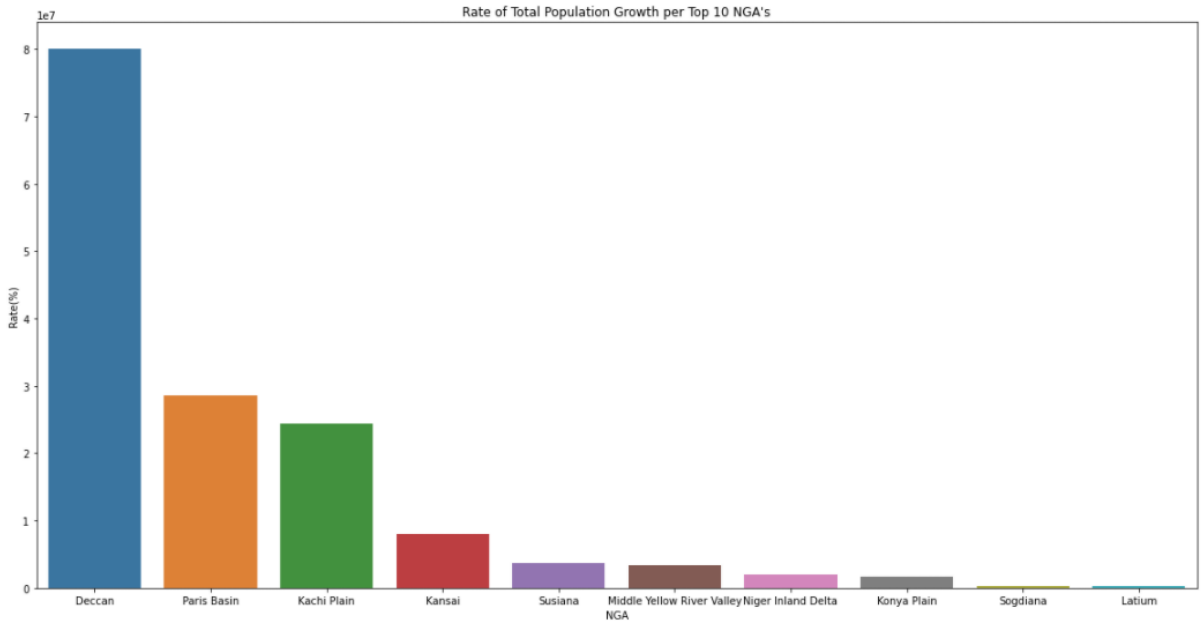




As we can see, there was a significant volatility in the first millennium of the Common Era in terms of population counts, with most of the civilizations' populations swinging up and down dramatically while the last millennium saw much more stable growth of the populations with those in China and India enjoying a very steep spike in the last 300 years (the beginning of which coincides with the beginning of the colonization era of those territories, which is an interesting fact by itself).

What is more important for the purpose of testing our hypotheses stated above is that we can easily **disprove hypothesis A** as false while **confirming hypothesis B**. Without making a calculation, we can clearly see that not all regions have similar rates of growth (hypothesis A). On the contrary, the plot confirms that the civilizations (or rather geographical areas within them) differ in terms of population growth (hypothesis B).

Now, let's take a look at yet another plot:



Here we see the rates of population growth of the same 10 regions we have examined before.

Unexpectedly, the NGA named “Paris Basin” (the area around Paris, France) comes second as the fastest growing region in history. By that we simply mean that based on the Seshat data for the whole time interval the data exist the Paris area had second highest growth of population - even if for just several centuries and peaking at 28 million 500,000 (in the year 1789 after which there are no data in the dataset for this particular region) as opposite to thousands of years of development and hundreds of millions of people in the regions of China and India for which there are records.

Thus, we can conclude that **hypothesis C is wrong**.

Conclusion

We have disproved hypotheses A and C and confirmed hypothesis B and can now state the following:

1. The rate of a population growth in different geographical regions along the whole time interval of observations/estimates are NOT universal
2. The rate of a population growth in different geographical regions along the whole time interval of observations/estimates DO vary significantly
3. The higher rate of population growth has not been found confined in the areas that are currently known for belonging to the most populated areas in the world such as China, India and Pakistan.

Quality of Data

Naturally, correctly estimating, let alone, scientifically confirming levels of populations of societies existing hundreds and thousands years ago present a tremendous challenge for the historical research. Naturally, any analysis is as good as its underlying data. The findings of this research are based on the data currently available from the Seshat project. Unfortunately, there cannot be any other source, or at least any better source than Seshat for the analysis presented in this paper. Even more unfortunately, there are, of course, many holes in our knowledge about ancient societies and, therefore, lots of missing data in the Seshat Databank that would otherwise lead to more interesting insights. That said, any new data about human populations that becomes available in the Seshat Databank can be analyzed using the same framework described here which, hopefully, may provide more food for thought in the future.