US Migration Simulator Analysis Report & User Manual

By Python SuperSonics

Branavan N, DJ W, Gahl G, Tarcisius H

Github: https://github.com/djwadhwa/USMigrationSimulation

USER MANUAL

In order to run our simulation model, a user needs to have two files with data representing their city.
One is a .csv file, in which each row contains data about one year in that city. Rows are formatted as follows:

<year>,<avg. rent cost>, <avg. tax>, <# of crimes>, <population size>, <# of jobs>

In addition to the .csv file, a user will need a .py file called <city name>.py. This code file will follow a specified template. It parses the .csv file for information, and also provides our program with information about distribution of adults vs. children within the city's population.

Once the required two files are assembled, the user will need to decide for how many years they want to run the simulation, and also how many trials they want to do. More trials bring greater accuracy. If the user selects not to provide a number of years or trials, our program defaults to 20 years and 100 trials.

In order to use the program, the user must import their city.py file city at the top of the simulation.py file. From there, calling the function runModelTest(city, output_file_name, num_years, trials) will run the simulation. The program will output 3 graphs predicting the amount of food consumption, water consumption, and population growth. It will also print to output the yearly population values, as well as absolute and relative error of the model relative to the actual yearly data.

We have also included a driver to make it easier to run the simulation and print out the results.

MODEL DESCRIPTION

In this simulation, we are simulating 4 cities - Seattle, Chicago, New York City, Los Angeles - and how their populations change over time with respect to migration into those cities. In this particular simulation, we are simulating a 20 year period, however, the model has the ability to scale the simulation even further. Now, in order to estimate future populations, we have to take into account several different factors. We have determined after research that these are the most important factors that will affect a city's population: Jobs, Crime Rate, Average Cost of Living (calculated with rent prices), and Taxes.

We decided on these four cities because while they are very similar in terms of having the fast, growing tech industry in these cities. They are also very different and each of these 4 factors we selected is completely different in each city. The crime rate in some of these cities is significantly higher than the others. Also, the average cost of living will be dramatically different between all four of these cities.

Our main focus is on Seattle and Chicago because these are two cities that are completely different from each other in terms of the four factors we are measuring. Seattle has less crime but the average cost of living is much higher in Seattle than Chicago. Seattle, Los Angeles and New York are very similar cities in many aspects so we will use Los Angeles and New York to confirm our model for validation. This will be a good way to see if our model is accurate since it can account for major differences in between city population.

We wrote our model first accounting for Seattle because we wanted to initially streamline our research method, only citing our data from reputable sources. Furthermore, this made the initial design and implementation simpler without much abstraction. We then chose to use our model to determine the population of each aforementioned city.

FACTORS EXPLANATION

In predicting what factors would be important to determine the change in population in a city over time, it took some extra research to determine what factors would be the best to use. For a simulation analyzing the change in a city's population, there is nearly an endless supply of factors which could be used to determine people's migration patterns There is also a wealth of data on all of those factors which could be used in making our model. Since we have a limited amount of time, it would be unrealistic to implement all of these factors, which is why we decided to go with the 4 most important factors.

How we decided to pick these factors was based on several different aspects. Jobs and cost of living were pretty obvious to pick because most people aren't going to be able to move to a new city if they can't financially afford it. The other two factors were the most challenging to pick since for these factors they aren't as obvious as the first factors, and because of that, we would have to narrow our choices down from a bigger list.

We first thought that education would be a good factor to use, because if you have kids then schools in the area will definitely be an important reason why you would move to a particular city. Our plan was to find data on school ratings in different cities across different years and cross-reference that data with population fluctuations in those cities. However, finding data for this turned into a large challenge for a few reasons. The fact that an individual city contains several different school districts, for one, muddled the process of reliably measuring any city's academic ranking. Another challenge we encountered was finding a reliable, unbiased, trustworthy source of data on school ratings, so as to uniformly rate schools as "good" or "bad." We thought finding data about schools' test scores could be a workaround to this issue, but it was also difficult to assemble a sufficiently large data pool for that factor.

We decided on the crime rate for a couple of reasons. One reason that we chose this factor is that, if you are deciding whether or not to move into a new city, the crime rate is indicative of what your decision may be. Although the crime rate is a factor which people do not always directly look at, it is indicative of the more acute "how nice is the area" factor which people almost always do look at. Another reason why we picked this factor is that there is a lot of very accessible official data. Each US city's police department or state releases the total amount of crime that has occurred on a single given year, and we were able to find a lot of this data and use it in making our predictive model.

Jobs

Jobs are one of the main reasons people move into a new city. In the modern day, migrants usually have a job ready for them by the time they chose to migrate into a city. We can, therefore, assume that most of the jobs in our model are fulfilled year-to-year. With some research, we also determined that jobs generally increased at a steady rate.

In the last 20-25 years jobs increased by about 2 - 2.5% per year and that is what we used in our simulation to calculate the potential jobs for future years. Here is the table for the data:

The fluctuation of job rates based on the previous table makes sense to us because in our process of gathering data, we found several reasons that can back up the changes of job availability rate in Seattle, Chicago, Los Angeles, and New York City:

Seattle:

Seattle has seen consistent growth in its job markets due to several different factors. Tech companies investing in the area, online retail trade, and the healthcare industry. All of these industries have been investing and growing in the city for the last several years. This has led to a constant influx of new people moving into the city due to work. Amazon has had a big impact on this due to its investment in the online retail space leading to an increase of 2.11%. All of these factors are the reason how we came up with the 2-2.5% increase per year regarding jobs for Seattle.

Source:
https://www.seattletimes.com/explore/careers/job-outlook-2019-seattle-has-a-very-innovative-workforce-and-no-one-wants-to-miss-the-party/

Chicago:

Chicago's job rate has been the opposite and it has been fluctuating up and down over the last several years. In fact, Chicago's job market has been increasing at a 1.5% rate which is lower than the national average of 1.7%. This is why we wanted to pick Chicago as one of our cities to examine because it has a lower job rate increase than the other cities and it also has a dwindling population over time. Using Chicago in our model can help us determine how much the job rate will affect the migrant population.

Crime Rate

The crime rate in a city is another factor regarding migrants moving into a new city. If a city is on a trend of an increase in crime, then obviously, migrants are not as likely to move to that city. In our simulation, if the crime rate reaches a certain amount then it will have a negative effect on the migrant population and fewer people will move into the city. On a general trend, the crime rate across the country has been going down and, as a result, the crime rate in many cities have been going down as well. We got our data for the crime rate from the police department of that specific city. This had the total crimes committed on a yearly basis in that city.

Seattle:

Seattle's crime rate was pretty interesting to see as it was on a steady decline until 2013 where there was a drastic increase from 35000 to about 40000 crimes committed in a year. This can be the cause due to the increase in the population in recent years. We believe this because the major increase in crime was regarding property crime rather than violent crime. Property crime includes things such as burglary, larceny, theft, motor vehicle theft, arson, shoplifting, and vandalism. These are all crimes that are more likely to happen due to an increase of population. If there are more people in the city then there will be more victims as a result of that increase.

Chicago:

Chicago's crime rate has changed over the last 15 years in different ways. From 2006 to 2014 crime has been on the decline. Starting at 165,235 total crimes in 2006 to 99432 in 2014. , The interesting thing is that the population has been increasing over this time and the crime rate was still going down. However, once the population began to decrease the total amount of crimes began to increase. In 2018, the total amount of crimes was 119,060. This example is why we thought the crime rate would be a good factor in determining population growth/decay as you can see that the population is decreasing while more crimes are happening.

Sources: http://www.city-data.com/crime/crime-Los-Angeles-California.html
http://www.city-data.com/crime/crime-New-York-New-York.html
http://www.seattle.gov/police/information-and-data/crime-dashboard

Cost Of Living (Rent)

The next big factor after jobs is the average cost of living. By the cost of living, we mean rent, as that is where most people's income will go towards and it is something that is easily verifiable online. Migrants aren't likely to move to a place with a higher average cost of living so we have to figure out how the rate of rent changes in the city over time. Similarly, to the other factors, this factor has also been increasing on a yearly basis. This is mainly because of the tech-centric aspect of these cities making it attractive to people who are working or plan to work in the tech industry. The rent varies from city to city but it will generally increase by 1%- 5%. This is where the data from the Department of Numbers comes in handy. This website has the rent for every major city in the US for the last several years. This data also has the countries average compared to the states average compared to the cities average over the years. In our simulation, we will assume that the higher the cost of living the less likely of a chance that a migrant will be willing to move into that city.

Source: https://www.deptofnumbers.com/

Seattle:

Seattle's cost of living has changed heavily within the past 10-15 years. The rent has been increasing at a 5-7% increase. This is due to the tech companies heavily investing in the area leading to people moving in and since there has been an influx in population the rent will increase as a result of that. This may or not have an effect on the incoming population and we would like to see if that is true.

Chicago:

Chicago's on the other hand, while the rent is increasing it is only increasing by about 1.15% at a time. This increase can be explained by inflation so perhaps this slow increase in rent could explain the decrease in the population. Since the rent is only increasing due to inflation, it could be a determining factor in the decrease of the population in Chicago.

Taxes

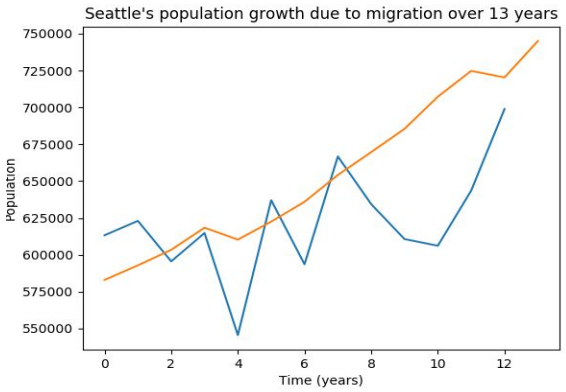
The final factor we will include in our simulation are the tax rates in each city. While the tax rate is not going to be as significant as the other factors regarding the migrant population increasing or decreasing we still feel that it is still an important decision regarding someone decides to move to a new city. How we determined this through our research we have found that certain countries that increase their taxes have had a drastic decrease in their population. We have also found in our research that taxes have generally been increasing over the years. We got our data for this from the Department of Revenue for each state. They have documents for the tax rate for each city through the years and we can use that data to find the trend and pattern of how often the taxes will change either positively or negatively. In our simulation, we will assume that if the taxes are higher than that will reduce the chance of a migrant moving to that city. However, since taxes aren't as important to people moving in as jobs and cost of living are, this factor will not be as significant as the others regarding the population changing.

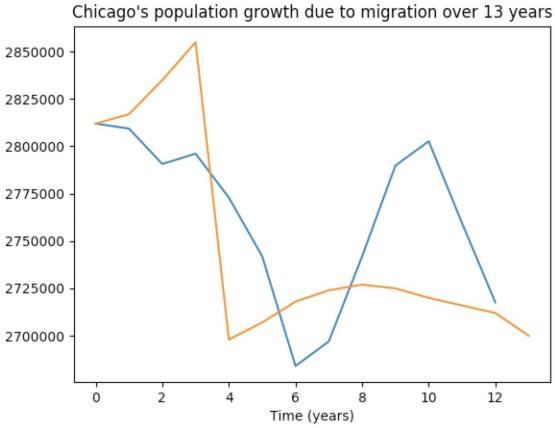
Sources: https://www.washingtontimes.com/news/2017/may/22/taxes-cause-population-to-leave/

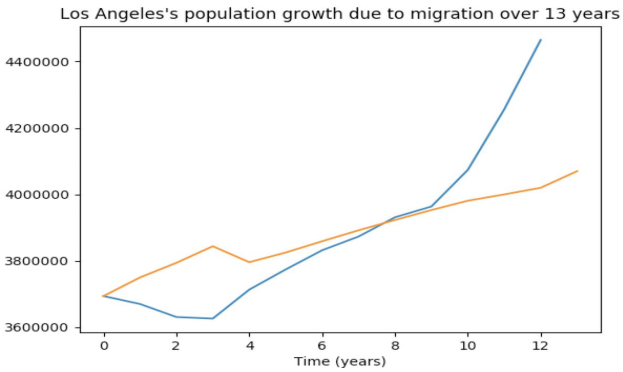
EVALUATING THE MODEL

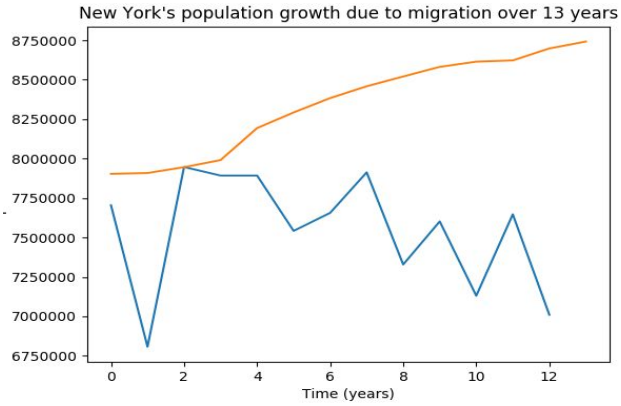
Before Bug Fix

All 4 Cities with default settings









In our first piece of analysis we will look at all the cities population due to migration over a 13 year period. This simulation had 1000 trials per year and calculated the mean for each year. We picked 13 years which is from 2006-2019 because that was the oldest data that we could find for all the factors from valid sources. We compare our model simulated population to the cities actual population and the orange line is the actual population while the blue line is our model's simulation.

The settings for our model in this graph are that the starting jobs, population, percentage of adults and children are all from the year 2006 and the years after that we will simulated and

compared to the actual data and then it will be graphed. Also, future jobs with be implemented at a 2-2.5% increase based on the national average of job increases.

As you can see, the graph varies between being accurate and inaccurate. Our model for Seattle and Chicago match really well are very close representation of what the actual value is. However, for New York and Los Angeles our model doesn't simulate their populations accurate and has some issues.

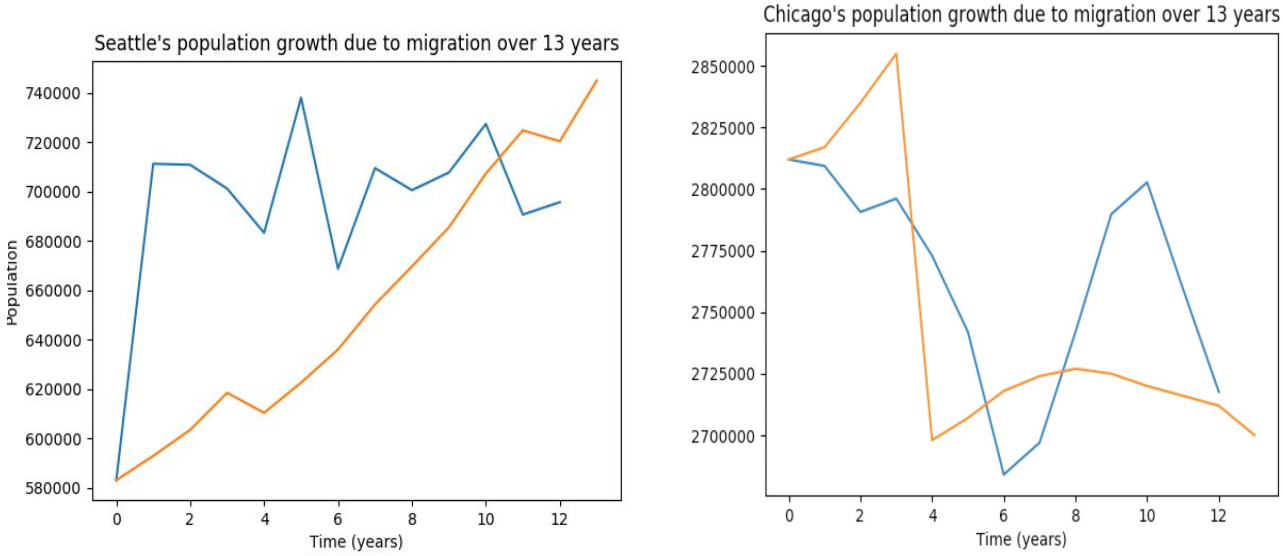
There are a couple reasons why we believe this. This is because we did most of our testing with Seattle first since that was the first city we implemented in our simulation and as a result our model is much more accurate to Seattle. Also, there is much more to a population change then the factors we picked and there could be alternative events that we cannot simulate to account for this population growth. Also, the job growth in Seattle is very similar to the job growth average in the country so this could be a possible reason why it works well with Seattle.

Now we wish to check if our variables are having an effect on our population so in the following sections we will change different variables and see how that affects our model. Will our model be more accurate? Will it stay the same? Or will it become less accurate?

In the following analysis we will just analyze Seattle and Chicago due to their difference in their factors with Seattle having a increasing population and Chicago having a steady declining population. Also, Seattle and Chicago are the city in our model that is the most accurate so it will be interesting to see how these different factors affect our model for those cities.

The first aspect we wish to test is the job increase rate. Obviously, the job increase rate is not uniform in each city and fluxuates so using the national average for the jobs like we did in our original sim would not be a good idea. Also, if we change the job rate to account for the city then the population growth/decay can be accounted for because if there aren't enough jobs then people aren't going to be able to move into the city. We made this change using the US census for each city and seeing the percentage of adults are working and use that to account for the total jobs.

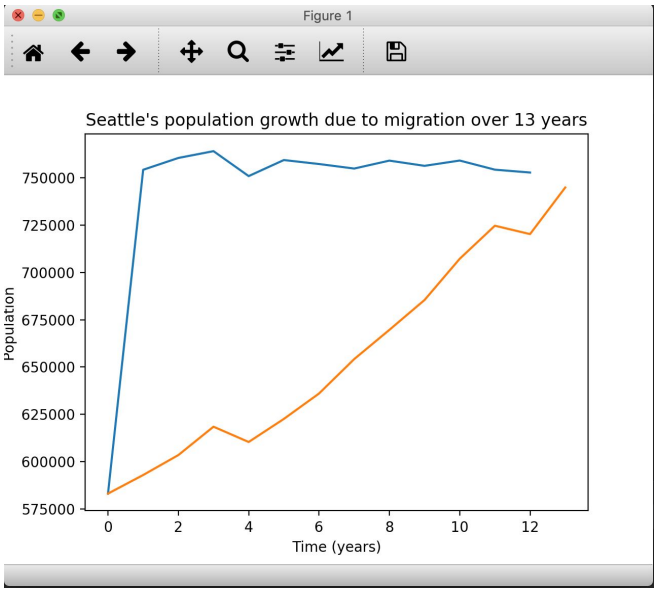
The city provides its own job rate

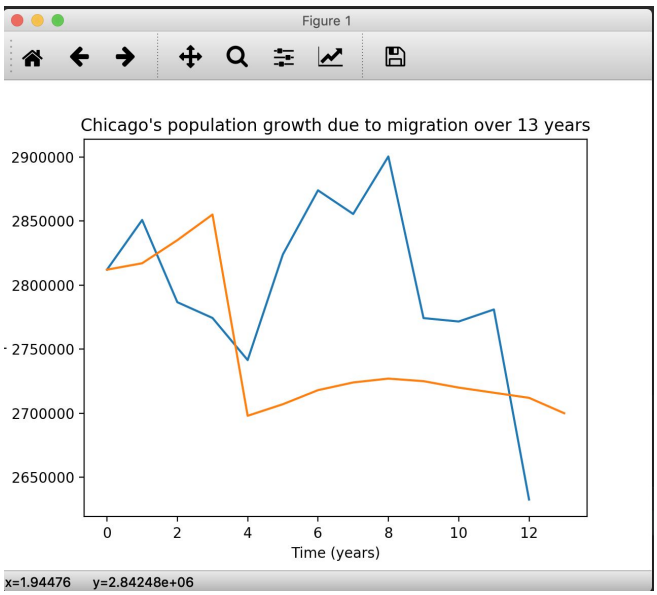


As you can see, Chicago's graph is much more accurate when we account for job rate and the absolute error was 41,432.243 compared 105,234.984 in the previous simulation but this also had an effect on Seattle's population. The relative error for Chicago was 5.432 for this simulation and the last simulation was 9.423. The error was very similar for Seattle between the first simulation but the graph was in the opposite direction. We believe this happens because by changing the job rate, our model can calculate population decay much more accurately

The absolute error in the previous graph for Seattle was 50,232.972 in this simulation when the city provides it job rate it increased to 60,320.534. This can be explained because the city job rate is very similar to the national average and we knew that when we used the national average Seattle worked just fine so changing to the provide its own job rate will have a very small effect on its population.

Changing the random distribution of jobs from normal to binomial





In this image, we changed the random distribution of jobs from normal to binomial.

Our thought process behind making the jobs in each city binomial is that realistically employers will generally hire at the same rate per year rather than mass hire one year and don't hire another year.

The results when we changed this was very surprising since from year zero to year 1 there is a super huge jump from 575,000 to 750,000. After that it alternates from increasing and decreasing for a smaller amount each year. The absolute error for this graph 95314.988 and the relative error was 12.583. This relative error is not very good and it shows that when we change the random distribution to binomial the error is about 12.5% off the actual number.

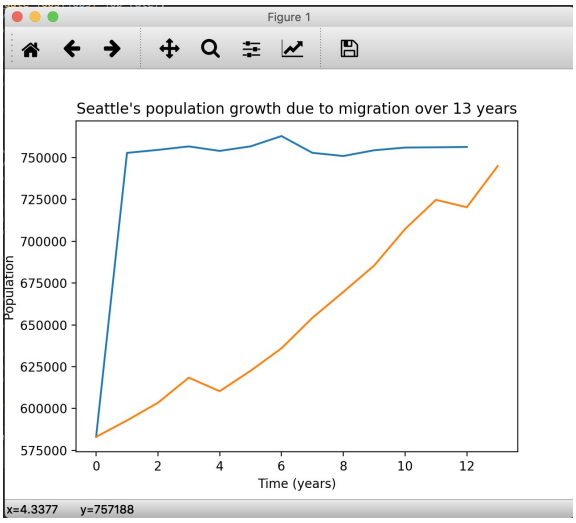
This is how Chicago was affected when we changed the random distribution of jobs from normal to binomial.

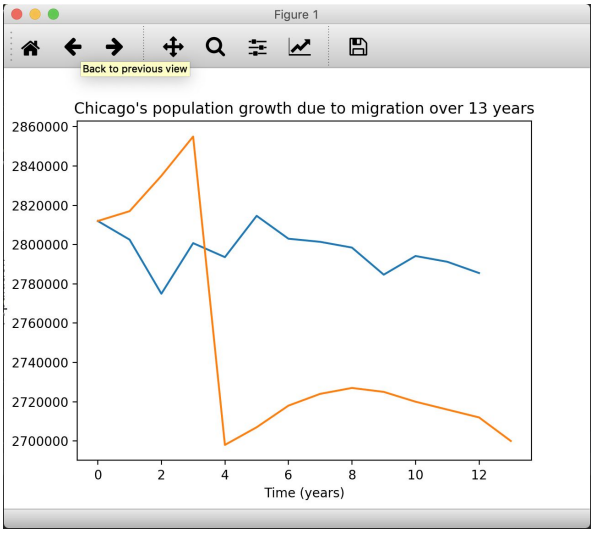
When we changed the job rate it started off very accurate for the first couple of years then afterwards there was a sudden surge in population which was unexpected.

The absolute error and relative error was 85,861.628 and 3.037. This is because even though the population sudden spikes very high it offsets this value by decreasing very low making it seem

like there is only a 3% of error. This seems to be an issue with our model in that it's very hard to accurate depict a good estimate of the future population due to the several different factors.

Changing the random distribution of crime to binomial





We decided to change our distribution of crime to binomial because the amount of crimes committed in a year is a discrete amount.

The absolute error was 66,066.589 and the relative error was 2.360.

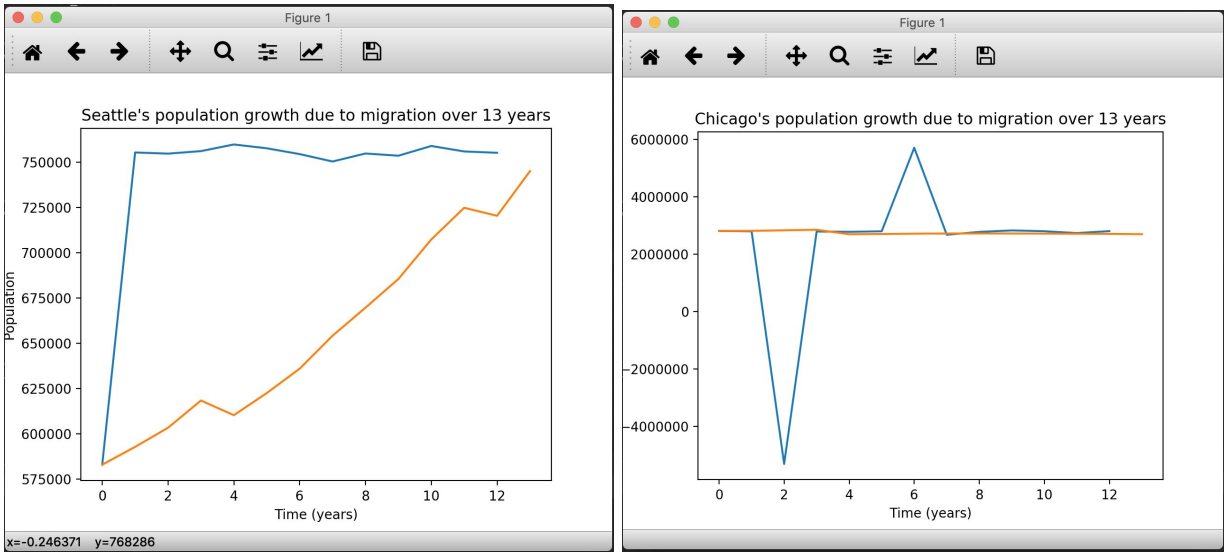
Changing the crime rate to binomial had a negative effect on our model and it actually made our simulation less accurate so we decided to change it back to normal.

Changing the crime rate for Seattle resulted in a drastic increase in year 1 but after that it levels out and in year 10 it begins to match the actual population.

The absolute error is 95,336.728. This is similar to when we changed the jobs from normal and binomial. We feel like this is happening because of that huge jump in year 1 it creates a big error.

The relative error is 12.594, this is not a very good sign that changing the crimes to binomial will make the model accurate.

Changing the random distribution of taxes from uniform to normal

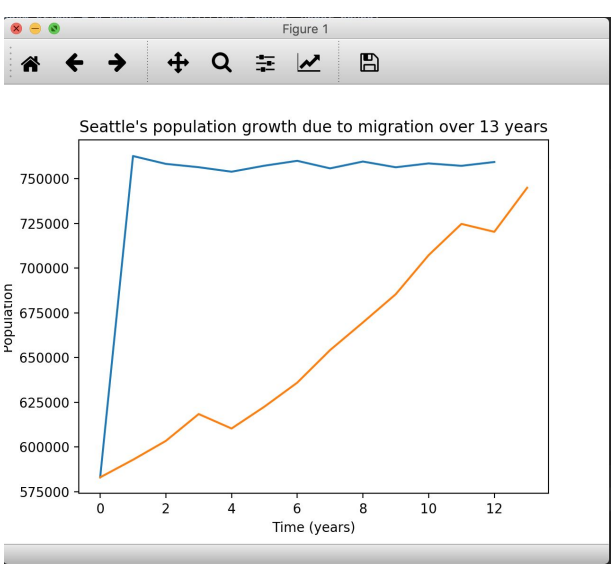


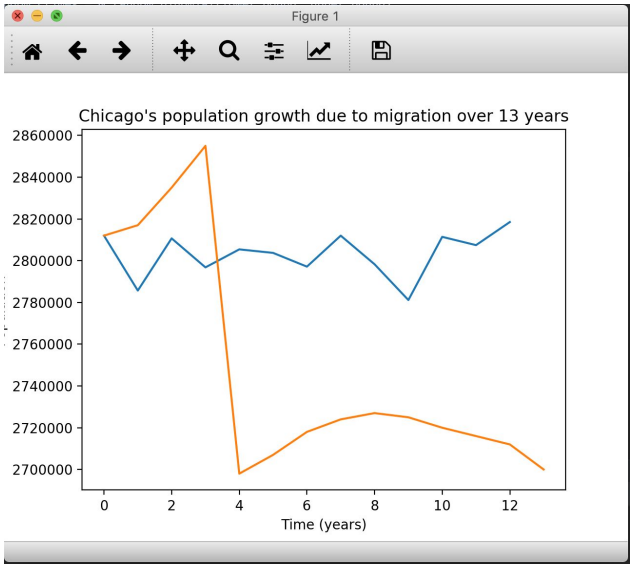
We decided to change the distribution of taxes from uniform to normal to see the changes that happen to the graph simulated from our model.

Our thought behind this change is that we assume the taxes for each city changes around the initial tax that we have in our simulation for each city. The taxes changes based on the certain trend that each city has for its changes in their taxes data.

As you can see, when we change the taxes' random distribution from uniform to normal, the graphs from our simulation that represents changes in population per year have bigger errors in comparison to the actual data that we gather for each cities.

Changing the distribution of rent from uniform to normal.

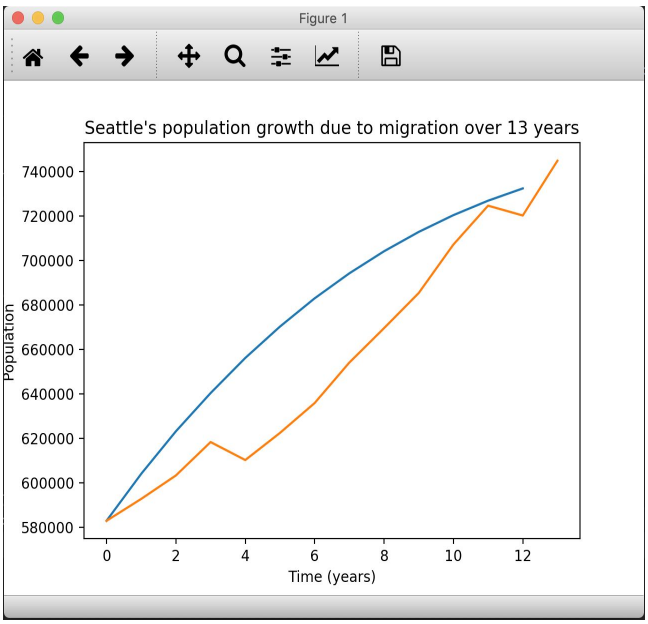


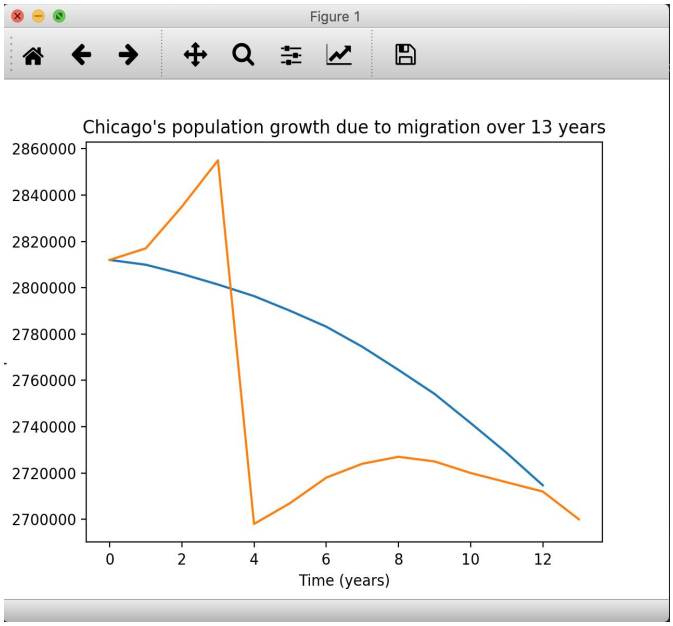


We decided to change the rent from uniform to normal because rent can be portrayed as a gradual change increase in an even way with no bias.

The absolute error is 95,799.667 and the relative error is 12.646. This graph is very similar to the graph when the jobs were changed from normal to binomial. So this factor has a similar effect to that change.

Fixed Major Bug in Model!





We just fixed a major bug in our simulation that makes our model of the cities much more accurate. We discovered a bug in the part of our code which was vastly increasing our error margin. The bug was rooted in our testing system, where we ran many trial simulations, and calculated the average value of the four factors for each timestep across all trials. When trying to find the average value of the entire list of the four values for each timestep, the numpy average function was averaging the entire list (all data is added and divided by the total of the data). After fixing this, it would only average values from the given timestep, the error rates improved and our model closely matches the actual data.

As you can see, the graph for Seattle and Chicago look a lot more accurate and the absolute error for Seattle is 24,914.225. The relative error is only 3.682. This means our data is only 3.6% off from the actual result and with the graph of our model looking realistic we can determine that this 3.6% is a good measure of accuracy.

The absolute error being 24,914 means that over our entire simulation our population is only off by 24,914 people. This is significantly better than what we had before which was over 100,000 and the relative error was over 15% for those simulations.

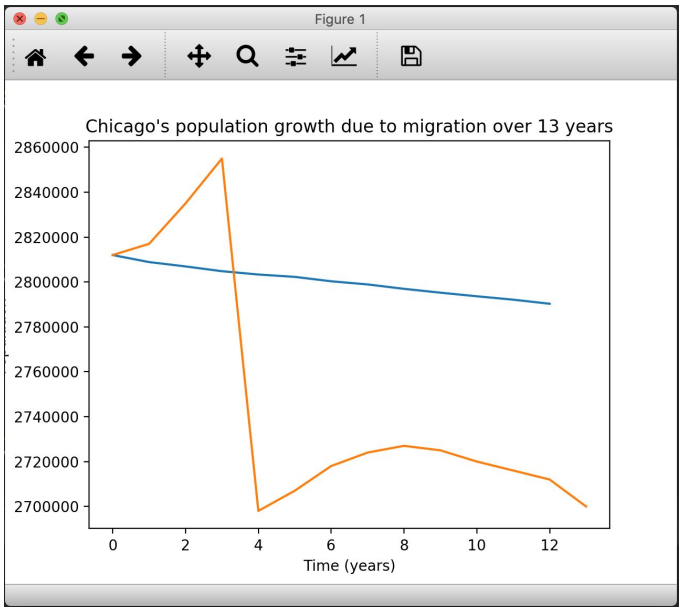
Considering we are using four factors to calculate the population, this is a very respectable result to have as there are several other factors that we haven't included that can also have an effect.

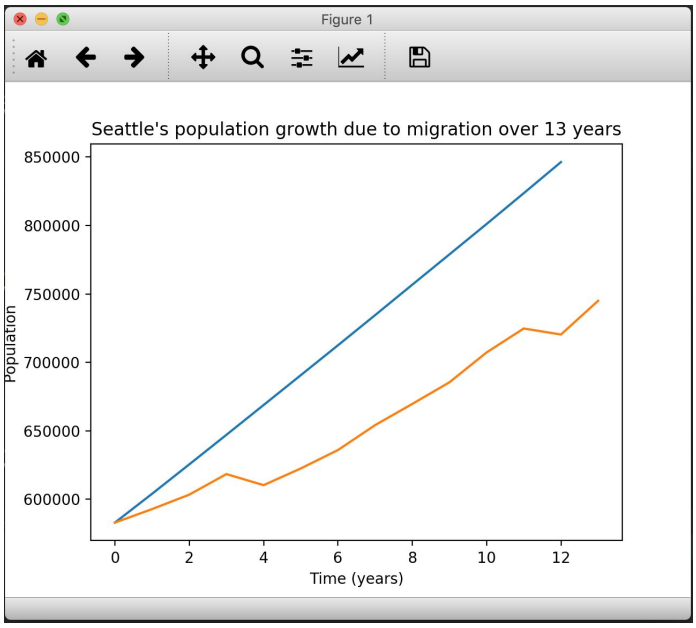
This accuracy is also seen in Chicago's model which is accurately depicting population decay as Chicago's population has been on the decline in the last 13 years.

The absolute and relative errors are 37,725.034 and 1.355 respectively. The relative error being only a 1% difference shows that our new model is a much better representation of the actual population change in a city over a period of time.

Re-Testing: Random Distribution of Jobs: Normal VS Binomial

Now that our model is much more accurate, we want to retest the previous factors we did and see if this will have a major impact on our model. Will it be better to have a normal distribution or a binomial distribution? First up is the distribution of jobs, with our older model it didn't help and actually lowered the accuracy of our model by quite a bit.

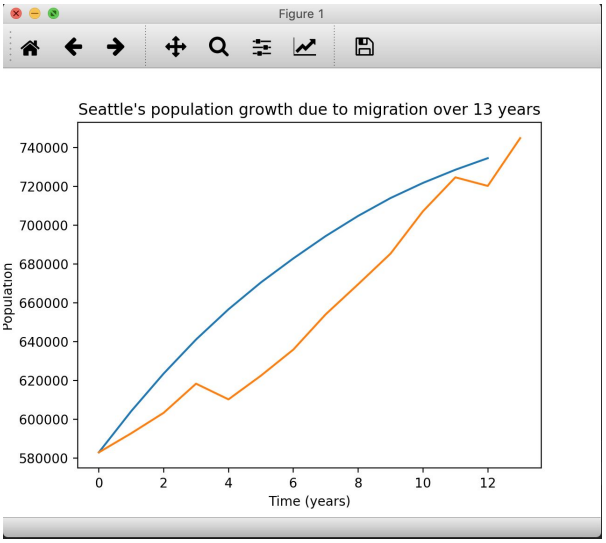


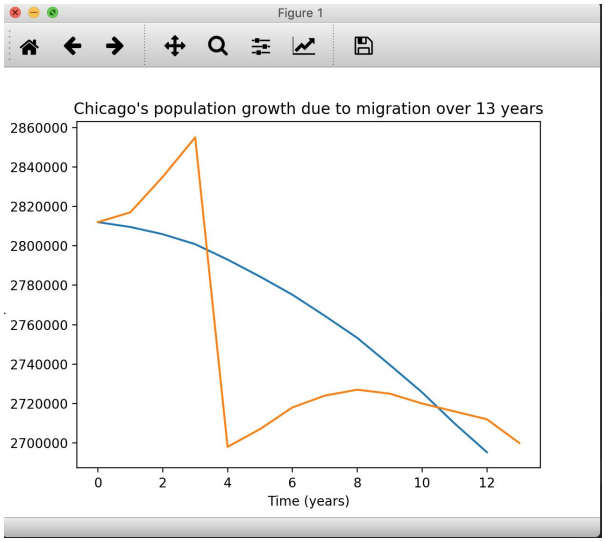


As you can see, the distribution being binomial actually also has a negative effect on the model and the relative error for Seattle in this graph is 8.6%. Compare this to the previous graph where the relative error was only 3.6%. The population of Seattle increases much more rapidly than it should.

In the Chicago graph, there are similar issues as well. Instead of this graph decreasing rapidly, it is actually decaying at a slower rate. This could be because by making the distribution binomial it is actually adding a lot more jobs than it should. This is why population growth increases much faster and population decay decreases much slower. The relative error for this graph was 2.2% compared with 1.4% for the previous graph. So we can determine that having a binomial distribution is not a good idea with our model pre-bug and post-bug.

Re-Testing:Random Distribution of Crimes: Uniform VS Binomial

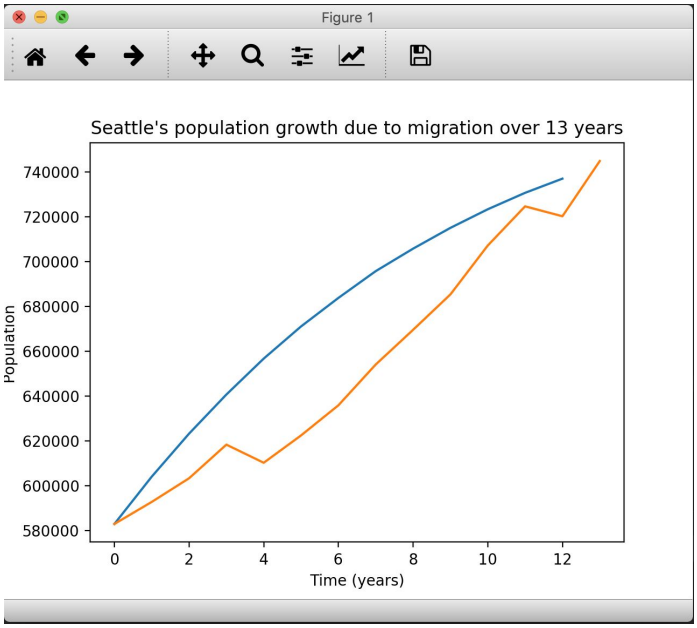


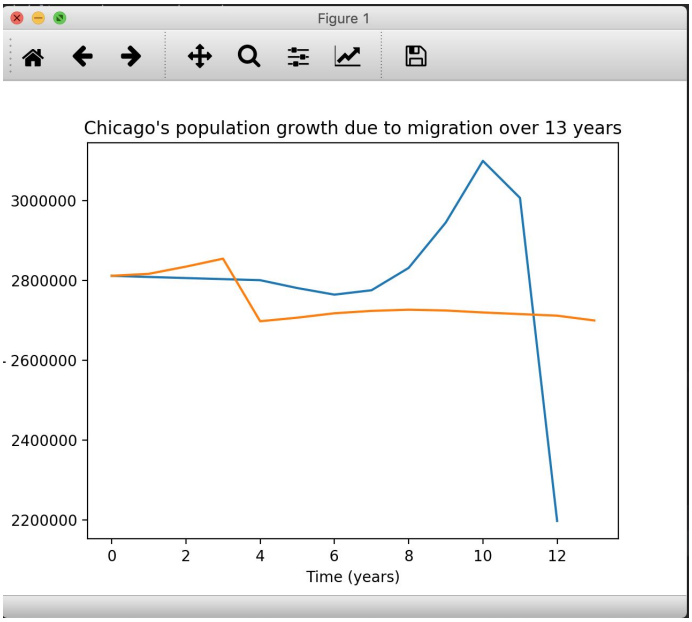


In our previous model when we changed the distribution to uniform it had a negative effect and increased the amount of error that would occur in the model. We decided to retest this and see if it will have a different effect now.

The relative error in the previous simulation for Seattle was 3.683 and for Chicago it was 1.356. When we switched to a uniform distribution this number decreased slightly to 3.5 for Seattle and 1.3 for Chicago. While it wasn't a major decrease, it was still a decrease. This indicates that our model will be more accurate if we change our distribution of crimes to be uniform instead of binomial. As a result, we will change our model to calculate crime as a uniform distribution for greater accuracy.

Re-Testing: Random Distribution of Taxes: Uniform VS Normal



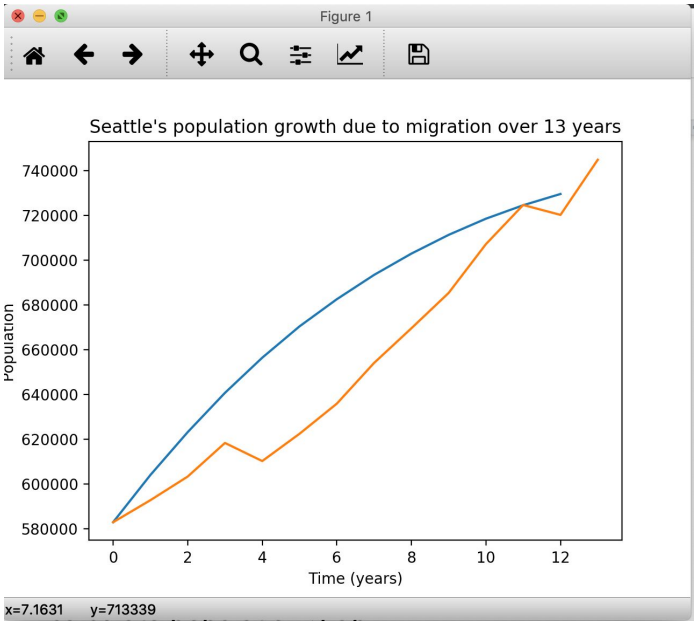


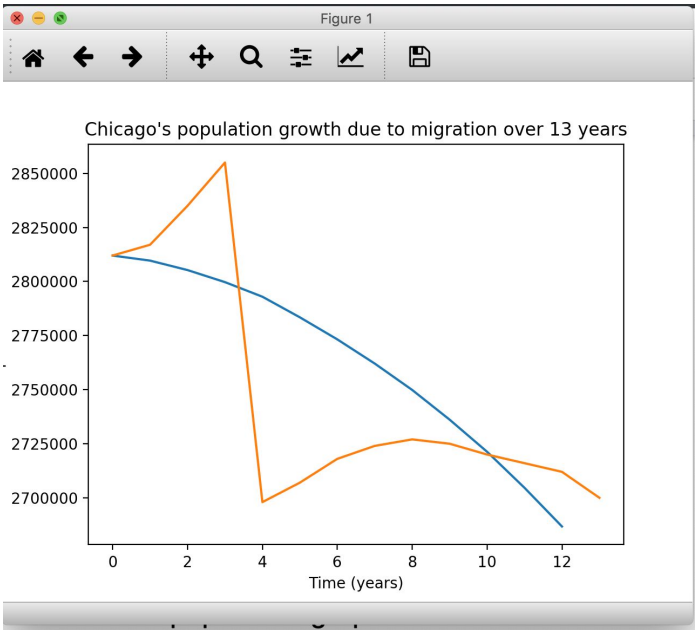
In our previous simulation, we found that changing the taxes to uniform instead of normal actually helped with our model making it more accurate. We wanted to see after we fixed our model that this would still be the case. This graph shows the normal distribution and the relative error is 3.889 for Seattle and 5.355 for Chicago. With a uniform distribution, Seattle was 3.489 and Chicago was 1.289.

Seattle's population is still pretty accurate a normal distribution and looks very close to the actual representation. However, Chicago is a different story. Chicago's population dips very heavily towards the end of the simulation and this increased the error by 4%.

Overall, this change has just made the error and model less accurate. This is why we decided it would be better for our simulation to just keep the random distribution of taxes as uniform instead of binomial.

Re-Testing: Random Distribution of Rent: Uniform VS Normal





In our previous simulation we found that random distribution of rent performed better when it was a normal distribution vs a uniform distribution. Will this still be the case after or fixed model? That is what we wish to answer.

With a normal distribution, the relative error for Seattle was 3.489 and Chicago's was 1.289. Once this was changed to a uniform distribution the errors for Seattle and Chicago became 3.578 and 1.189 respectively.

This change was pretty minor and didn't really change much in the grand scheme of things. Rent being uniform or normal has little effect in our model's population. As a result, we decided it would be better for our simulation to just keep the random distribution of taxes as uniform instead of normal.

Conclusion

Throughout the design and implementation process of our project, we faced a lot of ups and downs due to the fact that we lacked the experience in building heavy data-driven simulations (especially for calculating migration). Moreover, we encountered several miscalculations in while building the algorithm using merely four factors to determine total amount of people migrating to the city (total of jobs, average of rent costs, tax rate, and total crimes).

In terms of making the values for the four factors based on the initial data that we have, we see that the random number distribution affects our population simulation greatly. Take an example of the part where we see the difference of the population plot by using different random number distribution for total jobs per year (normal vs binomial). The population plots for Seattle and Chicago have bigger error values in comparison to the actual population values by using binomial distribution rather than by using normal distribution for calculating the total job available per year. Not just for jobs, there are more invalid simulation because we use different distribution for other factors as well.

We also learned that there are significantly more factors to determine migration to a city and each of those factors have very high variance and as such are really hard to predict. Our calculation is controlled in a such a way that it shows the general trend of how the population is changing, but in order to predict future populations we need more data.

In similar to predicting the value of a company's stock on the stock exchange, the population of migrants to or from a city differs greatly from year to year. As seen in one of our graphs, the population of chicago drops by 200,000 people. After doing a bit of research as to why this might have happened, we learned that it was primarily due to the 2008 recession.

Source: https://www.chicagotribune.com/news/ct-met-2010-census-20110215-story.html