Analysis of determinants that influence urban migration

Dheeraj Reddy Pailla, Rajiv Wadhwa, Riya Bakhtiani
Georgia Institute of Technology, Atlanta, Georgia

Project Repository: https://github.gatech.edu/dpailla3/ModSimProject

Abstract

Infrastructure Modelling is the backbone of any development project. Commonly analyzed key performance indicators towards the success of a new project are parameters such as population growth, economic growth, etc. of the region. There are various factors that affect these parameters, hence the success, of a city development project. We propose an approach to model the growth in population of a medium sized US city using a combination of regression and cellular automata. We use this approach to simulate the impact of these infrastructure elements on a city's population growth during real scenarios such as a pandemic, economic crisis, etc.
Keywords: migration, regression, cellular automata, *socio-economic factors*

1. Description of the system

In this project, we simulate the migration of the human population from rural or semi-urban areas to a city which is our object of study. All the regions of this system under investigation are considered to be graded on various socio-economic factors that drive migration. Based on these properties, the population of every region changes with each simulation. This model thus analyses the rate of population growth by changing the values of these determinants with each simulation.

In the real world, a city council could use this model for budget resolution and informed authorization of public improvements. Population growth generates new jobs, new income, increased tax revenue, and higher property values. [10] A city council could thus benefit from this simulator and invest in the right part of the infrastructure in the city.

Another application of this simulation is to determine the feasibility of investments in large-scale urban development. The builders and developers of any new city project invest based on the prediction of the timeline of the occupancy of their project, and that determines the success and failure of that project. Our study of the rate of growth of population or rate of migration based on different factors could prove beneficial to them in order to make various infrastructure modeling decisions.

2. Literature review

2.1 Factors

- *Healthcare*

  The primary indicator for determining whether or not people are able to “be" or “do” desirable things in life is the Human Development Index (HDI), by the United Nations Development Program (UNDP), which combines the performances of countries based on health, education, and economics. Migration due to lack of fundamental human rights such as health care is prominent as location matters when it comes to health.[13] To measure this, we look at the cost of care and life expectancy rates for different regions. A more holistic approach to quantify this could be an amalgamation of several other factors as well. A region can prioritize its citizen’s health care by promoting wellness, keeping medical services affordable & accessible, or just maintaining cleanliness, even. The most pressing issue in the current scenario is the COVID-19 pandemic, implying that how well a city handles such a situation can also be a factor to rate it based on healthcare services as it largely impacts the physical and mental wellness of its residents.

- *Infrastructure & transportation :-*

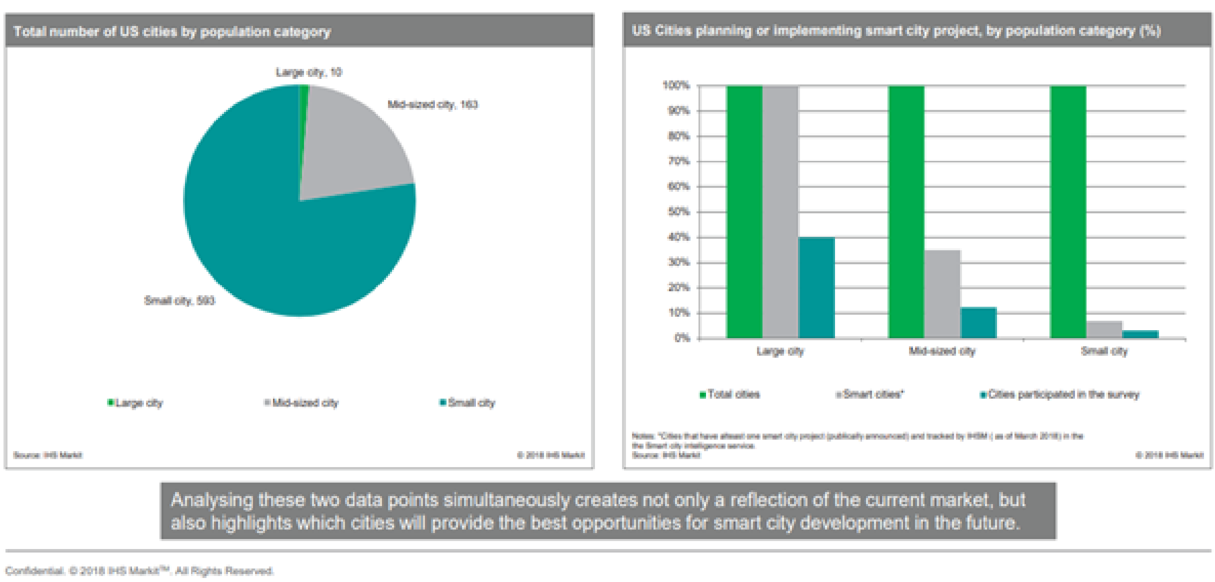
  For years, it has been evident that the increase in connectivity has always witnessed a rise in the growth of the region. The paper [1] has found a quantitative effect of infrastructure development on growth as well as income inequality. And in doing so, they have provided a comprehensive evaluation of that impact. The data set which they used covers 100 countries. It is with respect to the quantity indicator of each infrastructure network and making a linear regression framework model of each infrastructure impact. To evaluate the impact, they used GMM econometric methods to minimize their fluctuations in the data set. They found positive and significant correlations of three types of infrastructure assets – telecommunications, transport and power. It signifies the contribution of each sector from the 3 shortlisted infrastructure networks. The data such as telephone density change, power consumption data were taken for the study. Lastly, they identified that conceptually, along with growth in infrastructure, the underprivileged individuals and underdeveloped areas obtain better connectivity. This indeed opens up opportunities for jobs or economic growth and reduces economic inequality.

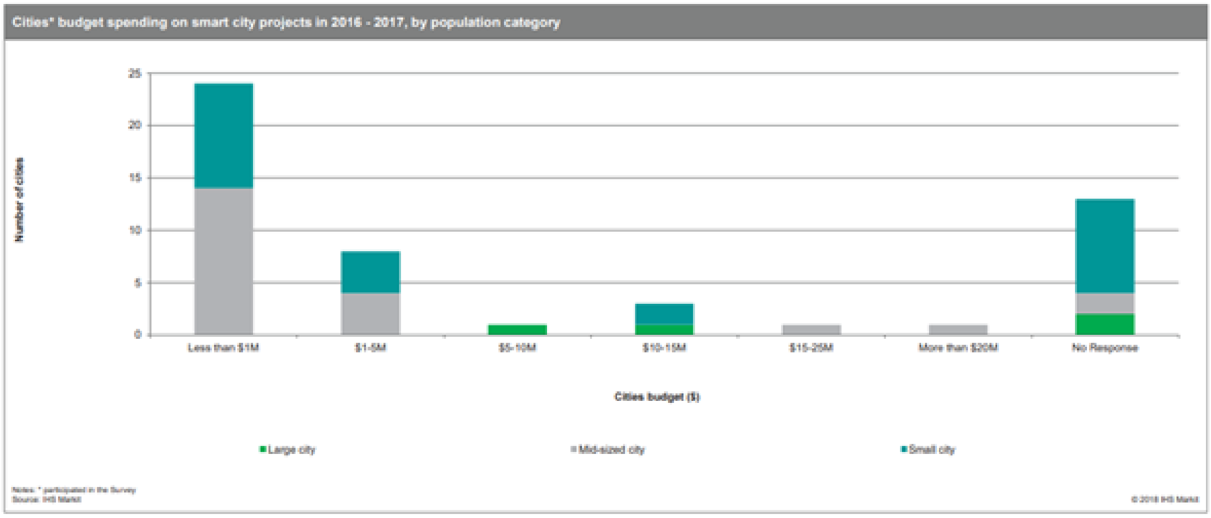
- *Taxation*

  It is widely believed that an increase in taxes will eventually lead to an exodus of people from a city. This belief has been strengthened in the last few years due to the large outflow of people from the state of California to other parts of the US with lower taxes, such as Texas or Virginia. However, research [9] conducted by the Political Economy Research Institute at the University of Massachusetts has shown that people are likely to remain in regions of high taxation because they value the public services financed by taxes. Additionally, the cost of moving is often fairly high with not much to gain from it. The main reasons for moving are employment, family related matters and education. This suggests that migration can be heavily influenced by improvements in infrastructure, even if it is funded by higher taxes.

- *Smart City*

Smart city development projects are a subset of infrastructure projects but since the sector is not flourishing and it is growing, the need to study its impact becomes of higher-order priority. This will help to make decisions on future investment in the right kind of projects. Here the indicator we have identified to study the impact on the economy due to a smart city is by growth in spending on smart city projects. The above literature forms our basis of references, which shows the distribution of smart city development projects trends concerning the population of the city. In their survey, they have found that all large cities at least have 1 smart city project and a higher population range in mid-cities shows great potential to invest in smart city projects. This shows that regions with higher populations favour having smart city projects. Since it is proved in the paper [2] that budget remains the major barrier towards implementing smart city projects, we consider the indicator of smart city growth to be an increase in the amount of spending on smart city projects by a city. This study also shows the trend that mid-size cities are willing to invest more in smart city projects (Fig 1)





- *Education*

If simplified to a certain extent, we can say that people migrate for security concerns, livelihood opportunities, and economic incentives. They base this decision on a comparison of expected lifetime earnings in the current region and in the alternative region. Education and skill acquisition are closely related to this ultimate goal. However, the literature survey [6] points to the fact that education is a primary factor only in certain groups of the population. For the rest, it is a secondary driver or a socio-economic factor that is considered as a family investment in children's education. These groups can be identified by dividing the population on the basis of age and gender.

Age: Migration is mainly often observed among the working-age population and education is not a predominant force in that case. Better schooling is a major issue for mainly two broad groups: youth aged 15-25 and people with children. For students, education is their chance to improve their economic situation and for young parents, this is an implicit factor towards building a better life.

Gender: A detail that could potentially add weight to education as a driver for migration is more freedom to seek better and more opportunities for skill acquisition. Here education acts as a push factor instead of a pull factor.

- *Immigration leading to job creation/economic improvement*

  The National Foundation for American Policy in 2016 released a study titled "Immigrants and billion dollar startups" in which it was shown that more than half (55%) of America's billion-dollar startups had at least one immigrant as part of their team of founders. While the topic of immigration is quite contentious politically, from an economic standpoint the research is quite clear - immigration leads to more jobs and economic growth. It is a misconception that immigrants are taking up jobs that would have otherwise gone to Americans. In fact, the jobs taken up by immigrants (whether it is high or low skilled) make it possible for more jobs to be created which are taken up by Americans - for example in the service sector where Americans enjoy the advantage of soft skills over immigrants.

3. SUI and Data Description

Our system consists of a region (could be scaled from city to state or nation). We aim to understand the behaviour of migration with respect to various socio-economic factors of that city. We have considered that these factors of the city influence the migration. For example, the GDP, access to healthcare & education and so on. Along with that, we have also considered the combined behaviour of the entire population of the city. Thus the influence will also be based on each other's decision of migration. We have tried to combine both systems in order to get the cumulative effect of both influences.

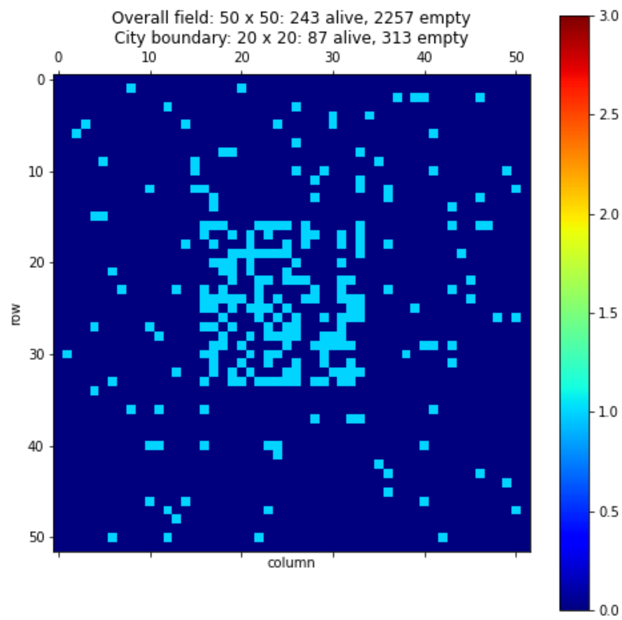
In the first part, we have incorporated the effect of various factors related to a region. We have implemented a regression model that is trained with all factors of a particular city. Using this, based on the input data, we can predict the change in population of the city. We feed data from www.data-z.org to the regression model respective to the factors mentioned in the literature survey. Based on the experiments and data analysis, we have incorporated additional features that are not covered in the literature survey.

The sample dataset of factors are illustrated below. Based on our literature survey, we identified that these factors could be crucial regarding the decision taken to migrate or not. This is a real data set of a US city collected from reliable resources.

Atlanta									
Year	**Population**	**Migration into City from Different States**	**Bachelor's Degree or Higher Share of 25+ Population (Total) %**	**Health Insurance, Percent Insured in City (%)**	**Violent Crime Rate in City Per 100,000**	**Unemployment Rate**	**Home Value Median (Zillow)**	**Income Per Capita in City**	**City Real GDP Per Capita**
2010	429410	16547	44.8	80.4	1071.63	10.3	152504	30688	52315
2011	437812	16850	47.2	78.8	1432.79	9.9	134362	33117	52808
2012	449016	20024	47.4	83	1379.05	8.8	131764	35829	52783
2013	453990	16294	48.4	82.8	1223.23	7.8	147272	36257	53217
2014	461154	19140	48.9	86.1	1227.43	6.7	165460	36936	54852
2015	468303	15962	48.3	88.9	1119.62	5.7	175916	39660	56590
2016	479174	23192	50.5	89.2	1083.63	5.1	187967	40882	58195
2017	491670	20051	49.2	89.5	935.72	4.5	202235	44690	59847
2018	498183	22481	53.4	89.8	768.79	3.8	221222	48869	61608
2019	506811	19052	56.5	90.4	770.30	3.2	235574	54414	61761

In the second part, we have implemented a cellular automata like model to visualize the nature of migration. Here, we have tried to demographically represent the city as the central grid and outer regions are analogous to exterior regions of the city. To simulate the 2nd part, we have translated the output of the first part to give as an input to our 2nd part. To do the translation, we have taken help from the measure of change in growth rate that influences other regions around the city. This is like assuming that the growth in migration influences others to move to a region, which is also validated in our validation section.
As mentioned, the data in the 2nd part is translation from the 1st part. Using regression, we get a

population number which we pass from the translator. The translator checks the growth of that number with the previous year's number and gives output based on the scale of calibration we have defined.



4. Conceptual model

4.1 Structure

Each active cell in the CA model signifies a single person. We assume that this person represents a section of the population.

4.2 Initialization

The initial condition of each lattice in a 50 by 50 grid of a Cellular Automata model starts with a 20% probability of being active. The center of this grid would be a 20 by 20 region which we would describe as a city, and the remaining cells would be considered to be an abstraction of regions outside the city. This 20x20 region is filled 40% (denser since it represents a city)

We also assign a state to each active cell - 0 or 1. 1 represents that the person will migrate towards the city, while 0 represents no inclination to migrate.

4.3 Properties of the system

Majority rule Moore's neighborhood type model is our central simulation engine. At a particular time period, the quality of infrastructure of the city will be the factor that will influence the decision of a person. The effect of a factor will be determined by the Gradient Boosting regression model. That value will be translated to give an input to our CA model. Once initial values are given to each lattice based on a regression analysis framework, we run through our CA rules.

4.4 Transitions

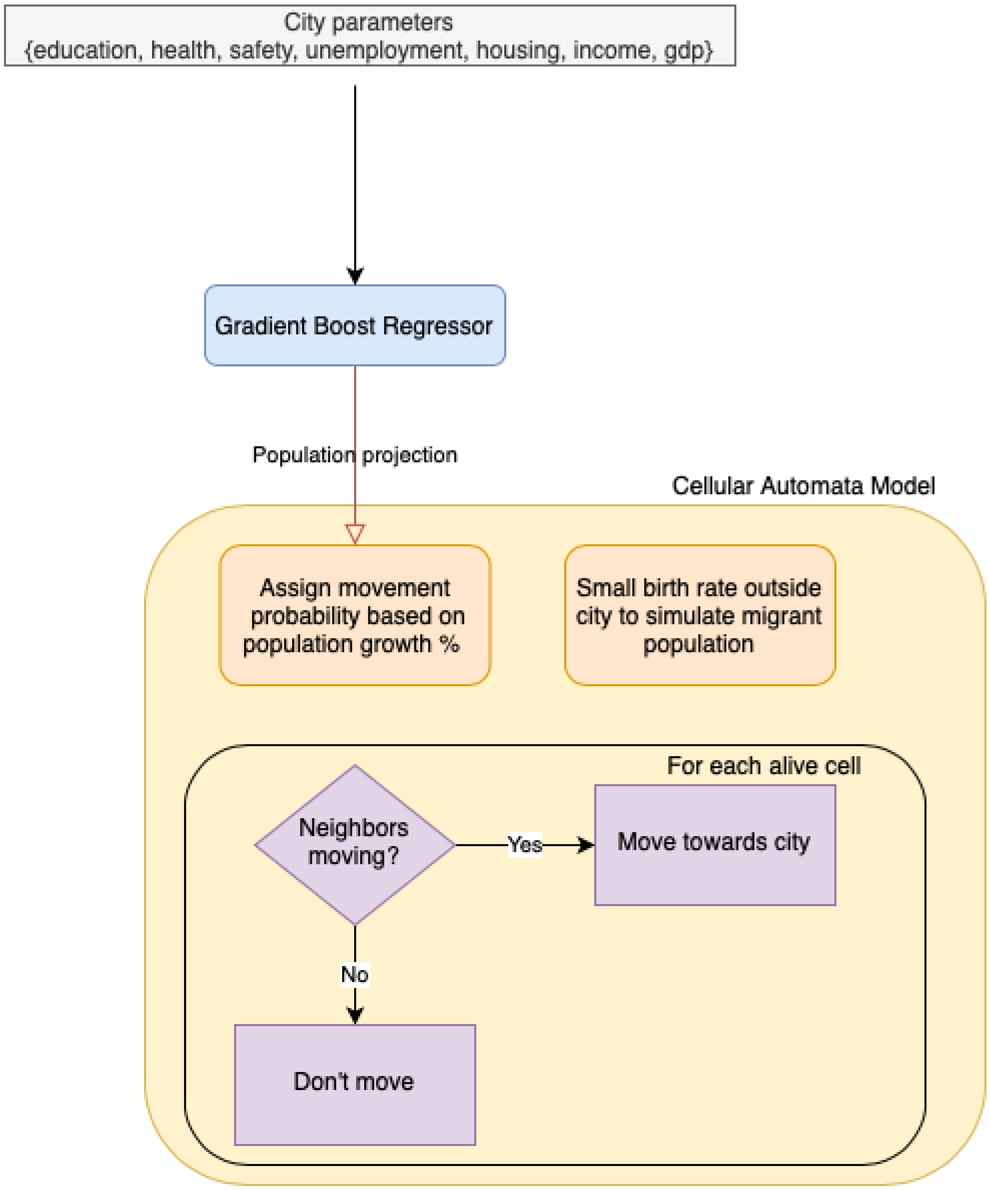
The system will go through transitions at every time step in the following ways:

- In each time step, first every active cell is assigned a state based on the population growth percentage predicted by the regression framework.
- Then, a vote is held for each active cell with the state 1 - it moves only if a majority of its active neighbours also are in the state 1.
- There are primarily 2 directions of movement - towards the city, and away from the city. Based on the output from the regression model, if growth rate is negative then active cells will move away. If it is positive, active cells will move towards the city.

4.5 Validation & Datasets

The simulation outputs will be validated using historical data like census reports obtained through the United States Census Bureau[11], HealthStats[12] & the Migration Data Portal[13]. Initial properties will be assigned based on cities for which the data already exists. The validation against known results will determine the feasibility to use this model for a new object of study.

5. Implementation



5.1 Description of Software

I. Population regression

The dataset indicates that each city has its own unique formula to optimize investment in infrastructure which would increase migration into the city. Hence, we decide to use a regression approach in order to estimate the importance of each infrastructure element for a given city. We first study 4 different regression approaches in order to determine which model fits our problem set and our pipeline - Linear regression, logistic regression, random forest regressor, gradient boosting regressor.

Based on the data we collected for the city of Atlanta, we have 10 data points to train our regression model on. We use Mean Absolute Error to determine efficacy and accuracy of each of these regression approaches. A train-test split of 80-20 was used to validate the results of each approach. The results from the Atlanta dataset are shown below -

Regression Model	MAE on test set
Linear	6924.86
Logistic	10562.0
Random Forest	12114.21
Gradient Boosting	3300.71

Based on these results, we picked the gradient boosting regressor and used this approach for regression across all our simulations.

II. Migration Simulator

The core simulator simulates the movement of population beginning from an initial world, W0. By default, the simulator stops when the world reaches a steady-state configuration. It returns all time steps if inplace is False which is also the default.

The building blocks of the simulation that results in a display of an interactive widget that allows you to look at each time step:

- create_world(n, q): Returns a new "world" of size (n+2)-by-(n+2), where the probability that any cell initially contains a living tree is q. The extra padding on the boundaries is initialized to "empty" (0 values), and is there to simplify some of the subsequent array indexing and slicing.
- show_world(W): Creates a "heat map" of the current world. The assumed states of each cell are "dead" (0) and "alive" (1).

At each iteration, we check neighbours and if the majority want to move then the person in question moves. Here, we count only votes of alive neighbours, not dead ones. If there are no alive neighbours, do whatever the current coordinate wants to do according to the fed input state. We use 4 move functions that result in movement in one direction, and then the other direction if the preferred field is empty. A cell can either move along the x or y axis - move along x axis if choice is 0, else along y axis. The cell will move nowhere if both directions towards the center are blocked.

We configure our model in such a way that the presence of more than max_neighbours (7 in our model) neighbours cause the death of the cell in question. We then introduce birth on the outer rim with a probability distribution. The model does not generate random new people inside the city. The growth rate is derived from our regression model described in the previous step. We determine likelihood to move based on city parameters:

```
W_move_chance = np.zeros(W_new.shape)
  for idx in range(len(alive_idx_x)):
     W_move_chance[alive_idx_x[idx], alive_idx_y[idx]] = np.random.choice(move_chance)
```

The model then loops through all alive cells to push them towards the center. To determine movement probability, we use the projected growth rate derived from our regression model to form a probability distribution. A high growth rate would imply a more favourable probability to move. Then, the active cell makes a decision to move towards the city only if a majority of its neighbouring active cells are also in favour of moving.

5.2 Code Performance

This simulator works with datasets for various cities ranging over a period of 10 years. The program, written in Python, takes up to a minute to execute all time steps on a personal computer. This kind of performance could be achieved because the model implements discrete-event simulation where each iteration corresponds to a certain number of days. This number varies with each city as each city has its own rate of growth of population. A continuous time framework would have to look at the state of every cell of the lattice and the value of scores at each point in time to determine the next state which would have reduced the time efficiency of the software.

The performance can be improved further by adding more input data to the regression model as the software is scalable and would generate more accurate results if fed with more input data. Code is inefficient since we relied on for loops rather than bitwise operators due to complexity of CA transition rules

6. Assumptions

In this project, we have modelled real world cities with highly complicated behaviour exhibited by its citizens. To simplify this system, we have made the following assumptions:

- We are considering migration related to a region which is similar to a city. This holds the assumption that based on the factors assigned to the city or region, a person will decide whether to move to the region or not.
- We are also considering the independence of any other factor other than the factors in our SUI.
- Surrounding rural area used as an abstraction of migrants coming from around the world to a US city.
- As with any computer simulation approach, we're relying on past data to make predictions for the future which may or may not be reliable since the model is a simplification of reality. Actual factors for migration may depend on a range of personal and other issues that we were unable to factor into our model.
- Assuming this for growing cities such as Atlanta and can only be extended to cities with a similar economic structure. Model does not account for migration out of cities back into rural areas since this is assumed to be low.
- Predictions are made only for the near future (within 1-2 years). May not hold up well for longer term
- No birth/death rates are considered. This is assumed to be stable and cancel each other out in the short term (1-2 years). In general, fertility rates have declined for most US cities so we think this is a fair assumption to make.
- Population density within cities is ignored because it isn't relevant to our discussion on infrastructure. Some elements that have been left out of our model are hard to quantify and hence

haven't been included in our model

- Doesn't consider negative factors affecting growth of population
- Can't validate intermediate time steps of CA model due to time and data limitations
- Our CA model is validated for population growth through the 2018->2019 population growth, but not for population decline (which is seen in some configurations). We think this is a fair assumption to make since it is tough to find recent data for a city that has both population growth and decline. The validity of our model depends on recency and relevancy of provided data, and the most recent population decline is several decades ago.

7. Data Analysis

After observing the collected data, we believe that there really isn't one single solution to every city's infrastructure issues due to varying levels of existing infrastructure and population inflow. Each has its own unique formula to optimize investment in infrastructure which would increase migration into the city. Hence, we decide to use a regression approach in order to estimate the importance of each infrastructure element for a given city. As an input to this regression model, we have gathered data from reliable resources such as the U.S. Census Bureau, Federal Bureau of Investigation, U.S. Bureau of Labor Statistics, Zillow and Bureau of Economic Analysis. The data spans over 10 years with an increment of one year for each city ranging from 2010-2019. We have considered 4 factors that are considered as the indicators of quality of life. To quantify each factor, we have collected various statistics from the resources mentioned as described below.

7.1 Economy

The values for this feature were collected from the sources U.S. Bureau of Labor Statistics[19] and Bureau of Economic Analysis[22]. The quantifiers are 'Unemployment Rate' of the city for each year in the considered time span and 'City Real GDP Per Capita'. The unemployment rate is the number of unemployed people as a percentage of the civilian labor force, which is the total number of employed and unemployed people in the civilian non-institutional population (the number of people aged 16 and older who are not incarcerated or on active duty in the military). City Real GDP Per Capita is the Gross Domestic Product of a given area divided by the resident population of the area. [25]

The data suggests that as these factors improve i.e. Unemployment rate reduces and the GDP per capita increases, we have a significant rise in the population indicating a positive coefficient in the results of the regression model.

7.2 Health care

The U.S. Census Bureau[12] website was used to get the percentage of people who have health insurance in the city corresponding to the feature 'Health Insurance, Percent Insured in City (%)'. From the data set it can be observed that this value directly indicates the degree of health care access that the citizens have. The Current Population Survey (CPS), jointly sponsored by BLS and the Census Bureau, is a monthly sample survey of about 60,000 households designed specifically to produce the current monthly employment and unemployment data and the annual data on income and poverty for the nation. [19]

7.3 Education

U.S. Census Bureau[16] releases the percentage of the population that is 25 years or older and have received a bachelor's degree or higher. This indicates the access to education facilities in a city to a level that can ensure higher income. The years given refer to the year in which the data was collected, which is also the year for which the data pertains. [25]

7.4 Safety

Federal Bureau of Investigation[18] releases the 'Violent Crime Rate in City Per 100,000'. The FBI's Uniform Crime Reporting (UCR) Program defines violent crimes as "offenses that involve force or threat of force. Violent crime is composed of four offenses: murder and nonnegligent manslaughter, rape, robbery, and aggravated assault." The violent crime rate in the city was calculated by dividing the number of violent crimes by the city's population and then multiplying 100,000. [25]

8. Verification & Validation

For verification of our model, we surveyed around systems of town planning data, and the conditions that are considered at the time of migration. The study suggests that infrastructure factors are considered as a priority to predict the nature of migration[24]. Another verification was done by a study of private developers making township projects. In order to predict the growth in occupancy of their project, the town planners give priority in investing in amenities that we have taken as our factors for a township to attract occupancy[26]. Lastly, the probability based on population growth was verified by results of psychological study mentioned in references, where we found that community engagement is also a factor which influences a person and gets influenced by the decisions of others around them[23].

The validation of our model consists of 2 parts. In our qualitative validation, we match population projections with current factor settings to population projections by city council or some other source like data of newly built township.

We validate our regression approach by using a test set consisting of data points from 2 years selected randomly from the input data, and a validation set consisting of the year 2019. For both cities that we tested - Atlanta and NYC, the Mean Average Error (MAE) on the test set was less than 5% and it is less than 2% for the validation set.

We use our Cellular Automata model to simulate growth in population over time. We define one time step as the total number of time steps taken for our model to arrive at 2019's population (when starting from 2018's configuration) divided by 12 months. This allows us to establish a baseline for the evolution of time in our system using which we can analyse the speed of population growth/decline in various situations.

9. Numerical Experiments

9.1 Baseline

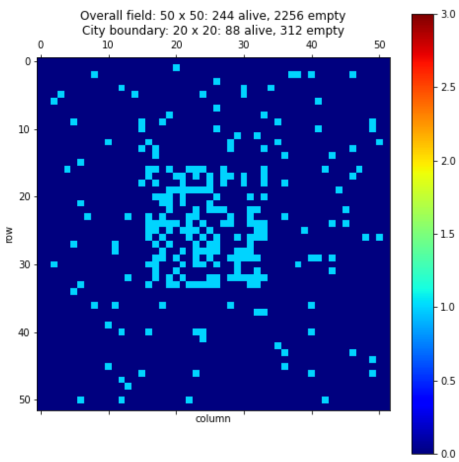
The dataset used for this project spans over a timeline of 2010-2019. Our model does not use 2019 data to learn the importance of each parameter. We use 2018 population numbers as a baseline for our cellular automata model and run our model to validate whether 2019 population number is achieved. Once we validate this, we modify the parameters of 2019 to simulate various situations (such as a pandemic, recession, increased education investment, etc).

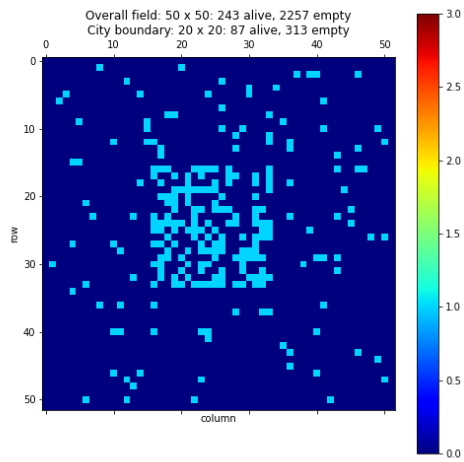
The data from the row corresponding to 2018 is fed to our cellular automata model which is run to achieve data for 2019. We then compare these results with the real numbers that are already a part of the dataset for validating our CA model for each city. In this process, we determine the number of steps required to reach the 2019 population from the 2018 setting. We call these number of steps as the baseline for a particular city. If population increases in the experiment situations in the same number of time steps (baseline for the city), it is considered a successful configuration.

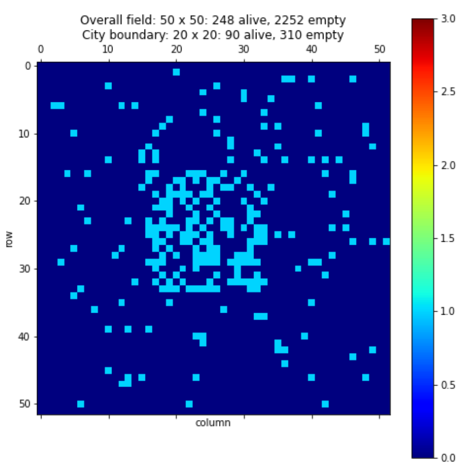
I. Atlanta

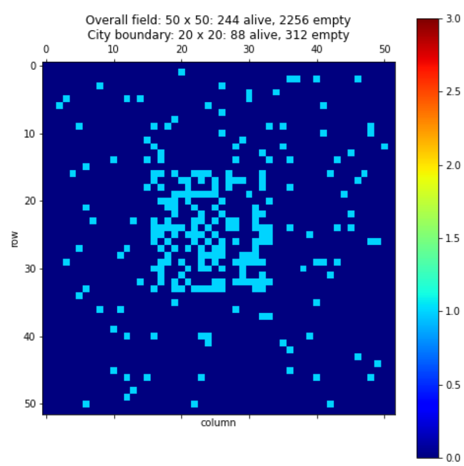
Each alive cell in the CA model for this city represents 6000 people. According to the dataset used, the population in 2018 is 498183 and that of 2019 is 506811 which indicates a growth rate of 0.17%. Our model predicts a population of approximately 507k which indicates a deviation of <1% from the actual results.

For this city, we find baseline = 14 as it takes 14 time steps in the CA model to reach the 2019 population from 2018. Illustrated below are snapshots after iterations 0, 5, 9 and 14 in sequence.







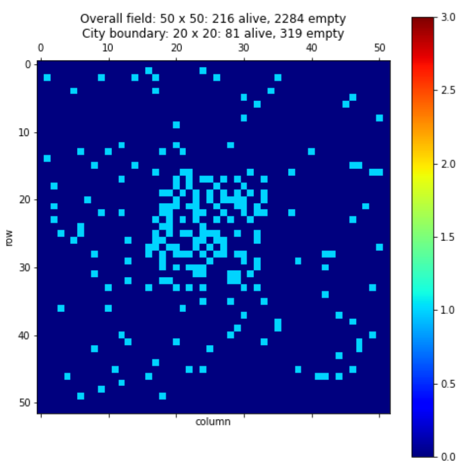


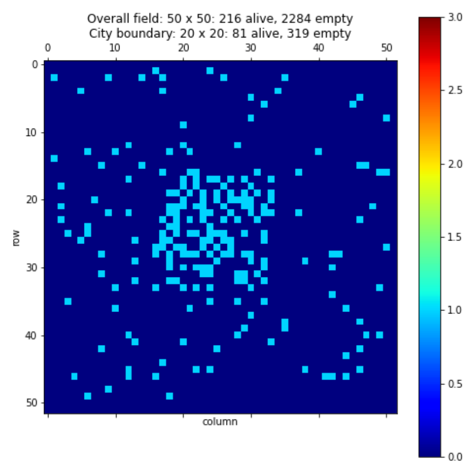
II. New York City

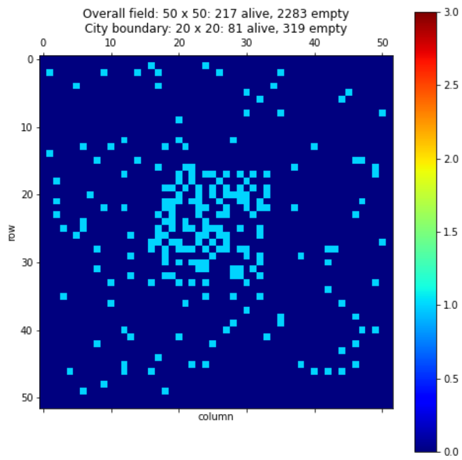
Each alive cell in the CA model for this city represents 60,000 people. According to the dataset used, the population in 2018 is 8390081 and that of 2019 is 8.34 million. Our model predicts a population of approximately 8.44 million which indicates a deviation of <2% from the actual results.

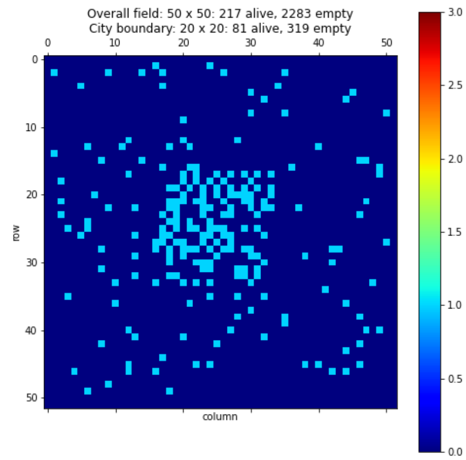
For this city, the population remains more or less stagnant, which implies that the CA model doesn't change as required so it's impossible to set a baseline. Illustrated below are snapshots after iterations 0, 3, 5 and 7 in sequence.

This is a major drawback of our modeling approach - we are unable to effectively model cities with a stagnant/declining population.









9.2 Experiments

We represent our input X as a combination of values in the following format -

[[Bachelor's Degree or Higher Share of 25+ Population, Percent Insured in City, Violent Crime Rate in City Per 100,000, Unemployment Rate, Home Value Median (Zillow), Income Per Capita in City, City Real GDP Per Capita]]

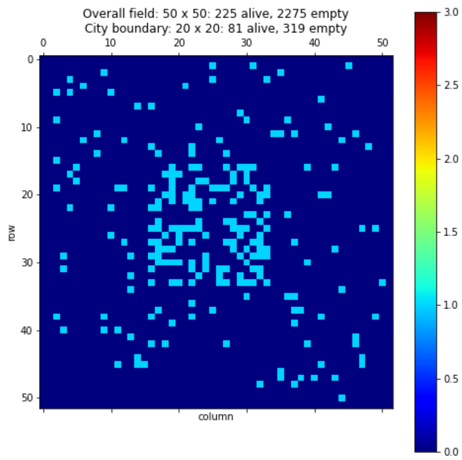
1. Recession / An economic crisis

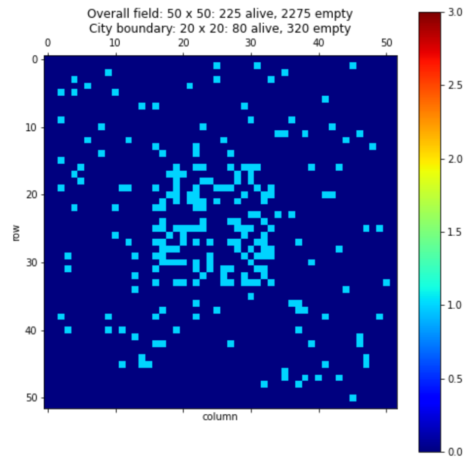
   We simulate this situation by reducing the GDP per capita and increasing the score that corresponds to the unemployment rate of the city.

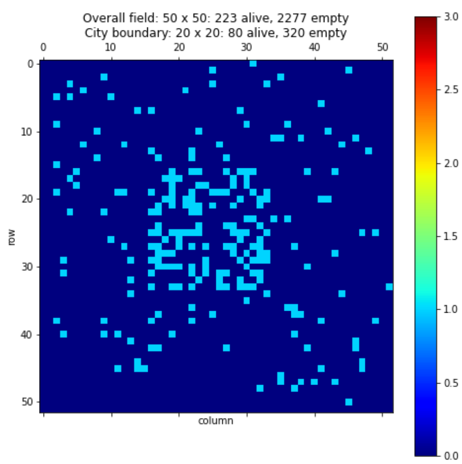
   I. Atlanta

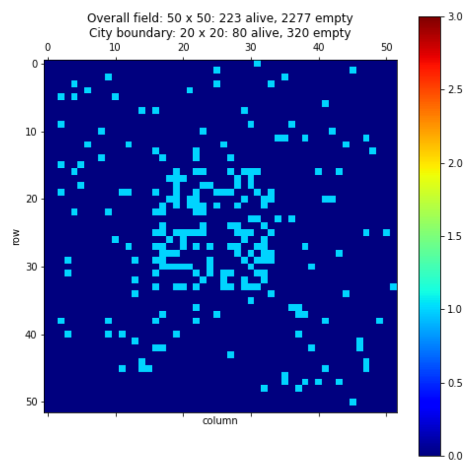
   X_economic = [[56.5, 90.4, 770.30, 6.2, 235574, 44414, 51761]] => 475690.64524543

   This suggests that the population declines in 7 steps. Illustrated below are snapshots after iterations 0, 3, 5 and 7 in sequence.



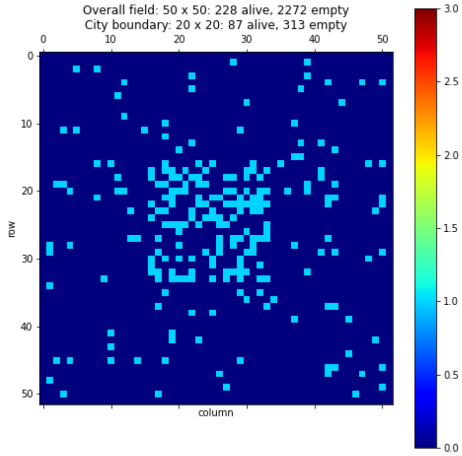


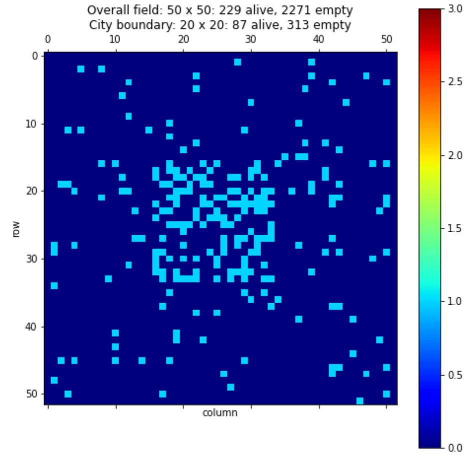




II. New York City

NYC_X_economic = [[39.2, 93.1, 570.72, 6.7, 478920, 33046, 71903]] => 8312822.03692795 - This suggests that the population declines in 4 steps. Illustrated below are snapshots after iterations 0 and 4 in sequence.





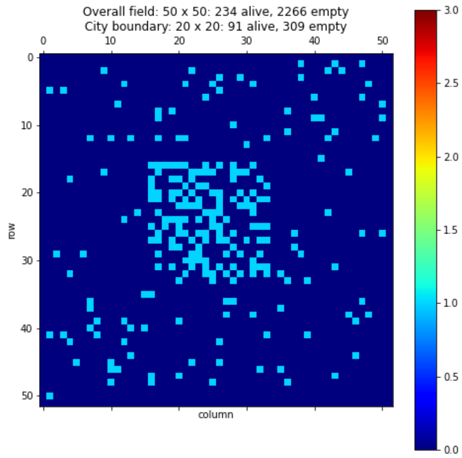
2. A Pandemic

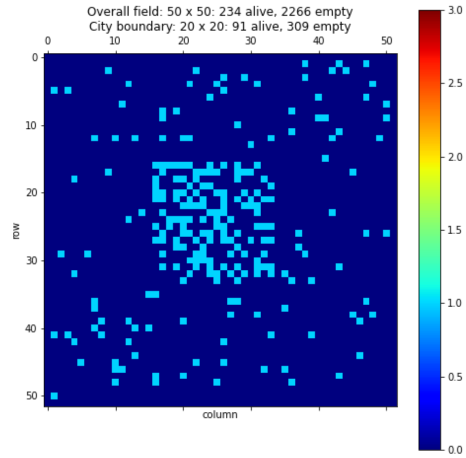
One of the biggest concerns in the world today is the capability of a government to deal with a pandemic efficiently. Whether a system is robust enough to save its citizens from major losses to health and economy has become a crucial question today. A pandemic can completely shatter a well established economy and put a major strain on the healthcare system. We simulate this situation by drastically decreasing the score that corresponds to health care of the city.

I. Atlanta

X_pandemic = [[56.5, 70, 770.30, 3.2, 235574, 54414, 61761]] => 498291.63081152

This suggests that the population remains stagnant, i.e. the CA model doesn't change because growth rate is too low.

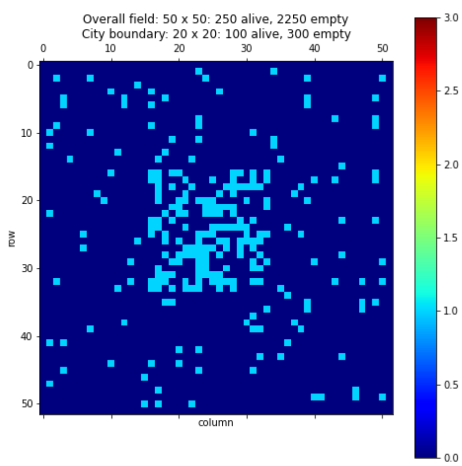


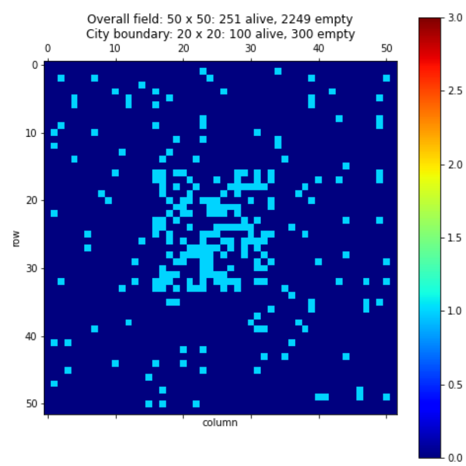


II. New York City

NYC_X_pandemic = [[39.2, 70, 570.72, 3.7, 478920, 43046, 81903]] => 8415092.48769164

This suggests that the population remains approximately stagnant, i.e. the CA model changes insignificantly as the negative growth rate is very low.



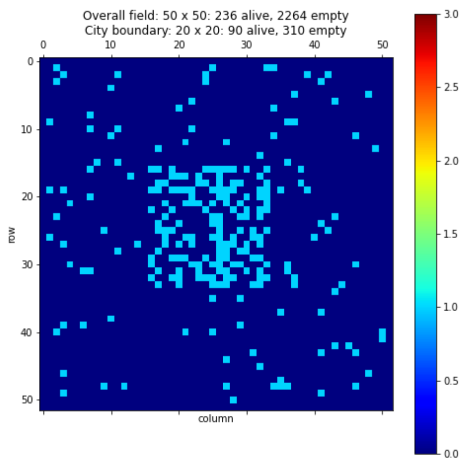


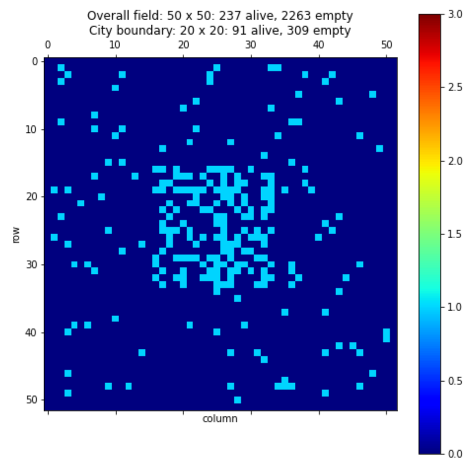
3. Investment in Education

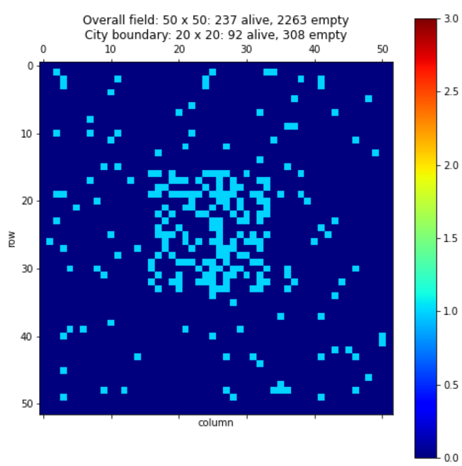
We simulate this situation by increasing the score that corresponds to the education of the city i.e. percentage share of population of age 25+ that have a Bachelor's Degree or Higher.

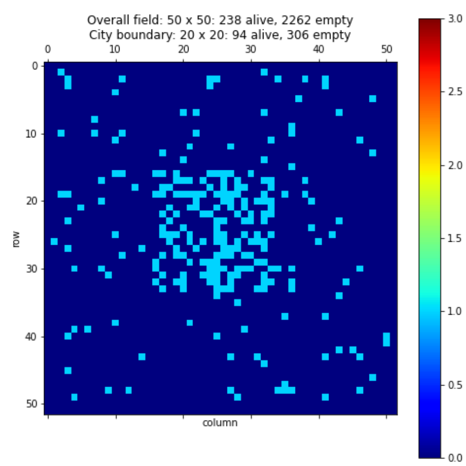
I. Atlanta

X_education = [[80.5, 90.4, 770.30, 3.2, 235574, 54414, 61761]] => 506810.01628531
This suggests that the population takes 7-8 steps to reach the desired outcome. Illustrated below are snapshots after iterations 0, 3, 5 and 8 in sequence.









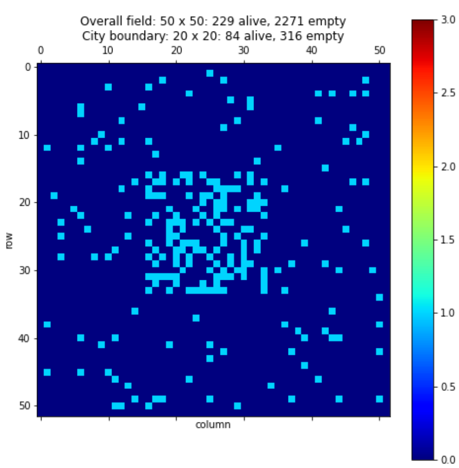
II. New York City

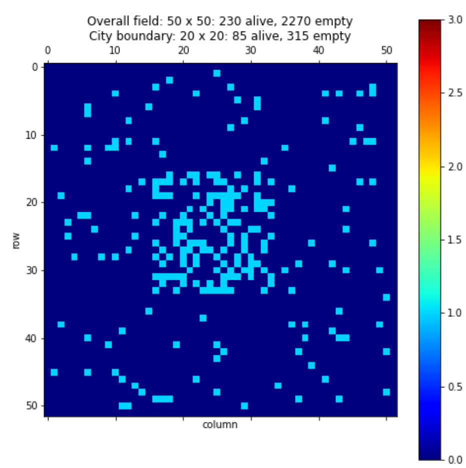
NYC_X_education = [[59.2, 93.1, 570.72, 3.7, 478920, 43046, 81903]] =>
8336818.51233633
The population remains stagnant.

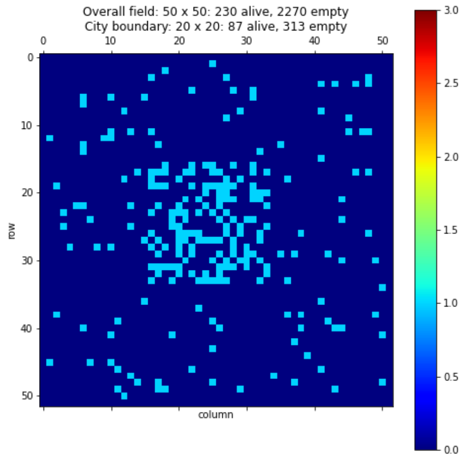
4. Investment in Safety
We simulate this situation by assuming an increase in policing and hence decrease in the number of Violent Crime Rates in City Per 100,000.

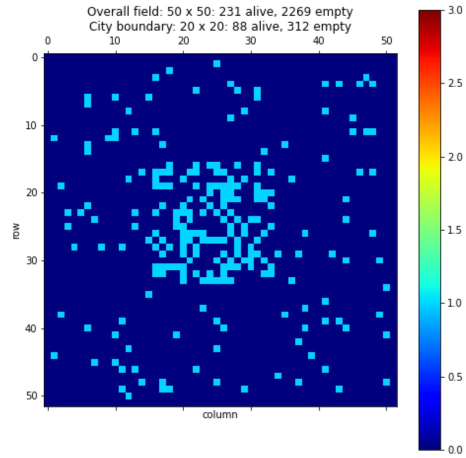
I. Atlanta

X_safety = [[56.5, 90.4, 500, 3.2, 235574, 54414, 61761]] => 505548.57804642
This suggests that the population takes 7-8 steps for the CA model to reach the desired outcome which implies that this investment has a positive impact. Illustrated below are snapshots after iterations 0, 3, 6 and 8 in sequence.









II. New York City

NYC_X_safety = [[39.2, 93.1, 370.72, 3.7, 478920, 43046, 81903]] =>
8337071.51139307
The population remains stagnant.

10. Challenges

Challenges made us revise our way of thinking. This led to some changes in orientation of our model from its initial condition to an extent that we need to select factors in order to validate it in some way. Validation of our model is not straight forward.
Abstract factors could have been useful for our model. But we wanted our model to be validated well enough. So to not fail the validation, we needed to gather data of factors which are available for nation, state and city too.

Another major challenge we faced was to quantify the influence of factors which can fit into our regression analysis model. With the real data, many factors we initially considered gave high mean error. It was not compliant with our model. So different iterations were made with different factors to conclude which factors to consider and which regression model will be the best fit.

Another task that we overcame after substantial efforts was to link the regression model by calibration of values which are practically found and not just based on hypothesis.
Our asymptotic behavior concludes that the city is abandoned due to the dynamic influence of one another. This challenge is also a limitation of our model to overcome in order to get perfect asymptotic behaviour.

However, we are unable to effectively model cities with a stagnant/declining population. Cities such as New York (as seen above in our experiments) aren't modelled well in our simulation engine.

11. Conclusions

The gains made in n time steps can easily be demolished within half that number of time steps. In a regular non-crisis year Atlanta experiences healthy population growth, but that can be erased in less than 6 months due to an economic/medical/any other crisis. We see this in real life as well since jobs gained over the course of 12 years since the financial crisis in 2008 were all wiped out within 3-4 months of the economic crisis caused by the pandemic in early parts of 2020. Population during the pandemic has stagnated, with fertility rates falling across the US [29] and travel heavily restricted.

However, there is a flip side to this as well. Population (and consequently the economy) can be salvaged through investments in infrastructure such as education. As seen in the ATL simulation, compared to the baseline which takes around 14 time steps (a year), an investment in education may result in the same immigration/population increase within 7 time steps (~6 months), and a 2-fold increase in overall population increase within the same 1-year time period.

12. Future Direction

One of the crucial jobs in the future will be to work on the connection between the regression model and simulation model. Since this remains the limitation of our model that we cannot have flexibility in our model. If in future, we can work on a function which seamlessly connects both parts of the model, this will increase the extent of our model and it can become a full-fledged methodology towards predicting the nature of migration.

13. Division of labor

Riya - Feature selection, data gathering/validation, documentation, simulation analysis

Rajiv - Conceptual model design, data gathering/validation, documentation, simulation analysis

Dheeraj - cellular automata simulator, deriving regression model, documentation, regression analysis

References

[1]Calderón, C., & Servén, L. (2004). The effects of infrastructure development on growth and income distribution. The World Bank.

[2]Driving Forces That Stimulate The Growth Of Smart Cities- https://www.smartcity.press/factors-affecting-development-of-smart-cities/

[3] Market Report, US City Decision Maker Survey – A Collaborative project run by IHS Markit and the US Conference of Mayors, US Smart Cities - June 2018

[4]Castelli, F. (2018). Drivers of migration: why do people move?. Journal of travel medicine, 25(1), tay040.

[5]Addressing drivers of migration, including adverse effects of climate change, natural disasters and human-made crises, through protection and assistance, sustainable development, poverty eradication, conflict prevention and resolution - https://refugeesmigrants.un.org/sites/default/files/final_issue_brief_2.pdf

[6]Browne, E. (2017). Evidence on education as a driver for migration.

[7]Liao, P. J., Wang, P., Wang, Y. C., & Yip, C. K. (2017). Educational choice, rural-urban migration and economic development (No. w23939). National Bureau of Economic Research.

[8] Anderson, S. Immigrants and Billion Dollar Startups. National Foundation for American Policy

[9] Thompson, J. (2011). The impact of taxes on migration in New England. *Political Economy Research Institute*, 255-284.

[10]The Benefits of Growth - Urban Land
Institute - https://uli.org/wp-content/uploads/ULI-Documents/BenefitsofGrowth.ashx_.pdf
[11]https://www.census.gov/topics/population/migration/guidance/calculating-migration-expectancy.html
[12]https://datatopics.worldbank.org/health/
[13]https://wallethub.com/edu/healthiest-cities/31072
[14]https://www.census.gov/programs-surveys/popest.html
[15]https://data.census.gov/cedsci/
[16]https://www.census.gov/
[17]https://www.census.gov/topics/health/health-insurance.html
[18]https://ucr.fbi.gov/crime-in-the-u.s
[19]https://www.bls.gov/
[20]https://www.zillow.com/research/data/
[22]https://www.bea.gov/
[23]https://www.psychologytoday.com/us/blog/after-service/201705/the-science-behind-why-people-follow-the-crowd
[24] https://www.shareyouressays.com/knowledge/10-major-causes-of-migration/111287
[25]https://www.data-z.org/glossary/detail/
[26] https://www.realync.com/how-to-increase-apartment-occupancy-through-resident-engagement
[27] The Effects Of Infrastructure Development On Growth And Income Distribution - Banco Central De Chile
[28] Cities of the 21st Century - June 2018- IHS Market - The United States Conference of Mayors -
[29] America's Falling Fertility Rate

Acknowledgments

We would like to acknowledge and thank Professor Vuduc for giving us the opportunity and inspiring us to do this cool project as part of CS 6730.