SIR & Markov Chain Disease Simulator for COVID-19

Group 2
By Devanshi Shah, Shrey Patel & Martin Graffigna
GitHub Repository: https://github.gatech.edu/dshah353/mc_sir_disease_simulator

Abstract

This project implements an SIR-Markov Chain simulator to simulate the spread of an infectious disease throughout a state. During the COVID-19 pandemic, the need to project the spread of an infectious disease became apparent, and the often default model is SIR. The SIR (susceptible-infected-recovered) model represents a closed population where individuals move from the categories over time based on an infection rate (beta) and a recovery rate (gamma). Although many public health interventions were established to stop the spread of COVID-19, movement was inevitable, and the disease ultimately spread to all corners of the world. The proposed model accounts for such movement by integrating a Markov Chain with the SIR model to simulate the transition of people from one county to another. This simulation can predict the spread over the 159 counties in Georgia over a given number of days, starting from any county. For validation purposes, the initial infected population was contained in Fulton County, the location of the first reported cases in Georgia, and the simulated spread of COVID-19 over three months reflected the actual reported spread in Georgia using the USAFacts.org database. This simulator can predict the spread of an infectious disease in any state, starting with any county, to help inform public health interventions.

Project description

Our goal was to develop a simulator to better understand the spread of infectious diseases between counties in a state. The standard SIR model assumes a homogeneous population and does not account for the time it takes for individuals to move between regions in larger territories. This aims to understand how transitions between areas can change the spread of the disease geographically instead of using a closed-SIR model. We aim to create a Markovian SIR model to simulate infectious disease spread. We can do this by creating a model of Georgia that is interconnected; Each county has a unique transition probability between connected counties and an SIR model representing the disease parameters in that county. At each time step, there can be an influx or efflux of people to and from regions, making this an open model. Our goal is to combine the SIR model with the Markov Chain to make a more representative infectious disease spread simulator.

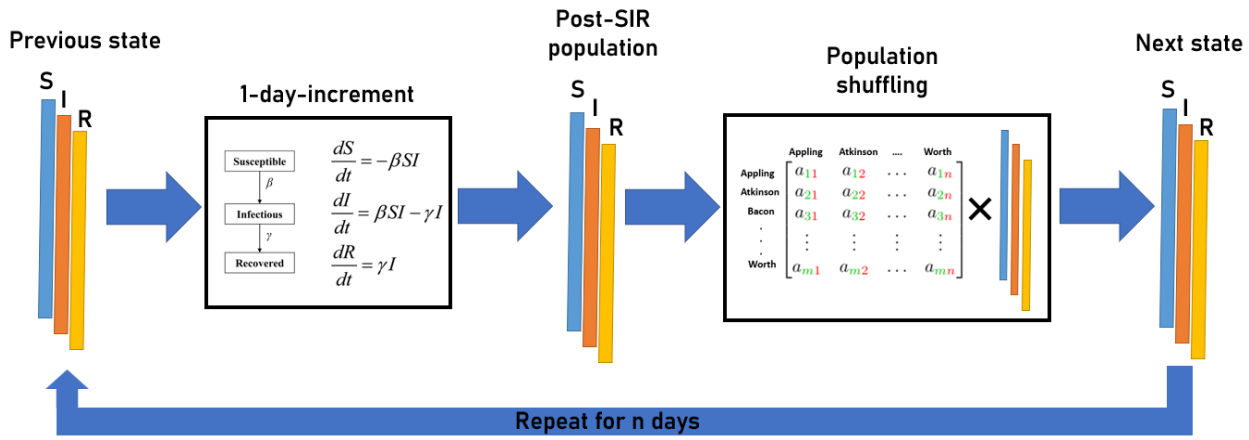

Figure 1. Pipeline of the iterative calculation using SIR + Markov Chain, the whole process repeats itself until the final time stamp is reached.

Literature Review

In Vivanco-Lira et. al, researchers developed a Markov chain probability model to scale total SIR model cases and estimate state-level distributions over time. Their findings highlighted that states with a higher mixing factor exhibited a greater risk of infection at equilibrium, emphasizing that population size alone does not determine infection risk. Instead, population migration plays a critical role in assessing regional disease spread and identifying high-risk areas.

Zakary et al. investigate the impact of travel policy implementation within a geographically connected regional model using a time-discrete SIR framework. Their approach incorporates population movement alongside birth and death dynamics, aiming to establish optimal control strategies for disease mitigation. The model represents interregional transmission by introducing an infection coefficient that links the infected populations of neighboring regions. However, this formulation primarily captures the increased exposure pressure from adjacent infected populations while neglecting the dilution of susceptibles or the influence of recovered individuals in neighboring regions. Consequently, the model implicitly assumes that the population remains in equilibrium.

Markov models have been used to study the long-term transmission dynamics of infectious diseases. Li et al. study the impact of asymptomatic carriers on infectious disease spread using Markov models. The individual interactions are characterized by continuous-time level-dependent quasi-birth-and-death processes. These processes can model interactions with different patient types (symptomatic, asymptomatic, recovered, deceased). There are also states (infection, recovery, asymptomatic, symptomatic) with transition rates defined for the change between states. The level of the Markov process represents symptomatic patients, and the phase represents asymptomatic patients. The paper defines several performance measures, including the average number of symptomatic patients, asymptomatic patients, and transition rate between states. Numerical analysis is conducted using COVID-19 data to verify the model's feasibility. Some experiments include fluctuations in infection rate and influx and efflux of patients. The generated model can be applied to other infectious diseases if they consist of asymptomatic and symptomatic patients. This study aims to better inform the spread of infectious disease between states, similar to the motivation behind our proposed project idea. This paper differs from what we want to do because we'd like to look at the transition between larger populations rather than modeling individual transmissions, but we can inform our approach using the transition models discussed.

Our proposed primary diversions from the classic SIR model are in our inclusion of changes in the states of the SIR model because of external influx to a location, which corresponds to an equivalent and opposite outflux of another location. The population changes occur with some probabilities as governed by the transition matrix. While our main challenge is having multiple SIR models connected as a Markov chain, some work has been done to address the inclusion of these factors. Discretizing the SIR model involves an extensive definition of probability distributions of the state changes and considering changes in time to change in steps as opposed to instantaneously (Yaesoubi and Cohen, 2011). With the discretization of time, the inclusion of multiple locations (multiple interconnected SIR models), and more states than the three states in the classic SIR model, our project hopes to combine previous models and create a generalized Markov SIR simulator.

Incorporating even more real-life factors is possible, with the inclusion of policies and recovery efforts like social distancing and hospital capacity. Previous work with policy intervention on COVID-19 simulations has included stochastic factors: the probability of infection, testing and quarantine policies, hospital capacity, and mortality probability (Palopoli et al., 2023). These factors make a more accurate

but complicated model. For the sake of simplicity, we may include none or some of these factors. However, it is good to know that a Markovian model can work with such real stochastic factors. For our purposes, Palopoli et al. suggest treating individuals as groups to scale their model to larger populations. Our idea of having abstract locations may be able to take inspiration from this suggestion.

Conceptual Model

Our model is developed in two stages to represent disease transmission dynamics and population movement across geographical regions.

Stage 1: Within-County Disease Transmission (SIR Model)
The first stage employs a classical SIR (Susceptible–Infected–Recovered) model to capture the transmission dynamics of the disease within individual counties. This model condenses the transmission and recovery processes into two parameters:

- **β (beta)**: the transmission rate, representing the probability of disease spread per contact between susceptible and infected individuals,

- **γ (gamma)**: the recovery rate, representing the rate at which infected individuals transition to the recovered state.

A key assumption of the basic SIR model is that the population is **homogeneously mixed**, meaning every individual has an equal probability of coming into contact with any other individual. While this assumption simplifies the dynamics and enables tractable modeling, it fails to account for real-world spatial heterogeneity and geographical boundaries, which often limit interactions.

Stage 2: Cross-County Movement (Markov Chain Model)
To incorporate the spatial structure of real populations, we extend the model to account for individual mobility across counties. The population is not modeled at the person-to-person level due to computational constraints. Instead, we adopt a **county-level granularity**, which balances model fidelity and computational efficiency.

We model inter-county movement using a **discrete-time Markov chain**, where each state corresponds to a specific county. The transition probabilities represent the likelihood that an individual will travel from one county to another within a single time step (e.g., one day). This stochastic process captures the "shuffling" of individuals across regions, thereby introducing a spatial component to the disease dynamics.

To reflect behavioral differences among individuals based on their disease status, we introduce separate **mobility adjustment matrices** for each SIR class. For example, susceptible individuals may exhibit reduced mobility due to risk aversion, while recovered individuals, presumed immune, may resume normal travel patterns. These behavioral differences are encoded through a **scaling matrix**, which is element-wise multiplied by the base transition matrix. This preserves the stochastic properties of the original matrix (i.e., each row summing to one), ensuring that population totals remain constant after movement.

Thus, at each discrete time step, the model proceeds in two phases:

1. **Intra-county transmission**, governed by the SIR dynamics, updates the state of individuals within each county.

2. **Inter-county movement**, governed by the adjusted Markov chain, redistributes individuals across counties according to their disease status.

This two-stage framework enables us to simulate the spatiotemporal spread of disease in a computationally tractable yet behaviorally informed manner.

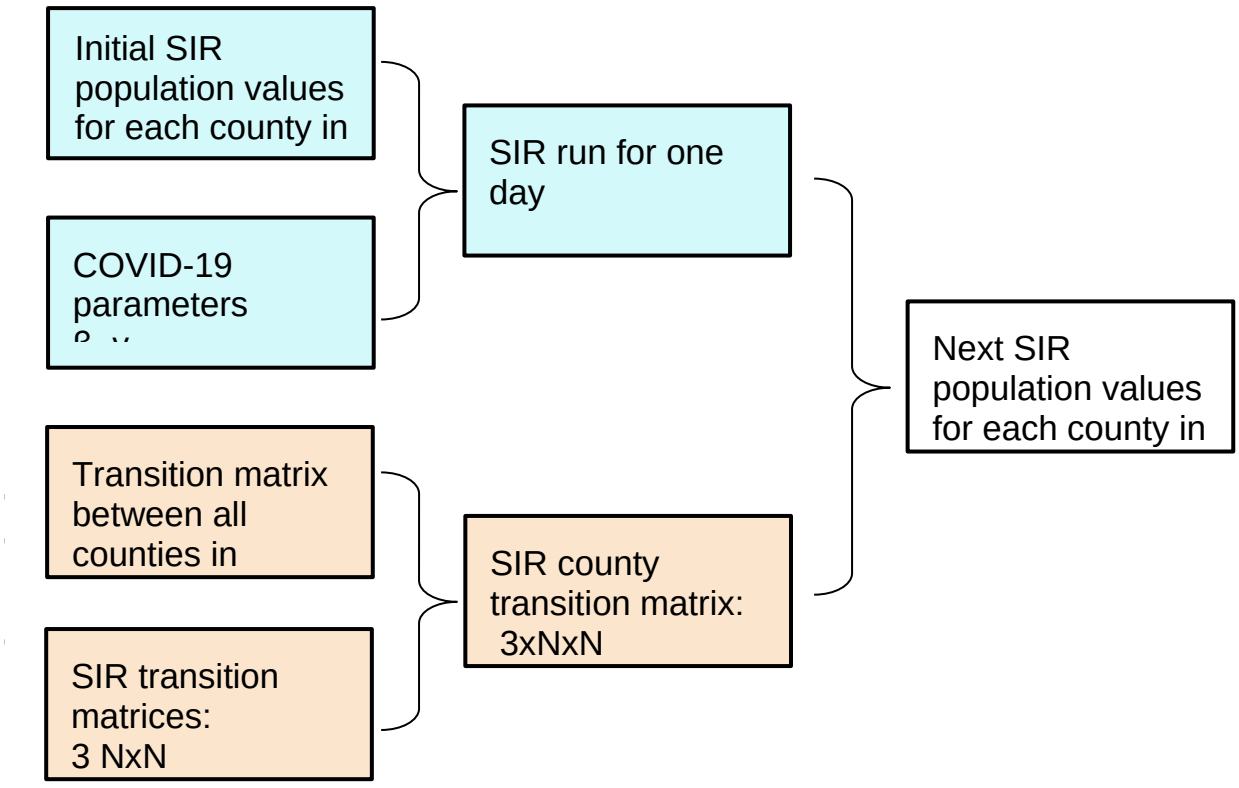

Figure 2. Explanation of the I/O for our SIR-Markov Chain model, showing where parameters β, γ, and the transition matrix are used explicitly.

Simulation

We implemented the conceptual two-stage SIR-Markovian model as a simulator that updates disease dynamics and travel patterns across counties in Georgia at each time step. The simulation proceeds in daily increments and is structured as follows:

County-Level SIR States:
Each County is represented as a row in a population matrix and has three columns corresponding to the proportion of Susceptible (S), Infected (I), and Recovered (R) individuals. The population matrix is stored

as an N x 3 matrix, where N is the number of counties in Georgia. The SIR dynamics within each county are updated using the following approximations of the differential equations:

$$\frac{\delta S}{\delta t} = -\beta SI$$

$$\frac{\delta I}{\delta t} = \beta SI - \gamma I$$

$$\frac{\delta R}{\delta t} = \gamma I$$

Figure 3. SIR set of coupled equations

where the infection rate (β) = 0.5 and the recovery rate (γ) = 0.13, based on COVID-19 SIR parameters [1]. This update is applied independently to each row in the matrix (i.e., each county), using county-specific initial population conditions and shared values of β and γ.

Markov-Chain County Transition Implementation

Population movement between counties is modeled using three N x N matrices:

- S_{travel}: Travel probabilities for susceptible individuals
- I_{travel}: Travel probabilities for infected individuals
- R_{travel}: Travel probabilities for recovered individuals

These matrices have the probability of staying in the same county across the diagonal, and one minus the probability of staying in the same county distributed evenly throughout the remaining counties. The transition matrix between counties is populated using values from Georgia commuter data. The transition matrix for each type of group is created by applying the population movement matrix to the county transition matrix, resulting in three transition matrices. These matrices are then stacked to create a 3 x N x N SIR-transition matrix, which serves as input for our Markov chain model that leverages tensor formats for computational efficiency.

Simulation Flow

The simulator operates as follows:

1. Initialize the SIR Population matrix based on the Georgia 2025 Population data
2. Initialize County Transition Matrix using US Commuter data
3. Initialize S_{travel}, I_{travel}, R_{travel}, matrices
4. Create a stacked SIR-transition matrix
5. For each time step:
    - Update each county's SIR values based on β and γ
    - Apply population movement by multiplying the SIR population matrix with the stacked SIR-transition matrix.

Parameters such as β and γ allow the user to simulate various types of diseases, and initial SIR populations can be edited to test the spread based on various starting counties.

Verification

We ran the apply_SIR function on dummy matrices, single counties, and the initial population matrix with no mobility modifications to ensure that the implementation reproduces expected epidemic curves. These results matched classical SIR behavior, validating the SIR transmission logic. We verified that all three population movement matrices ($S_{travel,}$ $I_{travel,}$ R_{travel}) are row-stochastic to preserve the total number of individuals in each of the respective groups. We also preserve zero entries where no connection exists between counties in the county transition matrix. This ensures valid probability distributions and preserves real-world spatial sparsity. After each complete time step (SIR run + Markov Chain shuffling), we verified that the total population (S + I + R) remained constant, confirming there was no loss/gain of individuals due to implementation error. These verification steps ensured our simulator faithfully implements the conceptual model and can reliably simulate spatio-temporal disease spread.

Experimental Results and Validation

Results

Our analysis compared the performance of the SIR-Markov hybrid model to a baseline pure SIR model, focusing on aggregated data across all counties in Georgia. The hybrid model consistently predicted a higher peak in infections, a divergence attributable to its incorporation of inter-county mobility through the Markov chain-based mixing stage. Unlike the pure SIR model, which treats counties as isolated systems, leaving some regions entirely unaffected once local infections subside, the hybrid approach captures the broader spread of disease through population movement, increasing effective exposure. To isolate the effects of mobility, we examined two contrasting counties: Fulton, which had initial infections at t=0, and Appling, which began with none. In the pure SIR framework, Appling's zero initial infections led to static, unaltered susceptibility throughout the simulation, as the model's isolated design prevented external transmission. The hybrid model, however, introduced infections into Appling over time through inbound travelers, simultaneously increasing the county's susceptible population due to incoming mobility. Fulton County exhibited equally revealing dynamics: the hybrid model showed an early, sharp decline in susceptibles, absent in the pure SIR results, driven by rapid outbound movement redistributing infection risk to neighboring areas. After this initial phase, Fulton's trajectory stabilized, aligning more closely with classic SIR behavior. Together, these findings underscore how population mobility reshapes disease spread, validating the importance of integrating movement dynamics into epidemiological models for more accurate forecasting.

Validation

To verify our model output, we ran our model with the same initial infected population as there was in reality. Our day 1 data was drawn from the number of infected reported by USAFacts.org on March 13, 2020. The following counties were reported with a total of 39 initial infections: Bartow (3), Charlton (1), Cherokee (2), Cobb (8), Dekalb (5), Fayette (4), Floyd (1), Fulton (8), Gordon (2), Gwinnett (2), Lee (1), Lowndes (1), and Polk (1). USAFacts.org provides data that tracks the cumulative number of COVID-19 cases across all the Georgia counties as time progresses. To check if the infections in our model progress as they did in reality, we used the same 39 initial infections as our initial state for our model.

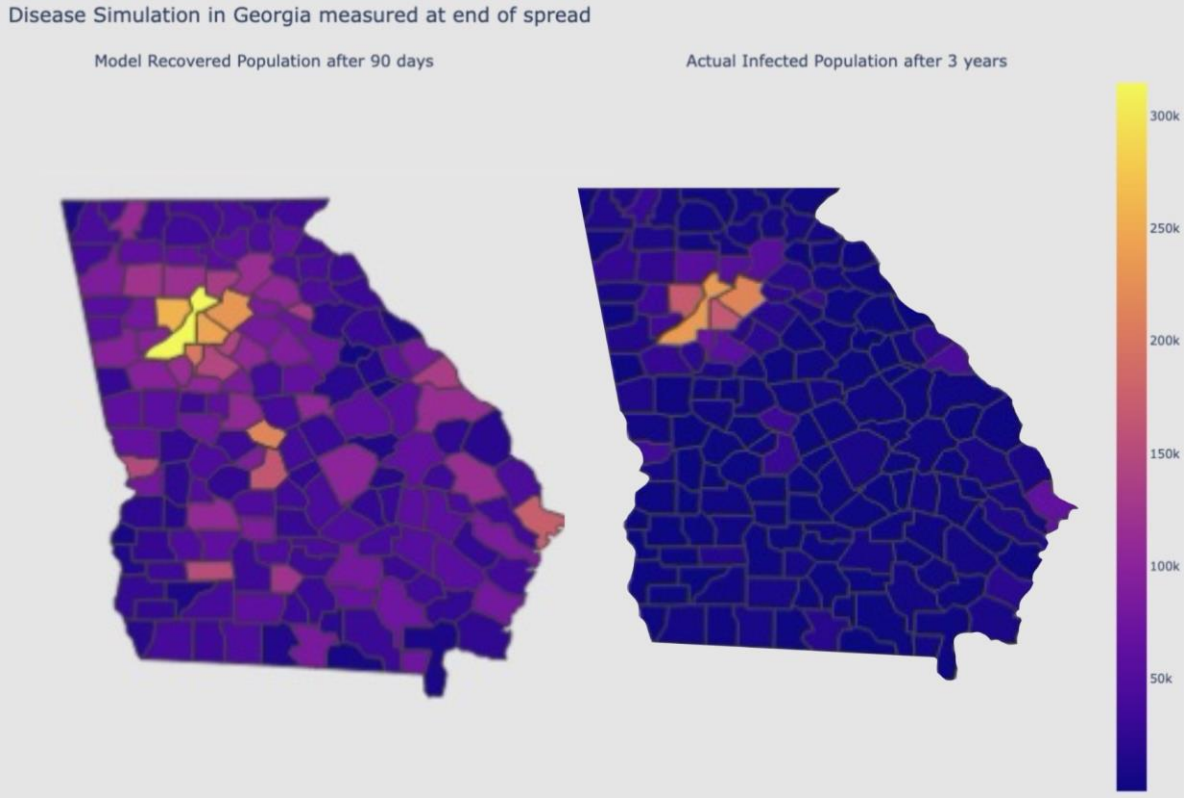

Figure 4. Choropleths of the cumulative infected populations in our model at the end/steady state of the infection spread, which was 90 days for our model (left) and 3 years in reality (right)

Figure 3 shows the resulting choropleths (heat maps) of our model after 90 days and of reality after 3 years, given the same initial state of infection. Because our model doesn't measure cumulative infected totals, we compare our model's recovered populations to reality's cumulative infected populations. Since our model assumes that recovered people had to have been infected, using the recovered number per county is equivalent to finding the cumulative infected population per county. We can use this metric to directly compare to the cumulative infected population in reality. Our model progressed completely within 90 days, and the spread in reality progressed completely in about 3 years, so we were obligated to use different time periods to capture the entire spread progression.

The glaring result is that the infection spread faster and wider in our model than it did in reality. While we did not expect this discrepancy at first, some reasoning allowed us to form an explanation for the state of the result. The SIR component of our model considers only S, I, R, beta, gamma, and travel probabilities for susceptible, infected, and recovered individuals. In reality, more factors are in play to determine infection dispersion. One critical intervention that played a role in reality was people refraining from travel because of the lockdown. The lack of travel, whether it be for leisure or for work, was crucial in limiting the spread of COVID-19. Our Markov Chain is backed by commuter data (from home to work), which does not consider the implementation of a lockdown. This exclusion allows our disease to spread completely in 90 days as opposed to three years. Another consideration for the lack of infections is underreporting. In reality, not all actual COVID cases are reported, leading to lower numbers of COVID cases recorded than the actual number occurring. Our model assumes all infections are reported. If all cases were recorded and no interventions were introduced, COVID-19 would have likely spread more like our model. Given the rapid spread of the virus in our model, we show support for the effectiveness of

lockdowns and interventions in place to reduce the spread, as measured by the drastic difference in results between our model and the viral spread in reality.

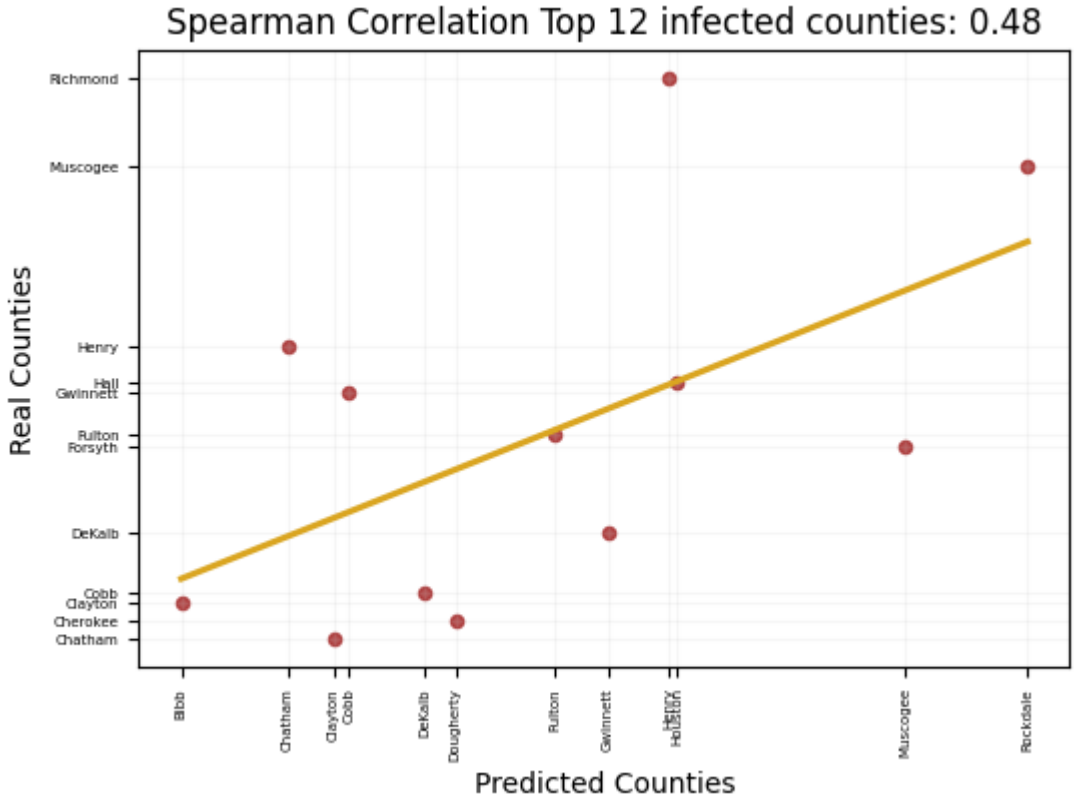

Figure 5. Spearman rank correlation coefficient computed for the first 12 most infected counties reported by our model and real-world data extracted from USAFact.org.

Due to the numerous confounding factors that cause our simulation to overestimate the number of infected cases compared to real-world data, we shifted our focus to analyzing the spatial distribution of the disease. One approach to assess this is by computing the Spearman rank correlation coefficient, which measures the strength and direction of the association between two ranked variables. As shown in Figure 5, the ranking of the most affected counties in the real data is moderately correlated with the rankings produced by our model. This suggests that, despite its limitations, our model captures the spatial dynamics of disease spread more accurately than a standard SIR model. This enhanced spatial resolution aligns with the goals of our study.

Discussion, Conclusions, and Summary

Discussion

The results of our simulation highlight the critical role that mobility plays in the spread of infectious diseases. By integrating a Markov chain mobility model with classical SIR dynamics, our simulator provides a more realistic representation of how a disease propagates across geographically distributed populations. This information can help better inform public health intervention policies before the disease spreads.

A key insight from our validation experiments is the contrast between the outcomes predicted by the standard SIR model and our SIR-Markov hybrid model. The pure SIR model, which treats each county as a closed system, fails to capture the risk posed by cross-county interactions. Such interaction can include work purposes, visitations, errands, etc. In this framework, counties without initial infections remain unaffected throughout the simulation, as seen in Figure 5, regardless of their proximity to the outbreak center or spread. This assumption can lead to falsely constrained pandemic trajectories and underestimates the total infections. This can also lead to overstated and misunderstood local public health interventions.

In contrast, the SIR-Markov Chain Simulation demonstrates how mobility can facilitate the transmission of a disease even into initially uninfected regions, as seen in Figure 6. Appling county started with no infections and eventually saw a spread of COVID introduced via inbound travelers. The hybrid model reveals higher and earlier peaks in infection prevalence compared to the SIR model. This modeling is consistent with the increased contact rates introduced by mobility, which facilitate faster and broaden exposure to infected individuals.

From a public health perspective, our simulator offers a practical tool for evaluating the potential spread of an infectious disease under various disease parameters and mobility scenarios. Public Health Officials and Policymakers can use this model to explore how restricting or modifying inter-county travel might delay or suppress outbreaks. Such policies/interventions can include lockdowns, quarantines, travel advisories, work from home, or transportation policies. The flexibility of the model makes it adaptable for a range of diseases, beyond COVID-19, and a range of states, given their county population and transition data.

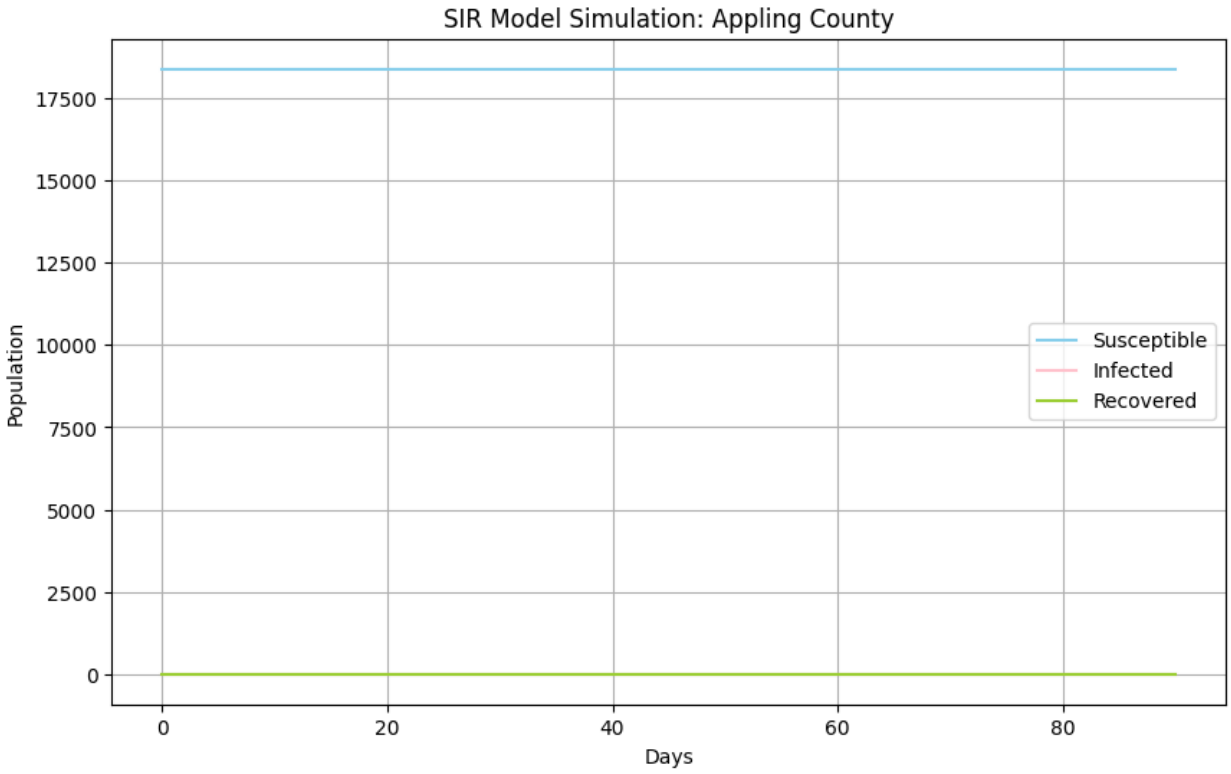

Figure 6. SIR model simulation of infections starting in the counties corresponding to those where COVID-19 began. Appling County was not an initially infected county and due to the closed SIR model, the predicted infection did not spread to Appling County.

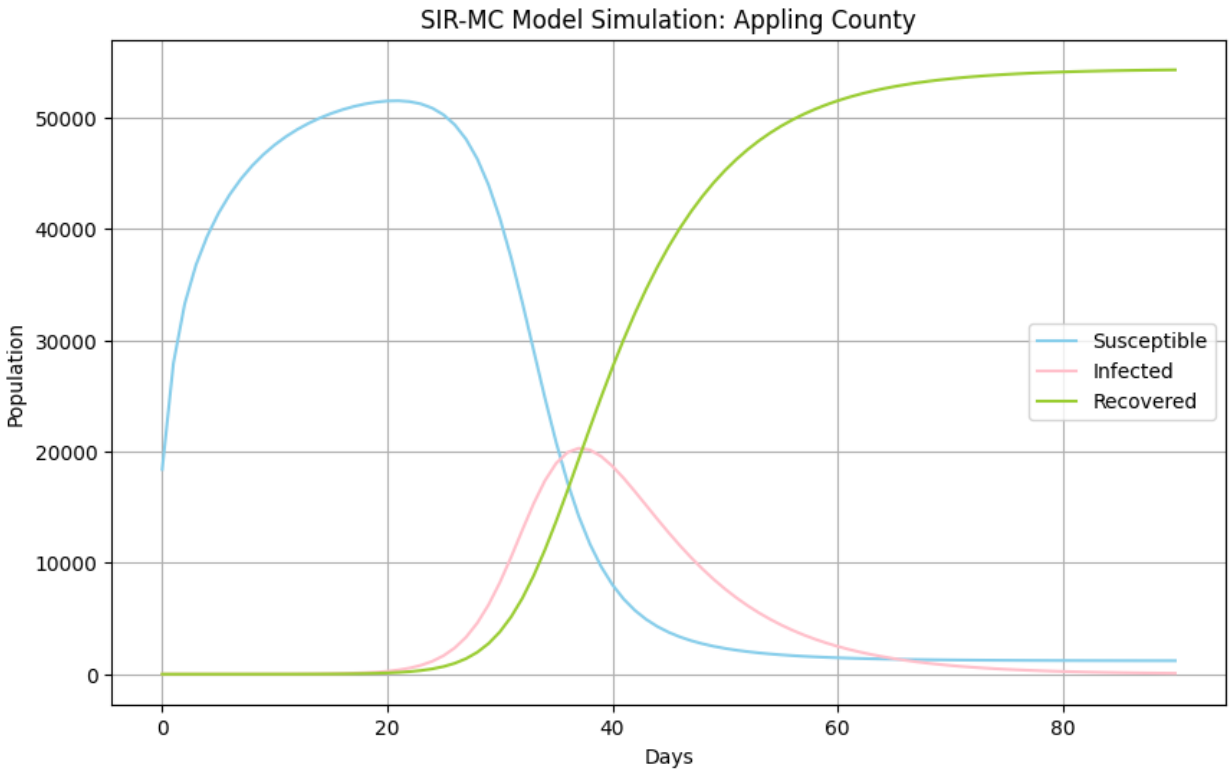

Figure 7. SIR-Markov Chain model simulation of infections starting in counties corresponding to those where COVID-19 began. Appling County was not an initially infected county, but due to the mobility aspect of the hybrid model, the predicted infection spread to Appling County via travel.

Future Work

While our model uses static mobility data derived from commuter data, more dynamic representations could include travel data for work, personal, and emergent reasons. Also, including mobility data from previous outbreaks could help better inform the transition probabilities for susceptible, infected, and recovered populations. This data would differ from the data we used because people would change their normal movement habits based on policies, health purposes, or interventions. This model is flexible as it can be applied to various diseases and therefore can be iterated on in multiple ways.

Summary

The goal of the project was to build a model that can account for granularity in population distribution in real-life scenarios to bridge the gap of spatial boundaries and decrease disease spreading due to far proximity of counties and population movement. A way to account for how strongly the counties are connected to neighboring counties, we associate a probability of traveling to neighboring counties. Also, we account for the proneness to travel of each susceptible, infected, and recovered. This proneness of travel reflects the fact that susceptible population might want to stay at home more to protect themselves, but preserving the overall connection of each county to its corresponding neighbours (we know they want to stay at home more but in case they travel, they will go to the same places as if there weren't any disease).

By accounting for this mixing factor between counties we are able to enhance a zero dimensional disease model like the pure SIR by incorporating a scheme to implement a 1 dimensional model, thus, if

resources need to be assigned to each county, and SIR model would only tell you the total amount of infected individuals expected but not their location, whereas with our model we can inform which county is the most likely to be compromised.
Comparing the distributions

References

[1] AlQadi, H., & Bani-Yaghoub, M. (2022, April 8). Incorporating global dynamics to improve the accuracy of disease models: Example of a COVID-19 sir model. PloS one. https://pmc.ncbi.nlm.nih.gov/articles/PMC8993010/#:~:text=Several%20studies%20indicated%20that%20the,initial%20value%20of%20%CE%B3%20%3D%200.13

[2] Bureau, U. C. (2023a, June 30). *2016-2020 5-year ACS commuting flows*. Census.gov. https://www.census.gov/data/tables/2020/demo/metro-micro/commuting-flows-2020.html

[3] Georgia coronavirus cases and deaths. USAFacts. (2025, April 15). https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/georgia/

[4] Georgia Counties by Population (2025). World Population Review. (n.d.). https://worldpopulationreview.com/us-counties/georgia

[5] Gov. Kemp, officials confirm two cases of covid-19 in Georgia | Georgia Department of Public Health. Georgia Department of Public Health. (n.d.). https://dph.georgia.gov/press-releases/2020-03-02/gov-kemp-officials-confirm-two-cases-covid-19-georgia

[6] Li, Q.-L., Wang, C., Yang, F., & Zhang, C. (2023). Markov modeling and Performance Analysis of infectious diseases with asymptomatic patients. Mathematical Biosciences and Engineering, 20(10), 17822–17848. https://doi.org/10.3934/mbe.2023792

[7] Palopoli, L., Fontanelli, D., Frego, M., & Roveri, M. (2023). A Markovian model for the spread of the SARS-CoV-2 virus. *Automatica*, *151*, 110921–110921. https://doi.org/10.1016/j.automatica.2023.110921

[8] Yaesoubi, R., & Cohen, T. (2011). Generalized Markov models of infectious disease spread: A novel framework for developing dynamic health policies. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2011.07.016

[9] Vivanco-Lira, A. (2020). Predicting COVID-19 distribution in Mexico through a discrete and time-dependent Markov chain and an SIR-like model. arXiv. https://doi.org/10.48550/arXiv.2003.06758

[10] Zakary, O., Rachik, M., & Elmouki, I. (2017). On the analysis of a multi-regions discrete SIR epidemic model: an optimal control approach. International journal of dynamics and control, 5(3), 917–930. https://doi.org/10.1007/s40435-016-0233-2

Appendix: Division of Labor

We all worked on all aspects of the model together during meetings with peer programming. Devanshi led the SIR implementation and active coding for the model. Martin led the Markov Chain implementation and mathematical modeling. Shrey led visualization efforts, implemented the heat maps for both experimental and validation data and handled validation analysis. Each member contributed to every aspect of the final model, final report, and presentation.

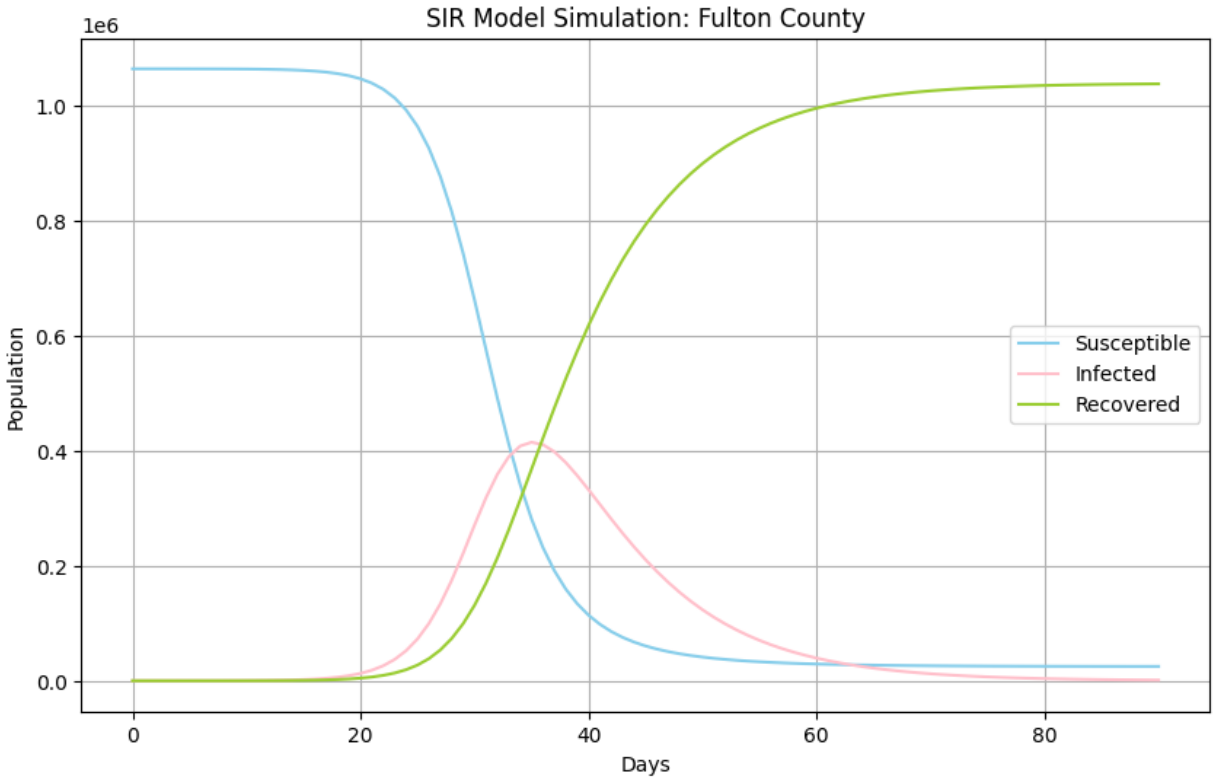

Figure 8. SIR-Markov Chain model simulation of infections starting in counties corresponding to those where COVID-19 began. Fulton County was an initially infected county. Without the mobility aspect of the hybrid model, the predicted infection spread follows a smooth curve according to pure SIR predictions.

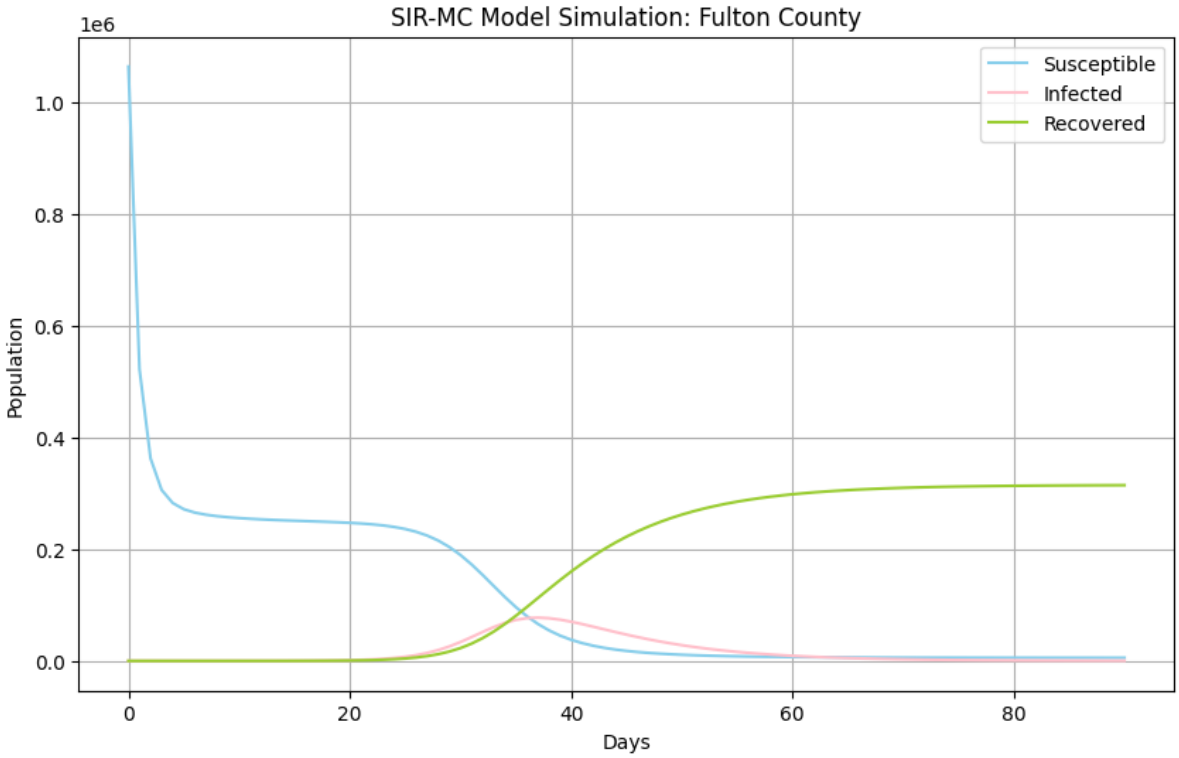

Figure 9. SIR-Markov Chain model simulation of infections starting in counties corresponding to those where COVID-19 began. Fulton County was an initially infected county, due to the mobility aspect of the hybrid model, the predicted infection spread drops suddenly until Markov-stationary state is achieved.

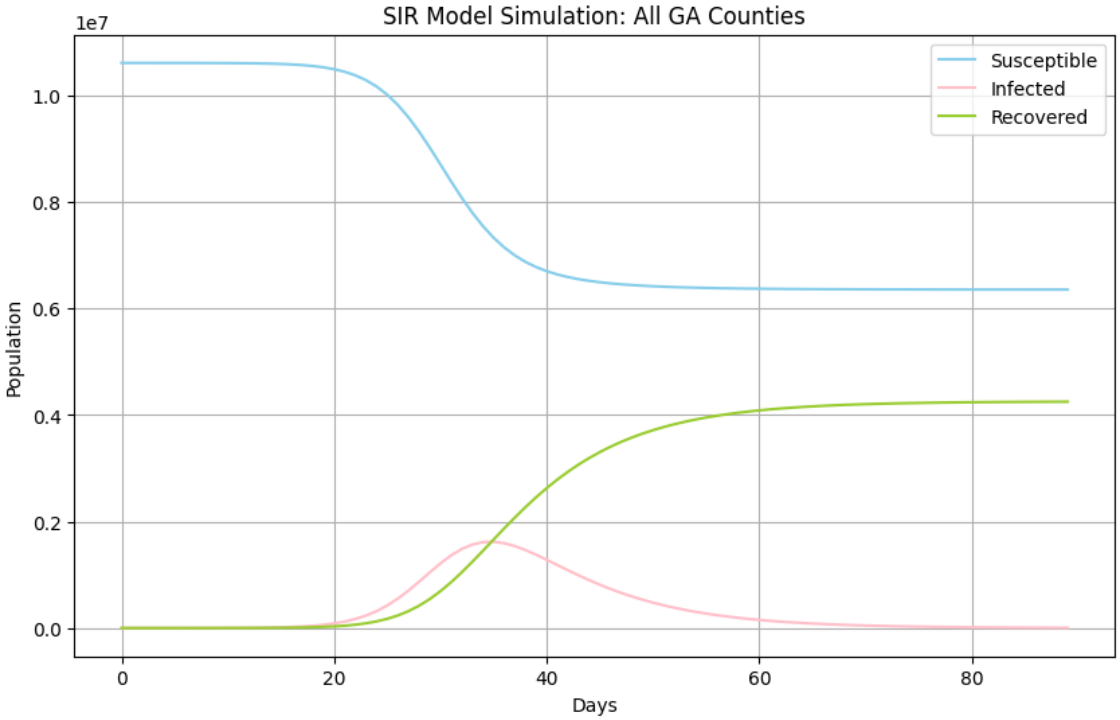

Figure 10. SIR-Markov Chain model simulation of all states in GA, without the mobility aspect applied to the model. The spread does not fully develop as the disease is not allowed to reach all counties.

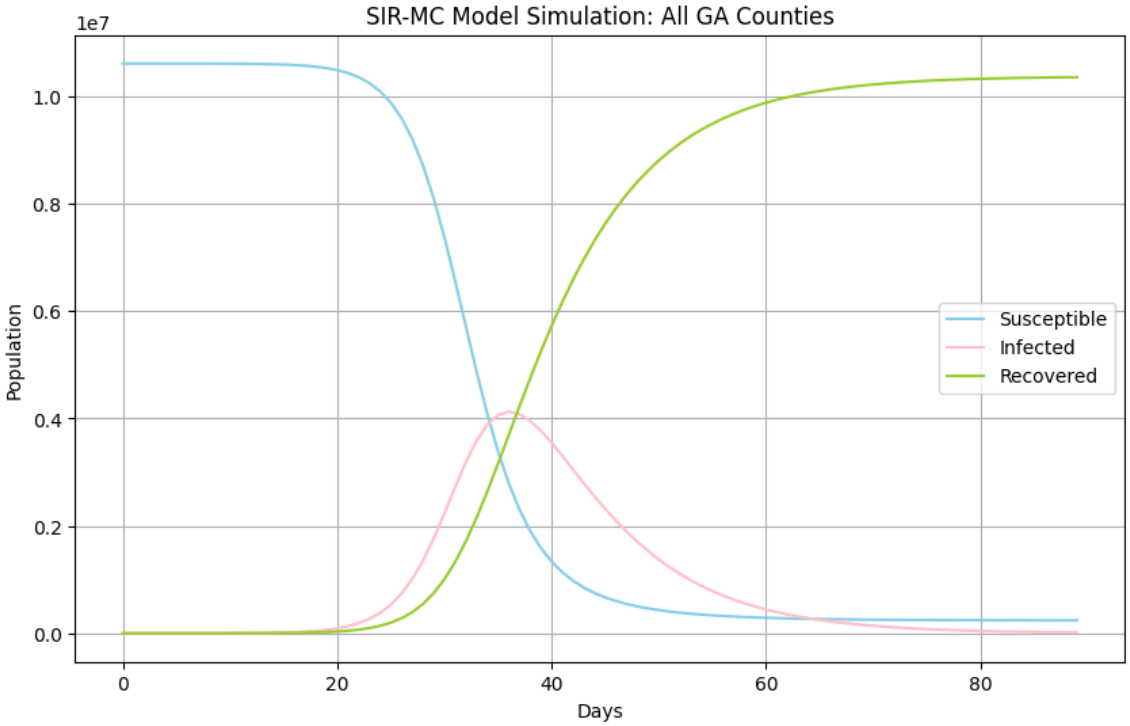

Figure 11. SIR-Markov Chain model simulation of all states in GA, with the mobility aspect applied to the model. The spread is fully developed, reaching all counties at the same number of days but with an accentuated effect.