Purdue University Industrial Engineering

IE332 Project: Phase I

September 10, 2018

Group 11
Michael Wang, Mrunmayi Dandekar, Julia Monti,
Parker Nichols, Karuna Srivastava, Richard Amayo

As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do.
Accountable together - we are Purdue.

Instructor
Dr. Mario Ventresca

1 Introduction

The objective of this project is to create a simulation within a web interface that would assist public officials in the greater Lafayette area in devising strategies to mitigate the spread of pandemic disease. Machine learning and simulation are powerful tools that can be used to model an outbreak with the goal of reviewing the effectiveness of potential disease mitigation policies. The absence of an efficiently designed outbreak protocol would have catastrophic adverse effects on a nation and its people. Phase I of this endeavor entails group policies and expectations, data collection and the synthesis of a realistic population, a brief outline of an algorithm that identifies a subset of K important nodes, and a prototype of our web interface.

2 Group Dynamics and Policies

1. From the formation of our group we set high expectations for each other agreed to strive for 100%. We established a consistent meeting time on Sundays at 1:00 PM while remaining flexible to attend an additional meeting during the week. Although unexpected circumstances do arise, we expect one another to communicate absences as early as possible and to contribute to the project outside of meeting times. All in all, our group expects transparency through communication.

2. For important design decisions, we will discuss all possible avenues and encourage open discussion within the group. If required, we will hold a majority vote. By taking this approach, we assure that no one member can steer the project in a direction in which the group does not feel comfortable.

3. Our goal is to encourage all group members to contribute to discussion and feel comfortable in doing so. In cases of conflicting interests, similar to our design decision policy, we emphasize respect and open discussion.

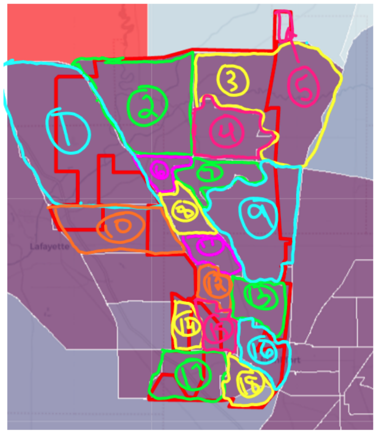
4. We will work on the project as a group and try our best to avoid compartmentalizing the project. To do this, our code will have extensive comments to maintain transparency and to ensure maintenance is user-friendly. If a meeting is missed, it is that individual's responsibility to find a way of contributing. The group also has joint responsibility in ensuring that particular member is kept in the loop.

5. Our vision for the solution is a website that welcomes the user with a brief summary of our purpose and the necessary instructions. Included will be a "Begin Simulation" button that directs the user to the disease simulation interface. There, we will have a section for the user to input and test several factors/policies to influence the outbreak. Once entered, an animation of disease spread in both the base case (no policy enacted) and in the case of the user's modified settings will be available for viewing. See section 5 for further details.

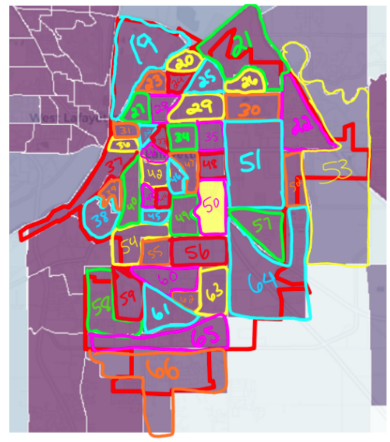
3 Population Simulation

Data Collection

To determine the data necessary to model the Lafayette region, we identified parameters that both describe a realistic population and correspond with the potential to spread the contagion. To properly represent the area, data collected for these variables were recorded by each region defined in Figures 2a and 2b below to account for possible variation. With this in mind, we created our simulated population with each node characterized by a household. We surmised that generating sets of individuals in households would be far simpler to model and analyze as a network. For the houses, we set a type category which had one/two parent families, roommates, single residents, elderly couples, and families with grandpa rents. Other categories include: sex, race, number of people per household, number of children, the schools they attend, seniors above age sixty, income, and public transit utilization. These statistics were utilized to create probability distributions in the model to create households meant to represent the entire region. Occupation and place of work were considered and will be implemented into a later iteration of the model after further research has been conducted.



(a) Division of West Lafayette into region number



(b) Division of West Lafayette into region number

Algorithm Overview

All aspects of this simulation were developed in R-Studio. The algorithm reads our data file and simulates a population of the 40490 households in the Lafayette and West Lafayette area with their respective characteristics. The simulation fills a data frame which then writes to a .csv file as output.

Algorithm Logic and Assumptions

1. Initialize data frame with number of rows equal to the number of households desired and one column for each of the parameters: "Region", "Households", "Income", "nChildren", "nPeople", "childAges", "Schools", "Race", "Males", "Females", "nOld", "PubTrans".

2. Simulate region number based on the probability per region i: $P_i = \frac{Population_i}{\sum(Population)}$. Students at Purdue University who live in on-campus dormitories as illustrated in Figures 2a and 2b are currently not considered among those impacted by the disease due to its uniqueness. These students will be added to our generated population in the near future.

3. Simulate household type based on two separate discrete probability distributions - one for West Lafayette and one for Lafayette. Possible household types include "HW" (Husband/Wife), "HW_O" (Husband/Wife (Old)), "A" (Alone), "AL_O" (Alone (Old)), "SP" (Single Parent), "Fam_O" (Family (Old)), and "R" (Roommates). We assumed that the type of household is independent of the region, but dependent on city. This assumption may need modification with more descriptive data since specific regions with both cities may have different distributions for household type.

4. Simulate number of children based on the number of households (Poisson distribution). We assume that the distribution of the number of children in each household is accurately represented by the country-wide data gathered in the most recent U.S. Census.

5. Calculate number of old people ($\geq$ 60 years of age) based on the household type. Household types "HW_O", "AL_O", and "Fam_O" all contribute to the data frame column "nOld". "Fam_O" entails a family with exactly one old person.

6. Total number of people per household is calculated in two parts - one based on the household type and the other based on the sum of columns "nChildren" and "nOld". For the household type, those living alone would only have one person and households with roommates sample from a uniform distribution between 2 and 6. Otherwise, "HW" families have two adults while "SP" households will have one adult.

7. Determine which houses are below the poverty line the use of public transportation based on region-dependent probabilities. We also assume that household income and use of public transportation are

dependent variables, but that income and race are independent variables. In the near future, we plan on sampling income from a continuous distribution and taking into account dependencies between race and income.

8. After determining the number of children, assign each simulated child to either "Pre-K", "K", "Elem", "Mid", or "High" using region-dependent probabilities.

9. After using the cities' school boundaries to formulate each region's respective elementary, middle, and high schools, assign school(s) based on number of children, their categorical ages, and which region they live in. We assume that all children attend the designated public schools within their region.

10. Determine number of Males based on household type and/or region-dependent binomial distribution. "HW" and "HW_O" households are assumed to have one male, plus the number of males sampled for the remainder of household members. The remainder of household types sample from the region-dependent binomial distribution for n number of times where n is the number of people within that household. For females, the number is determined by subtracting the number of males from total number of people in each household.

11. Other potential characteristics such as occupation, visitation public areas ma be added in a similar manner as described above by sampling from a probability distribution while considering the dependencies.

12. Write simulated dataframe to .csv file titled "generatedPop.csv".

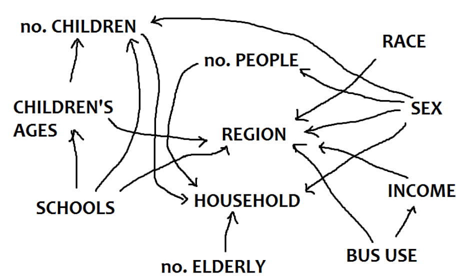


Figure 2: Network of dependencies of generated parameters. Connected parameters are non-independent and the algorithm was implemented based on these relationships.

4 Identification of Nodes

For users to simulate the implementation of various policies of disease mitigation, we require a list of "nodes" that correspond to the means of potential contact between households/individuals. Based on our current population characteristics, these nodes would contain all West Lafayette and Lafayette Schools, a list of public transit routes, and the sixty-six regions. Other possibilities for nodes include public meeting places, workplaces, and neighborhoods. We have developed a preliminary algorithm that formulates a list of households (indices) within the simulated population connected to each "node." This enables the calculation of the total households interconnected through a specific node, and as a result, the nodes with the most connections. The nodes with the most connections would be categorized as most important. With this algorithm, we can generate a network in which all households are connected to their respective nodes. Then, by manipulating the network through the removal of certain nodes, the user is able to visualize the magnitude of impact of various policies.

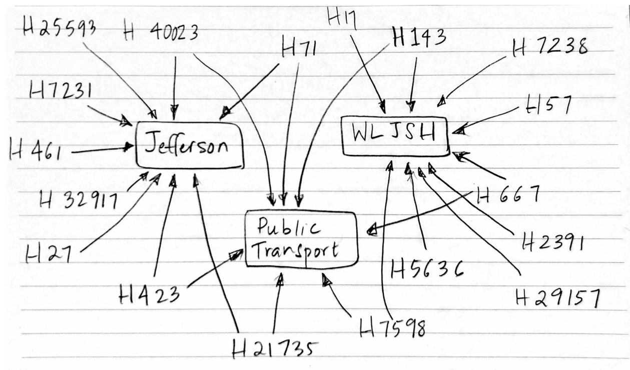
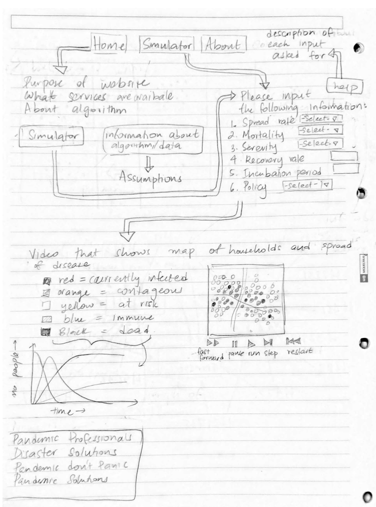
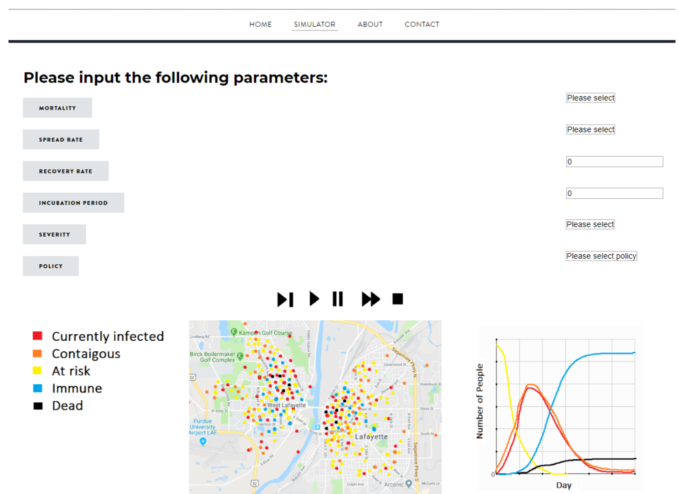


Figure 3: Example of potential nodes; Jefferson & West Lafayette Jr/Sr High Schools and Public Transport. Households are classified as separate kind of "node" and each connect not to each other, but to central nodes.

5 Website Design



(a) Website layout sketch outlining the components of each web page



(b) Prototype of user interface for disease simulator with various inputs for the user along with output animation and graph

Website link: `https://disastersoltuions.weebly.com/`

Conclusion

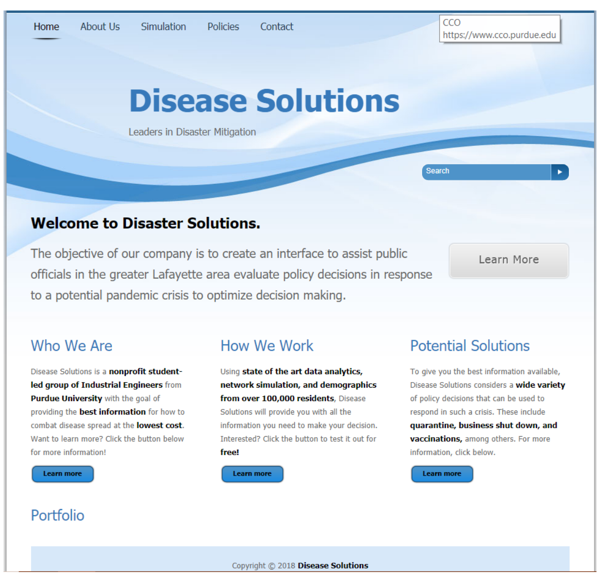
Designing a fully representational simulation is paramount to predicting potential epidemics. Phase I of the project laid the foundations for the research and development of the database, website, and optimization algorithms in Phase II. The simulation synthesized a population in which each household corresponds to a particular index, which in turn is connected to central nodes. These create a network that the user can interface with in order to model the effects of various implementation of outbreak protocols. With the solidification of the group dynamic and individual expectations, the group can be expected to perform exceptionally in Phase II undeterred by any potential difficulties.

6 Website

(a) Website Home Page

(b) Website layout

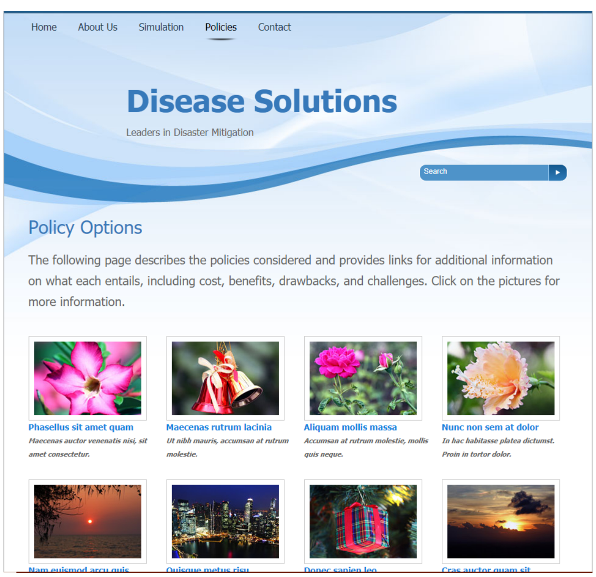
7 Website



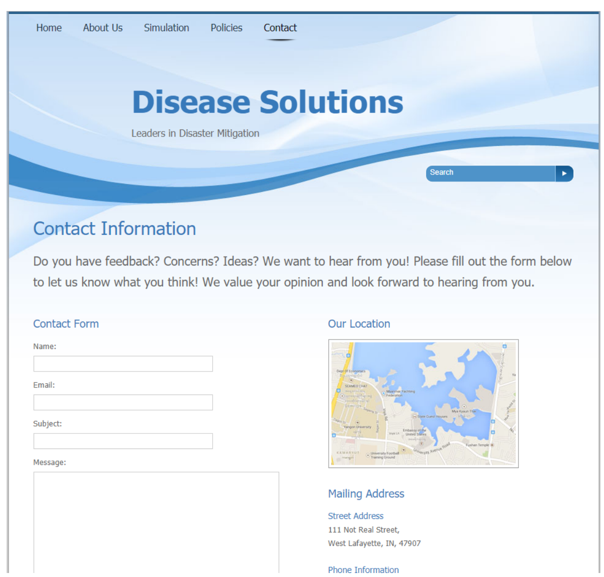
(a) Website Home Page

(b) Website About Us Page

(a) Website Simulation Page



(a) Website Policies Page



(b) Website Contact Page

As it currently stands, the website consists of five functioning web pages shown above.

References

[1] Disease Spread Simulation
https://www.learner.org/courses/envsci/interactives/disease/disease.html

[2] LaTeX Tutorial and Reference
https://www.overleaf.com/latex/learn/free-online-introduction-to-latex-part-1
https://www.overleaf.com/latex/learn/free-online-introduction-to-latex-part-2

[3] Monte Carlo Simulation
https://rpubs.com/Koba/Monte-Carlo-Basic-Example

[4] Population Statistics
https://www.neighborhoodscout.com/in/lafayette/demographics
https://www.census.gov/quickfacts/fact/table/westlafayettecityindiana/PST045217
https://datausa.io/profile/geo/west-lafayette-in/
http://www.city-data.com/city/West-Lafayette-Indiana.html