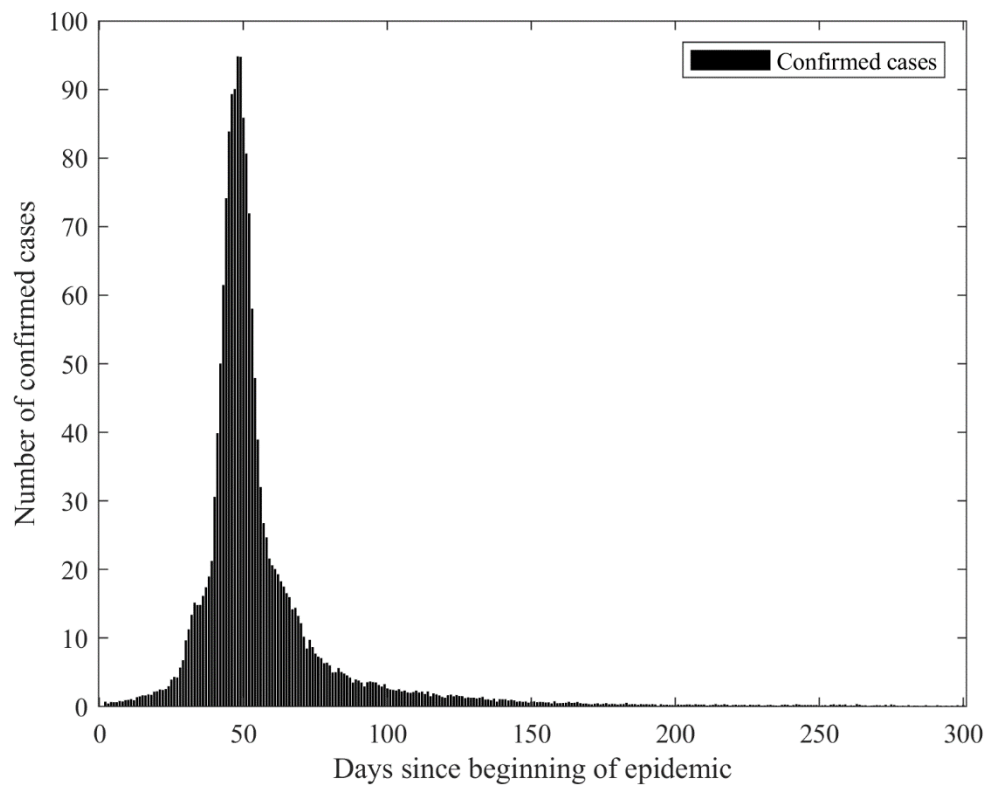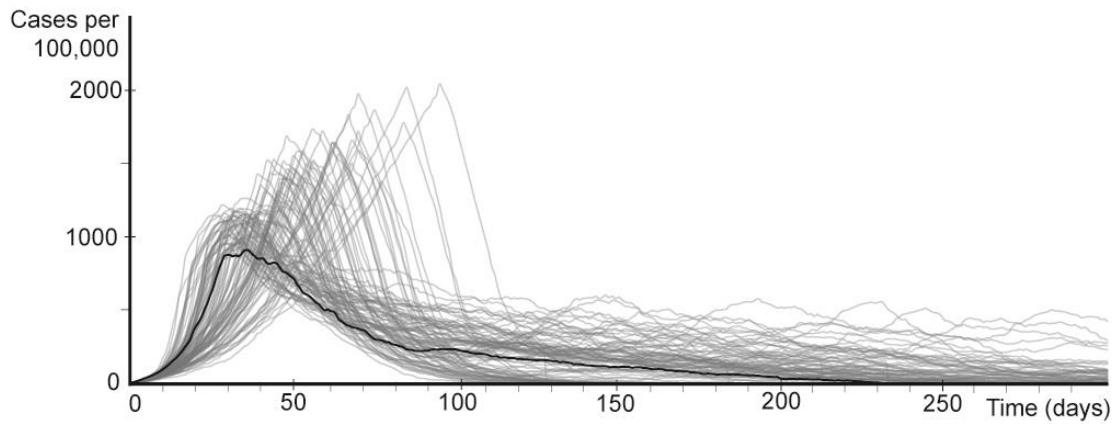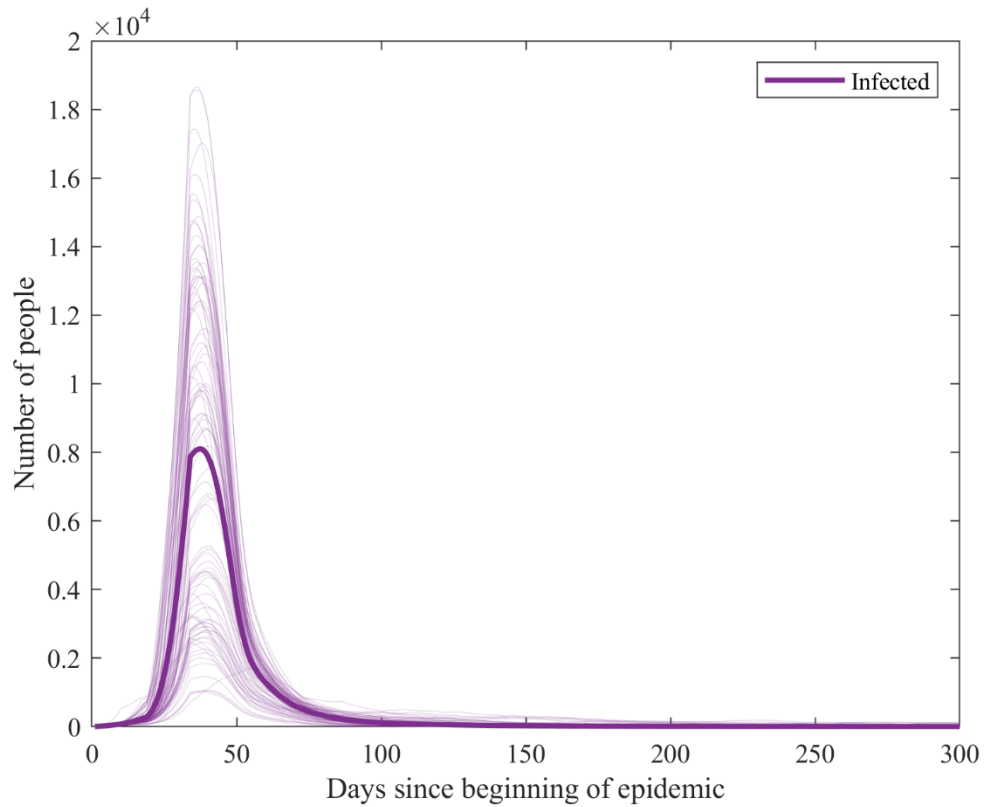# Quick start guide for geoSIR stochastic agent based epidimiology model

----

By Max Van Wyk de Vries and Lekaashree Rambabu, June 2020

Epidemic Progression

Legend: Unaffected, Infected, Recovered, Died
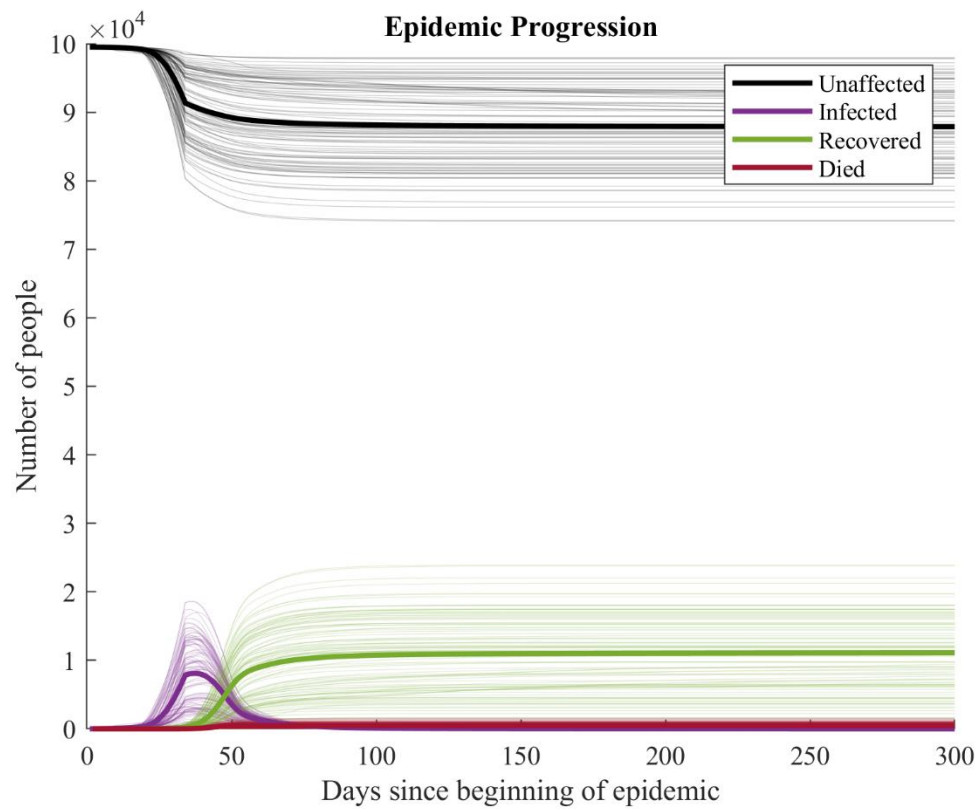
Figures: Theoretical Run 6 and combination of all theoretical runs.

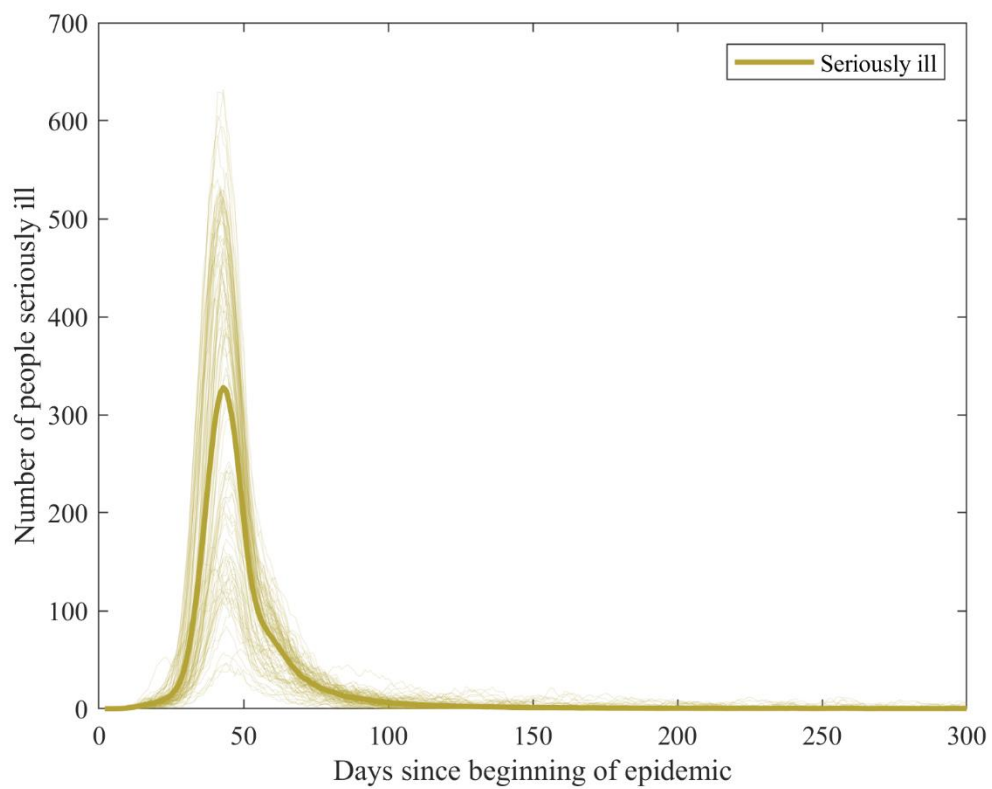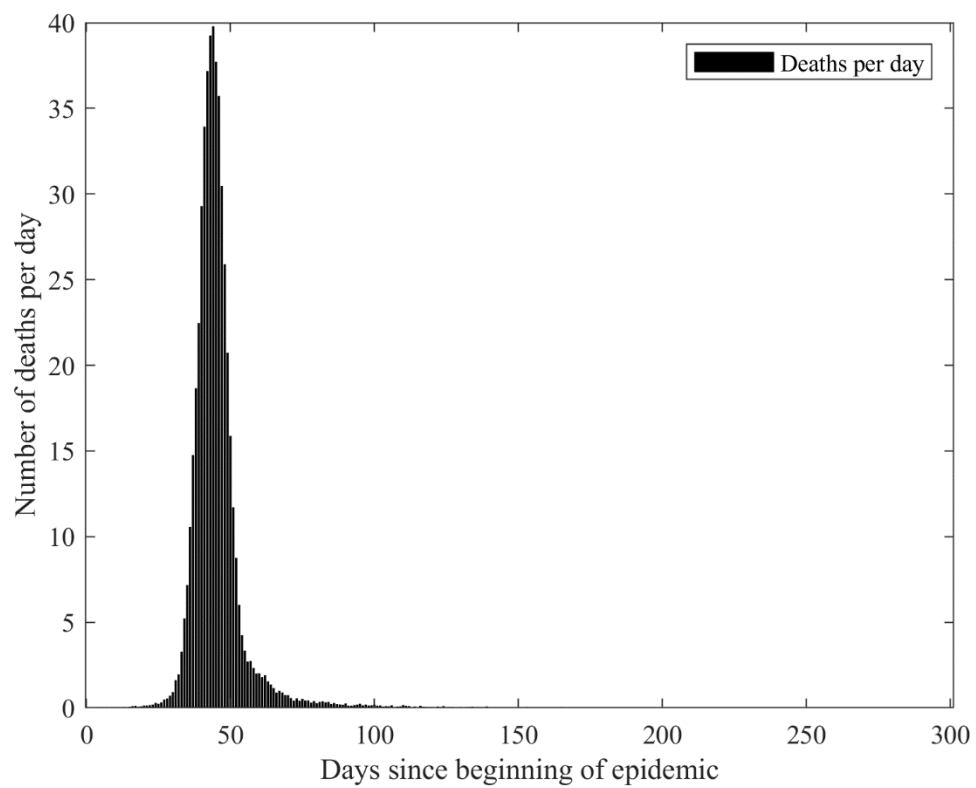--

Contents:

1 - Basic introduction

2 - Contents and model structure

3 - Input parameters

4 - Examples

5 - Licensing and other issues

----

1 - Introduction

geoSIR is an epidimilogy that follows the rough framework of all SIR models. Individuals are in one of several states. The basic states are susceptible (may catch the disease), infected (has the disease, may pass it on, may die), dead (died of the disease) and recovered (had the disease, now immune to it).

Unlike deterministic models that solve a series of differential equations to evaluate the spread of the disease, geoSIR models each individual. Infected individuals make a certain number of encounters each day, and may spread the disease to other susceptible individuals.

The inputs file (geoSIR_inputs) is the only file that needs to be modified prior to each run, and contains all the necessary input values.

----

2 - Contents and model structure

The model is designed with one main input file, one main model run file and one plotting file. This sequence may then be duplicated in a wrapper script in order to evaluate a number of different scenarios automatically. The script files included in this package are:

- geoSIR_wrapper

This is the wrapper script that is used to automate a large number of runs. It may for example be set to run a large number of scenarios overnight for large runs. The format of it is:

%% geoSIR_wrapper

> geoSIR_inputs

> geoSIR_run

> geoSIR_plot

%% end

or in the case of several sequential scenarios:

%% geoSIR_wrapper

> geoSIR_inputs1

> geoSIR_run1

> geoSIR_plot1

> geoSIR_inputs2

> geoSIR_run2

> geoSIR_plot2

> geoSIR_inputs3

> geoSIR_run3

> geoSIR_plot3

%% end

The different geoSIR_plot files will automatically save the outputs in different folders for future reference.

- geoSIR_inputs

This file defines all of the input parameters necessary for a succesful model run. Parameters may be chosen at random to represent a 'theoretical' disease, however runs will be most meaningful when data is taken from real world studies. The model is designed such that the input parameters can be observed in real world data.

See the 'input parameters' section for more details.

- population_density_generate

This file load real world gridded population data and hazard zone data and transforms it into a 1x1 km grid population density map and 1x1 km hazard map. These files are used to build the model simulation area.

Note that this script may be skipped if a theoretical region is being used.

- coordtom

Short script to transform geographical latitude and longitude coordinates to distances using the haversine formula. Used so that the lat-long of region corners is the only necessary input.

- geoSIR_run

This file simulates the spread of a disease throughout the model region.

The basic structure involves a time loop and a spatial loop as so:

%% geo_SIR_run

for time_loop = start_time (days) to end_time (days)

for spatial_loop = bottom left_grid_corner to top_right_grid_corner

 [CALCULATE ENCOUNTERS AND DISEASE SPREAD]

end spatial_loop

[RECORD DATA FOR THIS TIME STEP]

end time_loop

%% end

With the [CALCULATE ENCOUNTERS AND DISEASE SPREAD] portion depending on the specific input parameters. Infected individuals make a certain number of close and distant encounters each day (may be thought of as 'household/friends/office' and 'community/fomites' transmission), and may spread the disease accordingly. A certain number of statistics are recorded at each time step. These are:

1) Number of susceptible individuals (total at this time step)

2) Number of infected individuals (total at this time step)

3) Number of recovered individuals (total at this time step)

4) Number of fatalities (total at this time step)

5) Number of asymptomatic carriers (total at this time step)

6) Number of severely ill individuals (total at this time step)

7) Infection mortality rate

8) Apparent mortality rate (deaths / confirmed cases)

9) Number of self isolation + quarantined individuals

10) Number of confirmed (tested) cases

11) Number of test false negatives

12) Number of test false positives

13) Number of imported cases

14) Daily deaths

15) Daily new cases

16) Lockdown status (binary)

17) Travel ban status (binary)

18) Daily severely ill cases

19) Five day infection rate (mean and standard deviation)

20) Basic reproductive number (mean and standard deviation)

21) [if hazard zones have been defined] Number living in the high risk zone

22) [if hazard zones have been defined] Number living in the medium risk zone

23) [if hazard zones have been defined] Number living in the low risk zone

The full timeseries of the above variables is the fed into the plotting file to be saved as an outputs file.

- geo_SIR_determine_reproductive_number

This function iteratively calculates the mean and standard deviation basic reproductive number (R0) and five day spread parameter (R_5D).

R0 is given by the following formula:

R0 = [expected number of individuals infected by one single individual in one day]

    x [expected duration of infection];

and R_5D is given by:

R_5D = (1+[expected number of individuals infected by one single individual in one day])^5

The script is relatively long as the [expected number of individuals infected by one single individual in one day] parameter can vary depending on a number of conditions. In particular, this value depends on whether a natural disaster is ongoing, whether a lockdown has been implemented and on several input parameters.

A number of iterations for these calculations may be selected in the inputs. The mean and standard deviation of the results for all iterations is then recorded. The standard deviation provides a rough estimate of the spatial variability of the disease spread parameters within the study area.

- geoSIR_cropNmean

This short script calculates the mean of a number of different runs, removes any trailing Not a Number (NaN) values and either extrapolates the final value of a series or adds zeros. It is called for convenience to calculate mean values for different runs and generate summary plots.

Summary values may also be calculated manually from the saved outputs file.

- geoSIR_run_theoretical

The same as the other geoSIR_run file, however has been modified to generate an idealised region rather than to load in real-world data.

- geoSIR_plot

This file draws a number of summary plots, and saves the inputs and outputs to a given file.

First this function will check whether a results file for this run exists, and create it if it does not. This ensures different runs are saved in separate files and do not overwrite each other.

Secondly, it saves the .mat inputs and outputs files. These may then be scrutinised at any point, and re-run with identical input parameters.

Finally, it generates a number of output plots (Daily deaths and cases histogram, number of infected at any given time, number recovered, death rate, etc.). These plots are saved as high resolution .png or .jpg files, and may be suitable to include in reports, publications etc. See a number of examples in the 'Theoretical example results' file.

Note if running on a supercomputer cluser: you may wish to edit this file such that it only saves the inputs and output files. Generating and saving plots may cause the runs to fail, and plots can always be generated from the outputs file at any later date.

- geoSIR_movie_maker

This function may be used to generate a .mp4 or .avi video of the two dimentional spread of the disease at each time step. For instance a video of the number of infectious individuals, or the total number of deaths in each zone through time can be generated. This may help observe if cases are clustered in high or low population density areas, or around natural disaster prone hotspots.

----

3 - Input parameters

Suitable input parameters are key to a succesful simulation of real-world phenomena. In this section we run through the different possible input parameters in the geoSIR model. These are defined in the geoSIR_inputs file and loaded into a structure array. Stucture array parts are defined via dot indexing for convenince (e.g. for a structure array named "inputs", a new sub-section "name" may be defined as "inputs.name").

Inputs are:

Name

The name that the simulation is saved under. For example "Test case 1, no eruption, no lockdown".

Number of people

The total population of the model region. A value of 100,000 will lead to a model region with 100 thousand people in it being generated. Large population sizes will generally lead to more realistic results, however may be computationally expensive. As a rule of thumb, any simulation over 500,000 individuals may benefit from being run on a supercomputer node.

Population density

If a uniform population density is being used, and no gridded population map is being loaded. Defines the initial ratio of susceptible individuals to (susceptible individuals + empty space). This parameter may be ignored if real-world data is being loaded, or a specific theorietical population density is being generated.

Number of close encounters (min and max)

Defines a minimum and maximum number of close encounters that individuals may have. The distribution between the minimum and maximum is uniform, and the result must be a whole number. Min and max may be the same value.

Distance travelled for close encounters (min and max)

Defines a minimum and maximum distance travelled for close encounters. The distribution between the minimum and maximum is uniform, and the result must be a whole number. Min and max may be the same value.

Probability of infection for close encounters (mean and standard deviation)

Defines the mean and standard deviation for a normal distribution representing the probability of infection being transmitted during a close encounter. These values may in some cases be estimated from detailed real world case studies.

Number of distant encounters (min and max)

Defines a minimum and maximum number of distant encounters that individuals may have. The distribution between the minimum and maximum is uniform, and the result must be a whole number. Min and max may be the same value. Typically higher than the number of close encounters.

Distance travelled for distant encounters (min and max)

Defines a minimum and maximum distance travelled for distant encounters. The distribution between the minimum and maximum is uniform, and the result must be a whole number. Min and max may be the same value. Typically higher than for close encounters.

Probability of infection for close encounters (mean and standard deviation)

Defines the mean and standard deviation for a normal distribution representing the probability of infection being transmitted during a distant encounter. These values may in some cases be estimated from detailed real world case studies. Typically lower than for close encounters.

Number of incoming travellers

Number of travellers ariving from outside of the model region. Will be reduced to 0 in the event of a travel ban.

Number of outgoing travellers

Number of travellers leaving the model region. These are selected at random from the entire population. Will be reduced to 0 in the event of a travel ban.

Fraction of infected travellers

Proportion of infected travellers. For a value of 0, no infections are imported. For a value of 1, all incoming travellers are infected. Incoming travellers are assumed to be newly infected.

Risk of severe illness (mean and standard deviation)

Risk of any mildly ill (i.e. infected, not asymptomatic, not on incubation period) individual becoming severely ill. Our definition of severely ill is loosely based on a combination of the "severely ill" and "critically ill" categories from the large dataset study of COVID-19 cases in China by Wu and McGoowan (2020). Critically ill patients may take up hospital capacity, and may die. A normal distribution is assumed. Note: may be set to 1 if all individuals infected by a given disease are at risk of dying.

Proportion asymptomatic (mean and standard deviation)

Chance of a newly infected individual of being entirely asymptomatic, sampled from a normal distribution. May be set to 0 if there are no asymptomatic individuals. Asymptomatic individuals do not get ill, do not take up hospital capacity and cannot die. They can however infect other individuals.

Death rate of severely ill individuals (mean and standard deviation)

Chance of a severly ill infected individual of dying, sampled from a normal distribution. If an individual is severely ill for t days, they will have a [Death rate of severely ill individuals]/t chance of dying every day. May be increased up to a threshold if hospital capacity is exceeded. Note that this is not the the overall death rate, the case fatality or the infection fatality rate. This is the probability of dying for an individual that is already severely ill.

Hospital capacity

Total hospital capacity. May for instance represent the number of ICU beds, the number of ventilators or some other relevant metric. Death rate may then be increased if exceeded. May be set as equal to the total population if hospital capacity is irrelevant for this disease.

Sensitivity to overcrowding

Provides a numberical value for how rapidly death rate will increase when hospital capacity is exceeded. For a value of 1, a 10% exeedence in hospital capacity will result in a 10% increase in deaths.

Maximum death rate

Maximum death rate for severely ill individuals. Even if hospital capacity is exceeded by a large factor, no more than [Maximum death rate] of severely ill individuals will die.

Duration of symptoms (mean and standard deviation)

The duration of symptoms for any given individual, sampled from a normal distribution. During this period individuals may become severely ill or die. They also have a higher probability of self isolating or being quarantined.

Duration of incubation period(mean and standard deviation)

The duration of the incubation period for any given individual, sampled from a normal distribution. May be set to 0 for diseases with no incubation period.

Chance of self isolation for non-symptomatic individuals (min and max)

Chance of self isolation or quarantine for a non-symptomatic individual (in the absence of any lockdown). Non-symptomatic individuals may be susceptible, asymptomatically infected or infected and on the incubation period. Self isolating individuals make fewer or no daily encounters. Simulates social distancing measures.

Efficiency of self isolation for non-symptomatic individuals (min and max)

Efficiency of self isolation or quarantine for a non-symptomatic individual. Non-symptomatic individuals may be susceptible, asymptomatically infected or infected and on the incubation period. For a value of 1, no trips are made. For a value of 0.5, 50% of daily trips are reduced. Simulates social distancing measures.

Chance of self isolation for symptomatic individuals (min and max)

Chance of self isolation or quarantine for a symptomatic individual (in the absence of any lockdown). Symptomatic individuals may be mildly or severely ill. Self isolating individuals

make fewer or no daily encounters. Simulates social distancing measures. Typically higher than for non-symptomatic individuals.

Efficiency of self isolation for symptomatic individuals (min and max)

Efficiency of self isolation or quarantine for a non-symptomatic individual. Symptomatic individuals may be mildly or severely ill. For a value of 1, no trips are made. For a value of 0.5, 50% of daily trips are reduced. Simulates social distancing measures. Typically higher than for non-symptomatic individuals.

Number of tests conducted on non-symptomatic individuals (min and max)

Daily number of tests conducted randomly on non-symptomatic individuals. Non-symptomatic individuals may be susceptible, asymptomatically infected or infected and on the incubation period. Positive tests self isolate/are quarantined. Positive tests determine the number of confirmed cases.

Number of tests conducted on symptomatic individuals (min and max)

Daily number of tests conducted on symptomatic individuals. Symptomatic individuals may be mildly or severely ill. All positive tests self isolate/are quarantined. Positive tests determine the number of confirmed cases.

Test false positive rate (min and max)

False positive rate for a given test.

Test false negative rate (min and max)

False negative rate for a given test.

Lockdown status

Binary Yes / No. For yes, a lockdown is implemented once cases or deaths reach a certain threshold.

Lockdown entry condition: cases

Number of new daily cases which, if exceeded, triggers a lockdown.

Lockdown entry condition: deaths

Number of new daily deaths which, if exceeded, triggers a lockdown.

Lockdown exit condition: cases

Threshold that the number of new daily cases must be under for the lockdown to be released. May be set to -1 for a lockdown that is not released.

Lockdown exit condition: deaths

Threshold that the number of new daily deaths must be under for the lockdown to be released. May be set to -1 for a lockdown that is not released.

Lockdown chance of self isolation for non-symptomatic individuals (min and max)

Chance of self isolation or quarantine for a non-symptomatic individual during lockdown. Non-symptomatic individuals may be susceptible, asymptomatically infected or infected and on the incubation period. Self isolating individuals make fewer or no daily encounters. Simulates social distancing measures. Higher than for no lockdown, and may be set to 1 for a 'perfect' lockdown.

Lockdown chance of self isolation for symptomatic individuals (min and max)

Chance of self isolation or quarantine for a symptomatic individual during lockdown. Symptomatic individuals may be mildly or severely ill. Self isolating individuals make fewer or no daily encounters. Simulates social distancing measures. Typically higher than for non-symptomatic individuals. Higher than for no lockdown, and may be set to 1 for a 'perfect' lockdown.

Distance moved reduction factor.

Reduction of the distance travelled during any remaining encounters during lockdown. For a value of 2, the distance travelled during lockdown is halved.

Travel ban status

Binary Yes / No. For yes, a travel ban is implemented once cases or deaths reach a certain threshold.

Travel banentry condition: cases

Number of new daily cases which, if exceeded, triggers a travel ban

Travel ban entry condition: deaths

Number of new daily deaths which, if exceeded, triggers a travel ban

Travel ban exit condition: cases

Threshold that the number of new daily cases must be under for the travel ban to be released. May be set to -1 for a travel ban that is not released.

Travel ban exit condition: deaths

Threshold that the number of new daily deaths must be under for the travel ban to be released. May be set to -1 for a travel ban that is not released.

Maximum time

Maximum model time, in days. The run will end early if the number of infections drops to zero.

Number of iterations

Number of separate iterations for these input parameters. Due to the stochasatic set-up, different runs will result in different output data for the same inputs. Runs may be conducted in parralel for efficiency.

Parralel run status

Binary Yes / No. For yes, different iterations will be run in parralel. The default number of cores is the maximum number of physical cores available on the machine running the code. Parralelisation may greatly speed up results.

Number of iteration for the Basic Reproductive number

Number of estimates of R0 and R_5D calculated for the mean and standard deviation.

Natural disaster status

Binary Yes / No. For yes, a natural disaster will occur according to certain user defined parameters.

Natural disaster case

Several cases may be defined for different evacuation and mobility scenarios. The default value is 4, in this case individuals in the high hazard and 25% of individuals in the moderate hazard zones are evacuated, and mobility is increased.

Natural disaster timing

The start timing of the natural disaster in days.

Natural disaster duration

The duration of the natural disaster in days.

Record statistics (infected)

Record a 2D grid of number of infection cases at each time step that be used to generate a video using geoSIR_movie_maker script. May be memory intensive for large population sizes.

Record statistics (severe)

Record a 2D grid of number of severely ill cases at each time step that be used to generate a video using geoSIR_movie_maker script. May be memory intensive for large population sizes.

Record statistics (died)

Record a 2D grid of number of fatalities at each time step that be used to generate a video using geoSIR_movie_maker script. May be memory intensive for large population sizes.

Record statistics (recovered)

Record a 2D grid of number of recovered individuals at each time step that be used to generate a video using geoSIR_movie_maker script. May be memory intensive for large population sizes.

Record statistics grid size

Size of the grid (in km square) that the 2D statistics are recorded on. The statistics will be more reliable for large areas, but resolution of the outputs will be lower.

Record statistics time interval

Frequency of 2D grid recording in days. For a value of 1 the 2D grid will be saved every day, for a value of 2 the 2D grid will be saved every other day, etc.

Mimimum latitude

Minimum latitude of the study region in degrees (i.e. 55.500 rather than 55,30'00"N)

Maximum latitude

Maximum latitude of the study region in degrees (i.e. 55.500 rather than 55,30'00"N)

Mimimum longitude

Minimum longitude of the study region in degrees (i.e. 55.500 rather than 55,30'00"W)

Maximum longitude

Maximum longitude of the study region in degrees (i.e. 55.500 rather than 55,30'00"W)

Minimum population density

Minimum ratio of susceptible to (susceptible + empty space) in the model grid.

Maximum population density

Maximum ratio of susceptible to (susceptible + empty space) in the model grid.

Path to population

Path to the gridded population data on your local machine.

Path to red

Path to the high hazard zone data on your local machine.

Path to yellow

Path to the moderate hazard zone data on your local machine.

----

4 - Examples

The script in the code file may be run using the input files from the theoretical natural disaster case, or from the Campania case.

4.1 Theoretical case

Please select one of the input files from the "Theoretical region inputs". The file "Case 2 lockdown" is a good starting point, as it models the spread of a disease in the absence of any natural disaster.

Combine this file in a folder with the other files from the "code" folder. Make sure that the geoSIR_run_theoretical run file is included, as this is necessary to generate the theoretical region.

Open up the wrapper file geoSIR_wrapper, and modify the name of the input and run files as requires.

Finally, open up the inputs file and make any changes to the input parameters. For an initial test it may be preferable to reduce the population to 10,000 (run time should be at most a few seconds), set the number of iterations to 1 and turn off parralelisation. If this is succesful, try increasing the population to 100,000 and also conducting one single run. If either test fails, double check the file paths and any error messages provided by MATLAB.

4.2 Campania case

Please select one of the input files from the "Campania input files" folder. The files "V2_lockdown/wV2_1" and "V2_lockdown/V2_1" are a good starting point, as they models the spread of a disease in the absence of any natural disaster.

Combine this file in a folder with the other files from the "code" folder. Make sure that the geoSIR_run run file and population_density_generate file are included, as they are necessary to load real world data. This real world data for Campania may be found in the "Further data" file in the supplemetary data.

Open up the wrapper file geoSIR_wrapper, and modify the name of the input and run files as requires.

Finally, open up the inputs file and make any changes to the input parameters. For an initial test, it may be preferable to reduce the population to 10,000 (run time should be at most a few seconds), set the number of iterations to 1 and turn off parralelisation. Make sure to also modify the file paths at the end of the input file (

Path to population, etc) so that the model can find the real world population and hazard data. If this is succesful, try increasing the population to 100,000 and also conducting one single run. If either test fails, double check the file paths and any error messages provided by MATLAB. Do not conduct initial tests with a population of 5.8 million, as the run will take more than an hour on even high end computers. Note that large population sizes may also be memory intensive- a single 5.8 million population run will likely require around 10GB of RAM. 100,000 individual runs have very low requirements and should be possible on any modern computer, even when run in parralel.

----

5 - Licensing and other issues

geoSIR code is shared under a GNU public license. This means that it may be freely used and shared so long as the initial authors are acknowledged. Any derivative products cannot be given more restrictive licenses.

In its current state, geoSIR requires a MATLAB license in order to be run. In addition, a MATLAB parralel computing toolbox license is required in order for parralelisation to be used. If a large amount of potential users are unable to access the code due to a lack of MATLAB license, we may create an independant precompiled binary app of the model with no licensing burden.

Please contact the corresponding author at vanwy048@umn.edu with any relevant comments or questions, repeated questiong will be built into a FAQ page.