An R-Based Framework for Implementing Large-Scale Spatial Models of Infectious Diseases

Martina Curran
National University of Ireland Galway
University Road
Galway
m.curran8@nuigalway.ie

Dr. Enda Howley
National University of Ireland Galway
University Road
Galway
ehowley@nuigalway.ie

Dr. Jim Duggan
National University of Ireland Galway
University Road
Galway
jim.duggan@nuigalway.ie

ABSTRACT
Seasonal Influenza is a problem globally, with much research being carried out to find the best ways of preventing epidemics. Early detection is crucial in prevention, with mathematical models being used to allow a better understanding of infection rates, for a quicker response time. The programming language R is a powerful tool which can be used to develop these models, and creates the opportunity to create spatial models which will give a better understanding of different infection scenarios. The case study carried out highlights the importance of creating these disaggregated models rather than the commonly used aggregated models which look at the spread of infections at a high-level, within a single population.

Categories and Subject Descriptors
G.3 [**Mathematics of Computing**]: Probability of Statistics – *Statistical Software.* I.6.5 [**Simulation and Modelling**]: Model Development – *Modelling Methodologies.* J.3 [**Computer Applications**]: Life and Medical Sciences – *Health*.

General Terms
Algorithms, Design.

Keywords
Computational Epidemiology, Analysis, Disaggregated Model.

1. INTRODUCTION
Worldwide, seasonal Influenza is estimated to result in about 3 to 5 million cases of severe illness, about 250,000 to 500,000 deaths [13], and is an important public health problem in the industrialized world [6]. Modelling infectious diseases is an essential tool when faced with epidemics, as early detection is crucial in order to predict infection dynamics. Often used as early warning systems, models can aid in predicting the rate an infection, how it may spread, and give indications on how to proceed in the policy space. Being able to accurately predict the peak of the infectious curve, makes decisions easier for choosing the correct actions to take in order to minimize or prevent an epidemic from taking hold.

The current challenge in the epidemiology field, is fast response times: trying to tell when and where to best direct attention, and actions, before the spread becomes too large to manage. These models can help in making decisions on whether or not herd immunity (the percentage of the population needed to be immune in order to stop an infection becoming an epidemic) is possible at any stage of an epidemic, and if so, at what level. This creates a need for a lower level, detailed mathematical model to measure the spread of infections, by breaking a population down into much smaller areas, to ensure a more accurate prediction, which will allow for faster response times, and more directed interventions.

R is a language and environment which is used for statistical computing and graphics [9]. With many libraries and packages designed for these types of computations, it is a strong choice for this type of work. We created a model that can be used as a decision support system using R, which epidemiologists can use as a framework in order to model the spread of infections at a spatial level, which is scalable for any size and geography of population.

2. FRAMEWORK
The SEIR (Susceptible, Exposed, Infectious, Recovered) compartmental model is commonly used to measure the spread of infections in a population. It allows epidemiologists and public health analysts to visualise the spread of infectious diseases throughout a population, and see the impact of preventive measures like vaccinations or social distancing measures [10][11]. The model compartmentalises a population into four compartments (stocks), and tracks the movement of each individual through the model. These models are often used to look at a population in a country or area as a whole, and work by using many different factors including a recovery delay (the length of time taken to recover from infection), and the contact rate (the length of time needed for two people to be in sufficient contact in order for the infection to spread).

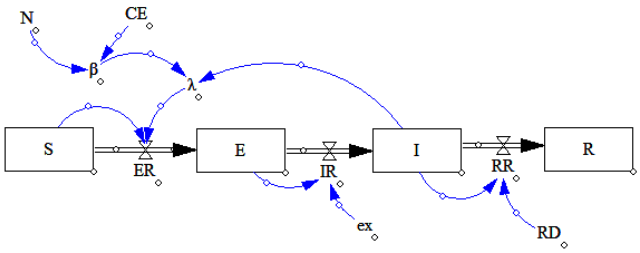

Figure 1. Diagram of SEIR model

The variables named below are updated during each iteration of the simulation, and allow the measuring of the differences in the stocks at each time step, along with the differential equations dS/dt, dE/dt, dI/dt and dR/dt required for these models. This gives a visual representation of how an infection can spread throughout a population, and when the infectiousness of the population will peak through the infection curve.

Constants:
CE: Contact Rate
N: total population
β: CE/N
RD: Recovery Delay
ex: exposure rate

Variables:
λ: β*I
ER: λ*S
IR: E*ex
RR: I/RD

$$\frac{dS}{dt} = -\lambda * S \qquad \frac{dE}{dt} = \lambda * S - E * ex$$

$$\frac{dI}{dt} = E * ex - \frac{I}{RD} \qquad \frac{dR}{dt} = \frac{I}{RD}$$

2.1 SEIR Methodology

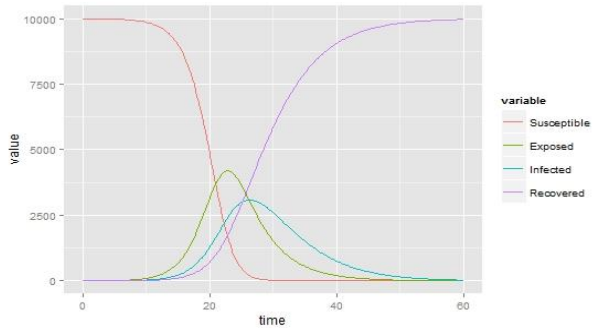

Figure 2. Graph of SEIR model for single population

The rate at which an infection is measured is heavily reliant on Lambda (λ), the force of infection. This value is reliant on Beta (β), which is the effective contact rate needed for the infection to spread. Beta is calculated as the Contact Rate (CE) divided by the total population (N). In order to calculate the force of infection, Beta is multiplied by the current number of infectious people in the population (I). This model gives a detailed representation of how infections spread within a population, however epidemics are large-scale social phenomena, involving populations of hundreds of millions of people, big geographical areas, and complex networks [12], and this makes it difficult to predict epidemics at a spatial level by only looking at a population as a whole.

2.2 Disaggregated SEIR Model

For highly-transmittable infectious diseases such as influenza, the travelling patterns of individuals play an essential role in geographical spread [2]. In order to see the geographical spread of an infection through a population, we break it down into sub-populations, which allows us to see how the factors from the single population model impact on each other in a continuous manner. This creates a need for alterations to the algorithm: as dealing with single figures is no longer possible, we now need to store multiple values in each stock, using vectors.

In R, the 'deSolve' library has an ordinary differential equations function 'ode', which does the differential equations needed to use the model. For the single SEIR model, this function expects a vector of length four, with one value for each stock. Once in the function, the values passed in through the vector are used for the calculations on the variables. These are then passed back in through the vector again, and added to the previous values to measure the difference for the next iteration or time step.

The Contact Rates for the model now need to measure the differences of contact in each constituency, and between each other, creating the need for a matrix that holds the Contact Rates between each constituency. This adds complexity as Beta also becomes a matrix, as each value of CE is divided by its relevant population numbers. In order to calculate Lambda, each relevant Beta value is now multiplied by its infectious population, and is summed to create a vector of Lambda values.

$$\beta = \frac{CEin}{Ni}$$

$$\begin{bmatrix} \lambda i(t) \\ \vdots \\ \lambda n(t) \end{bmatrix} = \begin{bmatrix} \beta ii & \dots & \beta in \\ \vdots & \dots & \vdots \\ \beta ni & \dots & \beta nn \end{bmatrix} \begin{bmatrix} Ii(t) \\ \vdots \\ In(t) \end{bmatrix}$$

As the stock and Lambda values are vectorised when modelling sub-populations, the power of R becomes apparent: having functions which make it possible to do calculations on these vectors without the need for loops which can be expensive in computational terms. This makes it more efficient, and easier to manage.

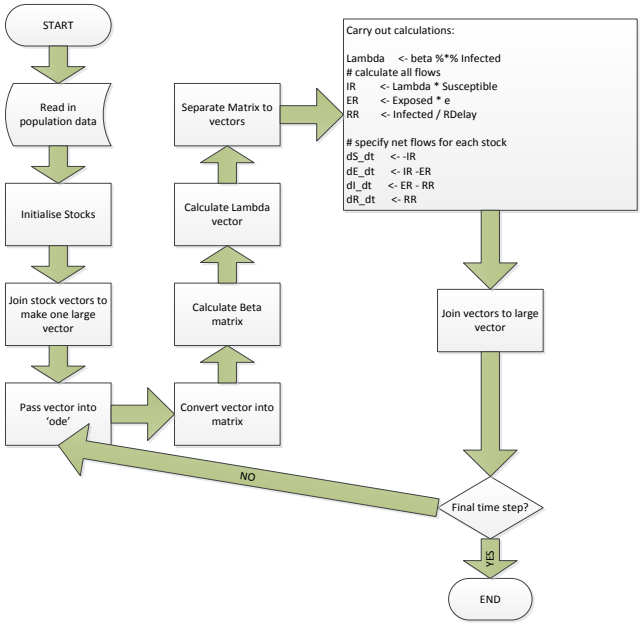

Figure 3. Flowchart of SEIR Model

```
# This is the callback function for ode
deriv <- function(t, state, p){
  with(as.list(c(state, p)),{

    # convert all state variables to an areas(rows)
    # by stocks (cols) matrix
    states<-matrix(state,
                   nrow=num_areas,
                   ncol=num_stocks)

    # extract required state vectors for flow calculations
    Susceptible <- states[,1]
    Exposed     <- states[,2]
    Infected    <- states[,3]
    Recovered   <- states[,4]

    # matrix operation to calculate lambda values
    Lambda      <- beta %*% Infected
    # calculate all flows
    IR          <- Lambda * Susceptible
    ER          <- Exposed * e
    RR          <- Infected / RDelay

    # specify net flows for each stock
    dS_dt       <- -IR
    dE_dt       <- IR -ER
    dI_dt       <- ER - RR
    dR_dt       <- RR
    return (list(c(dS_dt,dE_dt,dI_dt,dR_dt)))
  })
}
```

Figure 4. 'deriv' Function for Iterative Calculations

3. CASE STUDY

In Ireland there are 40 electoral constituencies, and using Ireland's national census data, it is possible to create an exploratory large-scale, disaggregated model.

For this vectorised model to track each constituency, each of these stocks for the 'ode' function need to store the values for each individual constituency, meaning 160 values are being manipulated at each iteration, adding to its complexity. Due to the nature of the vectorised model however, we only need to read in shape file data for population numbers, and the Contact Rate matrix, which now has 1600 values.

This disaggregated model now allows visualisation of the change in stocks for any number of sub-populations, with a clearer idea of the infections in each single population, and the peak of the individual infection curves, due to multiple Contact Rates and Beta values.

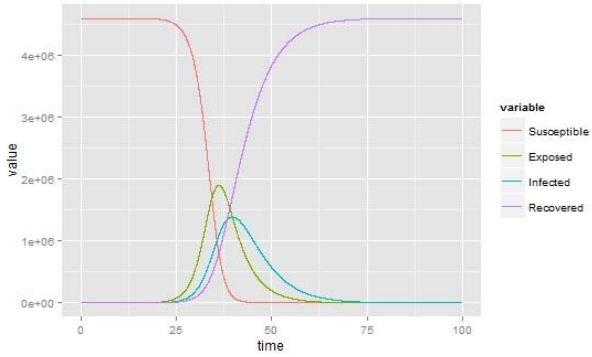

Figure 5. SEIR Model for Irish Population as a whole

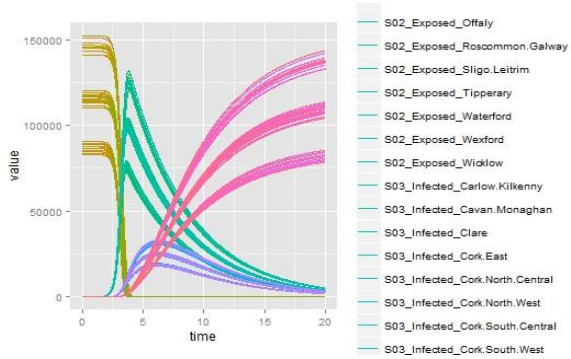

Figure 6. Disaggregated SEIR Model for Irish Constituencies

However, with increasing population numbers, it becomes difficult to visually separate each constituency from another in a graph. Using the shape files however, also allows the retrieval of coordinate data from Census, and using the package 'maptools', gives a method of visually mapping the spread not only within each electoral constituency, but also between the constituencies at the same time.

By adding a function to plot the map on each iteration, it is possible to easily visualise the spread of an infection of the country through each constituency. It also allows to see the difference of the rate or location of spread, depending where the infection originated from.

4. RELATED WORK

SEIR models are used to monitor and predict the spread of many infectious diseases, including different variations of Influenza e.g. Swine Flu [2], and other infectious diseases like Ebola [1][7]. Research carried out on these diseases include breaking populations into two subpopulations: health workers and non-health workers [7], and looking the impact of travel restrictions in order to delay a pandemic to give time for vaccinations to be available [2].

This research brings these ideas together, by breaking one population into multiple subpopulations, while also looking at the impact of contact between each subpopulation. This created the need for the Contact Rate matrix, which stores the information of how these subpopulations affect themselves and each other. The framework can be used for any infectious diseases which be modelled using SEIR models, by reading in population details and Contact Rates, while needing only minor changes in details around the infectiousness i.e. Recovery Delays, and Exposure Rates.

5. RESULTS

This research allowed the development of a spatial model which reads in real population numbers using shape files, allowing the visualisation of an infection spread using maps. Written in R, it allows epidemiologists and people in the health industry with some technical knowledge to monitor the spread of an infection, and track the differences of each stage of infection visually.

By creating a vectorised, scalable SEIR model, and using it with real data, it will be possible to predict the rise or fall of infections in each sub-population in an area. It shows the spread of infection to different subpopulations, and the differences in the infection dynamics depending where the infection originated from. In the event of an epidemic, epidemiologists will now have a framework on which to predict the spread of an infection at a lower level. This will impact on the way vaccinations can be distributed, in the event of a limited supply, as well as how other mitigation factors can affect the infection spread.

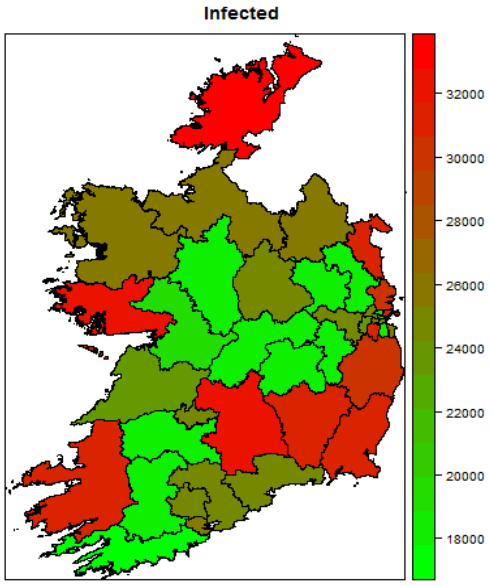

Figure 7. Number of Infectious People in each Constituency

Using shape files with the model allows it to be used for any number of populations, and any country or continent. It can be broken down into different spatial cohorts, depending on the shape files available. This allows a simplified use of the model, as all details regarding the populations (population names, population numbers, and area) are often included in shape files.

6. FUTURE WORK

This model was created using a static infectiousness, and an estimated matrix of contact rates for calculations. Further research will use real-time data gathered from the Irish participatory surveillance project FluSurvey [4]. Flusurvey, in collaboration with the Irish health service (HSE), is part of Influenzanet, a European-wide consortium aimed at introducing an innovative information and communication technology approach for a web-based surveillance system across different European countries [8]. Other participatory, community-based syndromic surveillance systems have also been introduced in Australia and in the United States to potentially address the limitations of existing healthcare based surveillance in a complimentary way [3]. Real-time data from Flusurvey.ie will be used to continually update any changes to the infection. Research will also be carried out to best detect contact rates for the prediction of influenza infection spread.

To the best of our knowledge, nothing like this has been done in R before, and although similar in idea to GleamViz [5], it is unique in that it uses contact rates rather than transport data to measure the differences in infection scenarios between sub-populations.

7. CONCLUSION

The aggregated SEIR model was implemented in R, and then adapted for scalability to allow for measuring infectious disease spread as closely as possible. Using a dynamic vector, while dynamically retrieving the names from the imported shape files, makes it possible to easily model the stocks for any number of populations, and allows reproducible use for industry impact. During each iteration of the model, it unpacks the data for calculations during the simulation, and repacks it for use in the next iteration, making the architecture slightly different from other models. Visually modelling the spread through maps during each time step makes it easier to predict high peaks of infection. Further research will use real-time data for up-to-date influenza infections, gathered from the Irish participatory surveillance project FluSurvey, which will make it possible to better predict these increases.

8. REFERENCES

[1] Chen, T., Ka-Kit Leung, R., Liu, R., et al. 2014. Risk of imported Ebola virus disease in China. *Travel Medicine and Infectious Disease.* 12, 6 (Nov, 2014), 650-658. DOI= 10.1016/j.tmaid.2014.10.015.

[2] Chong, K.C., Ying Zee, B.C. 2012. Modeling the impact of air, sea, and land travel restrictions supplemented by other interventions on the emergence of a new influenza pandemic virus. *BMC Infectious Diseases.* 12, (Nov, 2012), 309. DOI= 10.1186/1471-2334-12-309.

[3] Chunara, R., Goldstein, E., Patterson-Lomba, O., et al. 2015. Estimating influenza attack rates in the United States using a participatory cohort. *Scientific Reports.* 5 (April, 2015). DOI= 10.1038/srep09540.

[4] Flusurvey. 2015. https://flusurvey.ie/en/.

[5] GleamViz. Model. 2015. http://www.gleamviz.org/model/.

[6] Marquet, R., Bartelds, A., van Noort, S., et al. 2006. Internet-based monitoring of influenza-like-illness (ILI) in the general population of the Netherlands during the 2003-2004 influenza season. *BMC Public Health. 6, 1 (Oct, 2006), 242*. DOI= 10.1186/1471-2458-6-242.

[7] Okeke, I., Manning, R.S., Pfeiffer, T. 2014. Diagnostic schemes for reducing epidemic size of African viral haemorrhagic fever outbreaks. *The Journal of Infection in Developing Countries.* 8, 9 (Sept, 2014), 1148-1159. DOI= 10.3855/jidc.4636.

[8] Paolotti, D., Carnahan, A., Colizza. V., et al. 2013. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. *Clinical Microbiology and Infection.* 20, 1 (Nov, 2013), 17-21. DOI= 10.1111/1469-0691.12477.

[9] R. What is R? http://www.r-project.org/.

[10] Rizzo, C., Lunelli, A., Pugliese, A., et al. 2008. Scenarios of diffusion and control of an influenza pandemic in Italy. *Epidemiology and Infection*. 136, 12 (Dec, 2008), 1650-1657. DOI= 10.1017/S095026880800037X.

[11] Shulgin, B., Stone, L., Z. Agur. 2008. Pulse vaccination strategy in the SIR epidemic model. *Bulletin of Mathematical Biology*. 60, 6 (Nov, 1998), 1123-1148. DOI= 10.1006/S0092-8240(98)90005-2.

[12] Swarup, S., Eubank, S. G., Marathe. M. V. 2014. Computational epidemiology as a challenge domain for multiagent systems. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* (AAMAS '14). 1173-1176. http://aamas2014.lip6.fr/proceedings/aamas/p1173. pdf.

[13] WHO. Influenza (Seasonal) fact sheet. 2014. http://www.who.int/mediacentre/factsheets/fs211/en/.