|epcc|

THE UNIVERSITY OF EDINBURGH

## EPCC-SS97-08

# Spatial Modelling of Epidemics on High Performance Computers

**Jennifer Market**

**Abstract**

The spatial modelling of population evolution, particularly the dynamics of epidemics, is ideally suited for high performance computing. We have developed the stochastic SEIR model to account for spatial aspects, enabling us to simulate large numbers of interacting sub-populations. Using the CRAY T3D and the Meiko CS-2, we are able to study the evolution of such diseases as measles, over long periods (30-50 years), for many (10,000+) interacting sub-populations. The size of the sub populations varies between 1000 and 1,500,000, and the effect of population size on epidemic dynamics is examined.

# Contents

# 1 Introduction

Epidemic processes are a familiar feature in biology. In addition to characterising the spread of diseases, they also describe such diverse processes as fire spread, starfish outbreaks, food chain dynamics, and the invasion of exotic plants. A common feature to all epidemic processes is the point to point spread: one sick person infects the next, one burning tree ignites the next, water currents carrying starfish larvae from one reef to the next, and seeds spread from one site to another. [9] The SEIR (Susceptible, Exposed, Infectious, Recovered) model is generally accepted for characterisation of all these processes. This model has deterministic and stochastic variants, both of which will be discussed below. As well, as isolated populations, this model can be used to study interacting sub-populations, through the use of a coupling constant which depends upon the number of infectious subjects migrating between populations.[2]

In the past, attempts to use this method to model small numbers of interacting sub-populations have led to artificial features, such as synchronisation between the populations and the eradication of the disease.

A previous SSP project used larger numbers of subpopulations (i.e.1024) with variable coupling constants and population sizes to attain persistence of the disease. Both the deterministic and stochastic models were tested, with better results being obtained from the latter.[2]

We make use of both the Cray T3D and the Meiko CS-2 to model yet larger numbers of sub-populations (tens of thousands), which can interact on either a nearest-neighbour, or weighted global interaction model. We also attempted to make the code portable across platforms by using a combination of C and MPI.

Although the ideas in this report are applicable to many other diseases we will mainly concentrate on measles. This disease has been well monitored for the last century, with copious and complete data available for whole nations on a town-by-town basis. Measles, like most childhood diseases, spreads very quickly through airborne contact and the cycle is very short, i.e. days instead of years. This allows us to use short time steps (as short as five minutes) and evolve the data over many years, comparing the results to those actually witnessed.

# 2 Background

## 2.1 Childhood Diseases

Childhood diseases are the most common form of infectious diseases. These are diseases such as measles, mumps, chicken pox and such, to which children are born susceptible and usually contract within five years. Because young children are in particularly close contact with their peers, at school and play, the disease can spread quite quickly. The process follows a regular course of exposure, then an incubation period where the person is infected but does not yet show the symptoms, followed by the symptoms themselves. Between these last two steps, is the latency period, during which the individual is infectious to others. This varies from one disease to another. Finally the individual recovers and in most cases, becomes immune to the disease.[4]

## 2.2 SEIR Model

SEIR models are a class of compartmental prevalence models, with classes Susceptible, Exposed(Latents), Infectious, and Recovered. [4] This is perhaps the most generally accepted and most commonly used model, having several permutations and variants. The populations can be divided into four categories:

**Susceptible** - those capable of contracting the disease

**Exposed** - those who are infected but not yet infectious

**Infectious** - those capable of spreading the disease

**Recovered** - those immune from the disease

For the simplest version of the model we will use, we will make several assumptions.

• Birth and death rates are equal, leaving the population size constant.

• Recovered individuals to be truly immune (in reality, for measles, this is not always true.)

• The disease will have no, or negligible effect on the mortality rate.

• The mean latency period to have a duration $1/a$.

• For any individual, the probability of remaining exposed at time $\tau$ after the point of infection is $e^{a\tau}$.

• The probability density of recovering at time $\tau$ after becoming infectious is $e^{-g\tau}$.

• $1/g$, the effective infectious period, is the difference between the average incubation period and the average latency period.

For time dependent contact factor $b$, there exist two stationary states of the SEIR equations: (S,E,I)=(1,0,0) for extinction and ($S_0$,$E_0$,$I_0$) for endemic equilibrium, in cases such as measles. Another factor of interest is the reproductive rate $1/S_0$, the number of secondary cases one individual is capable of producing.[1]

The SEIR model is best visualised in figure 1.

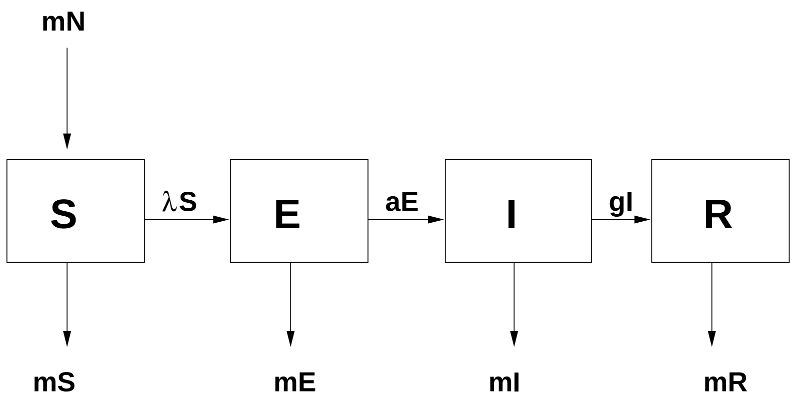


Figure 1: SEIR mode

The parameters for the model are:

•**m** - birth and death rates

•**a** - inverse latency period

•**g** - inverse of the difference between the incubation period and the latency period

•$\lambda$ - $bI/N$ , where b is the average number of infections caused by any one person,

•**N = S + E + I + R**

•$b(t) = b_0(1 + b_1 cos(2\pi t))$ -contact parameter with seasonal variance, $b_1$

### 2.3 Stochastic vs Deterministic

While our implementation uses the stochastic model, it is worthwhile to briefly detail the differences between the two variations of the SEIR model. By definition, the deterministic model is one in which the parameters and variables are not subject to random fluctuations, so that the system at any time is defined entirely by the initial conditions chosen. In contrast to this the stochastic model considers the presence of randomness in one or more of its parameters or variables. The predictions of the model therefore do not give a single point estimate but a probability distribution of possible estimates.

In the deterministic model, the system is represented by four first order differential equations:

$$\begin{aligned} \frac{dS}{dt} &= mN - (M + \lambda)S \\ \frac{dE}{dt} &= \lambda S - (m + a)E \\ \frac{dI}{dt} &= aE - (m + \lambda)I \\ \frac{dR}{dt} &= gI - mR \end{aligned} \tag{1}$$

where N=S+E+I+R. The difficulties with the deterministic model include the 'fractional people' problem. Here the disease does not die out as part of one person has the disease, which is hardly realistic.

The stochastic model deals with probabilities. We take the parameters in the above differential equations as the probabilities for various events. According to the model, there are nine possible events, represented by the arrows in figure 1. Depending upon the even, the SEIR values are changed as described in table 1.

Thus the total probability is

$$W_n^t = m(N + S + E + I + R) + aE + gI + \lambda S$$

At time $t_0$ the probability of an individual being in state n is 1, so $p_n$ = 1 and $q_n$ = 0 for all other n, $n_q \neq n_p$. At any time $t > t_0$, $p_n$ will decrease according to $\frac{dp_n}{dt} = -W_n^t p_n$. Integration gives

$$p_n(t) = \exp(-\int_{t_0}^{t} W_n^\tau d\tau) \tag{2}$$

| Event | Probability $W_n$ | S | E | I | R |
|---|---|---|---|---|---|
| Birth | mN | +1 | 0 | 0 | 0 |
| Death | mS | -1 | 0 | 0 | 0 |
| Death | mE | 0 | -1 | 0 | 0 |
| Death | mI | 0 | 0 | -1 | 0 |
| Death | mR | 0 | 0 | 0 | -1 |
| Infection | $\lambda$S | -1 | +1 | 0 | 0 |
| Becoming Infectious | aE | 0 | -1 | +1 | 0 |
| Recovery | gI | 0 | 0 | -1 | +1 |

Table 1: Stochastic method of determining SEIR values

If $W_n^\tau = constant$, the integrals are easily solved. Thus we have $p_n(t) = \exp(-W_n^t \Delta t)$ where $\Delta t = (t - t_0)$. If the time steps are sufficiently small, the assumption of constant $W_n^\tau$ is justified, and we can define a probability density for a transition from state $n_p$ to $n_q$:

$$Pr_t(\Delta t) = W_n^t \exp(-W_n^t \Delta t) \tag{3}$$

One issue with the stochastic model, in particular, is the development of herd immunity. Even though there are susceptibles within the populations the infected people are so well surrounded by immune non-carriers, that the disease can stochastically fade away. This can happen immediately after a huge infection has made nearly the entire population immune, and small stochastic fluctuations can eradicate the disease. This is known as Epidemic fadeout.

Another option, easily implemented, is to study the effect of vaccinations upon interacting populations. This could be extremely useful to health service coordinators. This could be done either at any point on the evolution, by simply subtracting the number of vaccinations from the Susceptibles list and adding them to the Recovered quantity. One could thus demonstrate the effectiveness of vaccinations.

On the assumption that most people -those who do not die along the way- follow the path of birth, infection, recovery, death, we can estimate the number of events in a year using equation[4]

$$timestep = 1/(N * m * 5) \tag{4}$$

For example, for a population of 1,000,000 with a birth/death rate of m=0.02 there are 20,000 births/deaths annually. Thus there are approximately 100,000 transitions in a year. The average time step for a population of 1,000,000 then, is approximately five minutes. For a one year simulation, we would need $365 * 24 * 60/5 = 105,120$ For large populations, then, parallelisation becomes essential. The use of exponentially distributed time steps is more efficient than using fixed time steps. The variable time steps can be calculated as

$$\Delta t = -\frac{1}{W_n^t} \ln(1 - r) \tag{5}$$

When considering parallel techniques, however, all the subpopulation would have to be synchronised after each time step, or group of time steps, so we really gain little, and equally incremented time steps are much simpler. When using variable population sizes in the simulation, it is essential also to remember that whereas all the populations are interacting at each time step, that time step takes on different temporal values within each population. For example, the same time step which corresponds to 5 minutes in a population of 1,000,000 is roughly 5 years for a population of 5,000. A graph of the time steps is shown in figure 2.

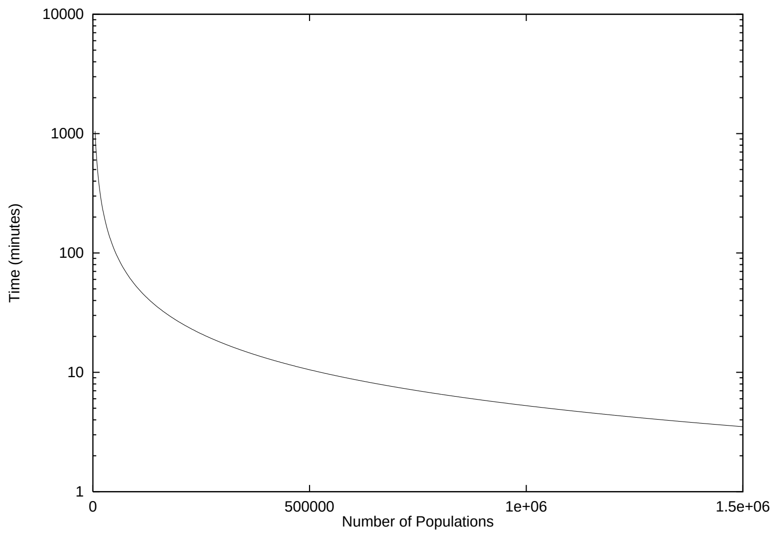


Figure 2: time steps

According to figure 3, we can thus simulate the evolution of the disease.

Infectious diseases often show a seasonal variation, often dealt with as a chaotic effect (see Chaos and Epidemics, below). Especially concerning schoolchildren, who spend the winter months in much closer contact with other children, the probability of infection can be increased dramatically.[1] The parameter which describes this effect has the form

$$b(t) = b_0(1 + b_1 * cos(2\pi t)) \tag{6}$$

## 2.4 Prevalence

Prevalence is defined as the proportion of the population infected at a particular time. One important value to examine is the critical parameter, $V$, associated with an epidemic process. The analogy is often drawn between percolation in porous media and epidemic spread. Imagine water trickling through porous media; if the amount and speed of the water flow are below a critical value single rivulets are sufficient. If the water flow is above the critical value the flow will spread out. Similarly, below a critical value, epidemics will not spread but will be eradicated, while at the critical value they will maintain their numbers. Whereas above the critical value the disease will spread to engulf all the population. $V$ is the probability that each cell will infect its neighbour.[9] In the case of measles, not only do we wish to maintain prevalence of the disease but we also wish to see continued periodicity, cycling due to the seasonal parameter. It is here that chaotic effects can complicate the model significantly.

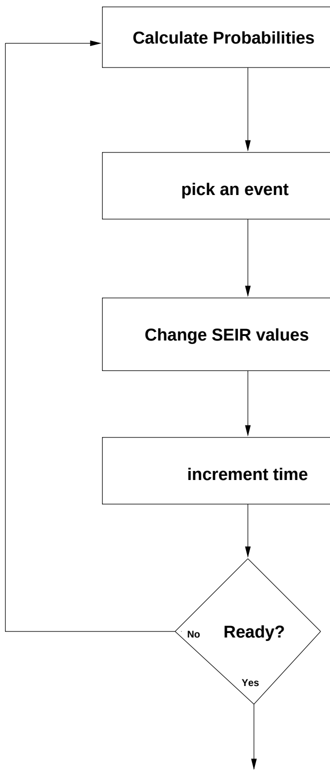


Figure 3: Algorithm Design

## 2.5 Measles

As measles is one of the best documented diseases we have chosen it as an example on which to test our programme. Measles, or Rubeola, is best known for its typical skin rash though it is actually a respiratory infection. The incubation period is 8-13 days with a latency period of 6-9 days. Individuals are then infectious for 6-7 days before becoming immune. [3] Table 2 shows the SEIR parameters for measles.[1]

| Description | Symbol | Value ($yr^{-1}$) |
|---|---|---|
| Birth/Death Rate | m | .02 |
| Inverse Latency Period | a | 35.84 |
| Inverse Infectious Period | g | 100 |
| Contact Parameter | b | 1800 |

Table 2: Measles parameters

## 2.6 Chaos and Epidemics

"Simple nonlinear systems do not necessarily posses simple dynamical properties."[1] While it can be argued that chaos has negligible importance, in a single population's evolution, it is likely to have an effect on nonlinear interacting systems. However there has yet to be undeniable proof that it is indeed chaos rather than environmental or demographic factors which causes the sometimes unpredictable patterns in population evolutions. At any rate endemic diseases like measles are ideal for studying chaotic behaviour.[1]

### 2.7 Spatial Aspects

The above discussion concerned a single population. Of course to develop a realistic model we must consider many interacting, or coupled, subpopulations. We can either concern ourselves only with nearest neighbour interactions (as most migrations are local) or consider a global interaction with the coupling constant weighted by the inverse square of the distance between the populations. More complex still we could use the available data for migrations between subpopulations (such data is available from the Human Systems Modelling Consortium).[6] In the past researchers have used small groups (i.e., ten equally sized, coupled subpopulations).[2] Even with forced seasonality they found a strong coupling between populations that led to the eventual eradication of the disease. They attributed this error to the global mixing of their model. We have chosen to deal only with nearest neighbour interactions. It is also the least communications-intensive and a valid approximation. For any number of neighbours we chose to include, the coupling constant, $\lambda$, is given by

$$\lambda = b(t)(I_i + v\sum_{j\neq i} I_j) \tag{7}$$

where $I_n$ is the number of infectious people in the subpopulation of interest, and $v$ is a coupling coefficient. $b(t)$, the contact parameter, is altered as well to reflect these interactions. With interactions.

$$b(t) = \frac{b_0(1 + b_1 cos(2\pi t))}{(1 + (n-1)v)N} \tag{8}$$

where n here refers to the number of subpopulations in contact with that of interest.[1]

## 3 Implementation

### 3.1 General

The simple serial model was translated from FORTRAN to C in preparation for implementation as a C MPI code on parallel machines.

### 3.2 Initial Conditions

We wish to start with non-uniform population size and SEIR values. We used four instances of the Gaussian random number generator, so that the percentages of Susceptible, Exposed, Infected, and Recovered individuals varied by population. The Gaussian distributions are shown in figure[] . Various distributions, with $sigma$ = .1,.3,.5, and .7 were used to test the prevalence of the disease with different start values. (Discussed in the RESULTS section).

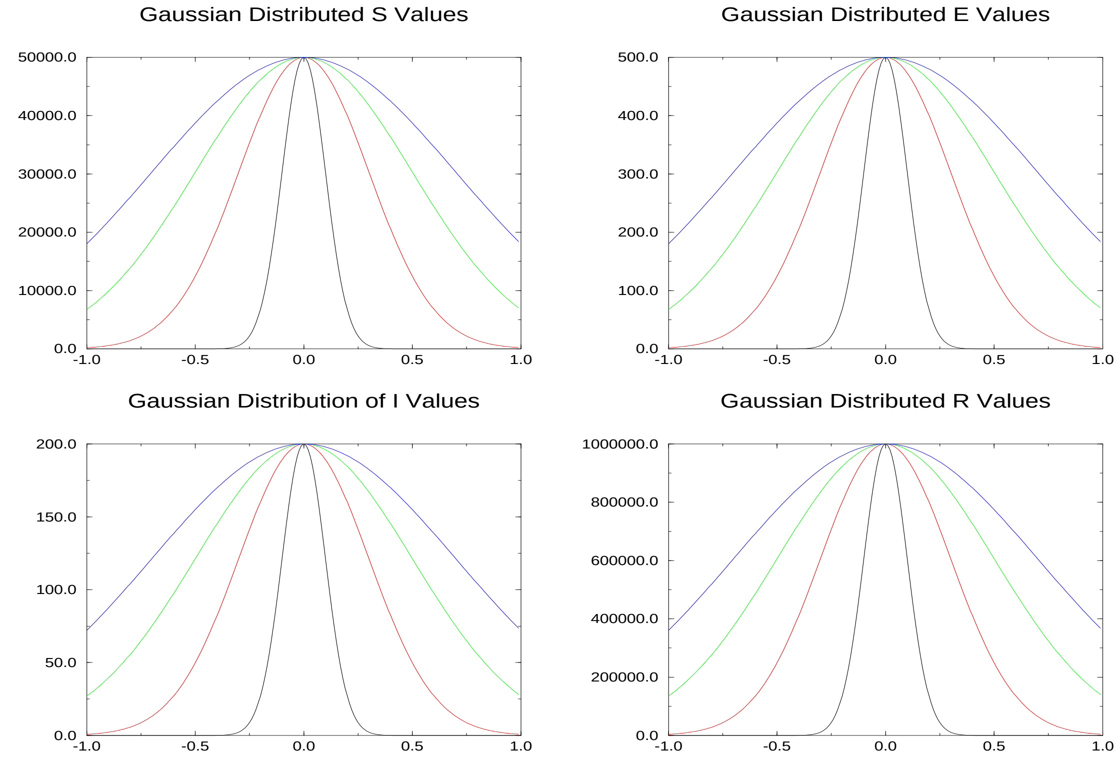


Figure 4: Gaussian Distributions for S,E,I,and R, with $\sigma$ = .1,.3,.5, and .7

## 3.3 The Cray T3D

The Cray T3D at EPCC contains four components - the CRAY J90, the CRAY Y-MP host, the CRAY T3D system itself, and an IBM 3494 mass storage system. The configuration contains 512 PEs, each consisting of a DEC Alpha 21064 processor at 150 MHz, supporting 64-bit integer and 64-bit IEEE floating point operations and delivering 150 64-bit Mflops/s. The DEC Alpha 21604 includes an 8 Kbyte direct-mapped data cache and an 8 Kbyte instruction cache. Each processor has 64 Mbytes of RAM giving an aggregate memory for a 512 PE system of nearly 33 Gbyte. There are two I/O gateways, which must communicate with the Y-MP for all I/O.[7]

## 3.4 Theory of Parallelisation

The data is evenly distributed over the processors in a simple fashion, sending the first n/gsize populations to processor 0, the second n/gsize populations to processor 1, etc., where n is the number of populations and gsize is the number of processors. Within each processor, the data is put, row by column, in a square grid of size sqrt(n/gsize)

* sqrt(n/gsize). This should give as even load-balance as possible, unless the user has some fore-knowledge of a concentration of large populations in one area, such as a large concentration of cities along coastlines, in which case the data should be distributed by hand.

Of course, the above process imposes the limitation that the number of populations must be a perfect square after being divided by the number of processors. In practice, this is not really a limitation, since we can just add 0 person populations to any number to bring the value up to one which satisfies these conditions. Thus for 16 processors the time would be the same for 9217 populations as for 10,000 populations. Each processor needs to know where its surrounding 8 neighbours are located. To facilitate this we placed the processors in a 2 dimensional grid. Thus, we can easily add and subtract to find the nearest neighbours in both the x and y directions.

The most challenging aspect of parallelisation is the halo passing necessary for communications between processors. Since each population will be interacting with its 8 neighbours, those populations on the edges of the sub-blocks will need to communicate with their neighbouring populations, which are located on different processors. We must also consider that, since our population grid is not periodic the edge processors will need fewer halos than the inner ones. To visualise the halo arrangement, see figure 5.

After creating the halos every effort is made to pass them efficiently by reducing the amount of communications involved. Instead of communicating the east, west, north, south, ne, se, nw, sw halos separately, we first exchange the east/west halos, including the empty top and bottom corners. After these exchanges are completed, we then pass the top/bottom halos in a similar fashion. By doing these in this order, we save communications overhead and create a simple, elegant passing routine.[5]

## 4 Results

We encountered a number of difficulties with the large amounts of data we were attempting to communicate, often running into problems with memory storage and execution time for very large numbers of sub-populations (50,000). In the end we have managed to create a code which can handle the data passing efficiently, though not quite as fast as was desired. We ran the code on both the Meiko CS2 and the Cray T3D to compare the results. A graph of the speed-up is shown in figure 6.

Because of the communications overhead, there is little advantage to parallelisation for small populations (64) on many processors (16). In such a case, more communications, i.e., halo passing, is occurring than actual calculations. It is not until we reach bigger populations sizes that it is worthwhile to parallelise the calculations. Figure 7 shows these results.

When we attempted to simulate random sized interacting population, we found that the method for randomising the start values was a bit unrealistic, and it thus took 10-20 years for the populations to settle, after which cyclic seasonal variation is present, but on a rather small scale. See figure for an example. Preliminary results upon investigating

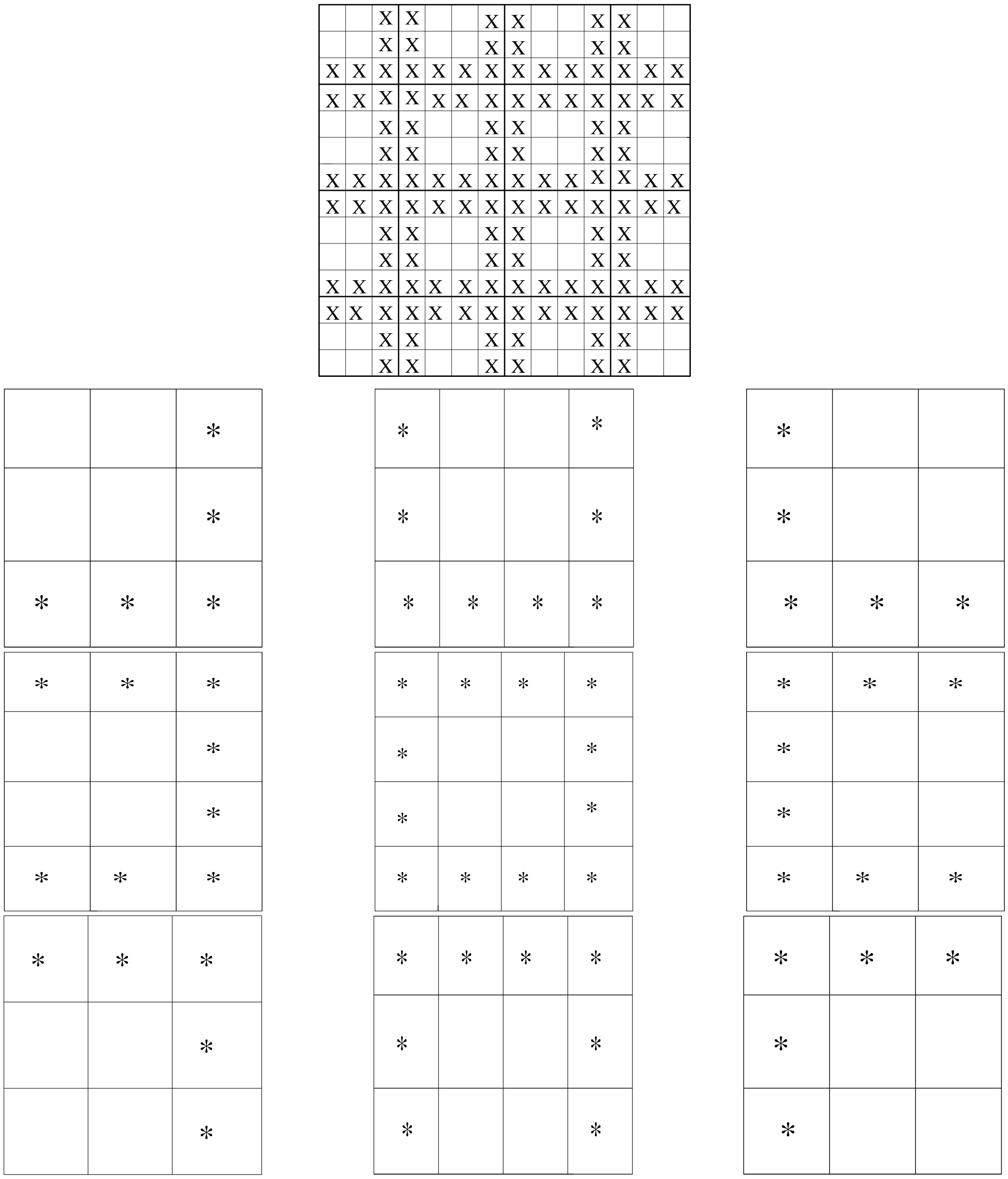

Figure 5: Halo Distribution

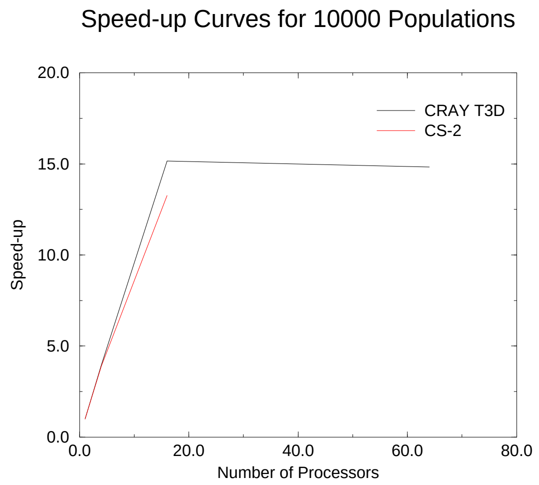


Figure 6: Speed-up time comparisons from the Cray T3D and CS2

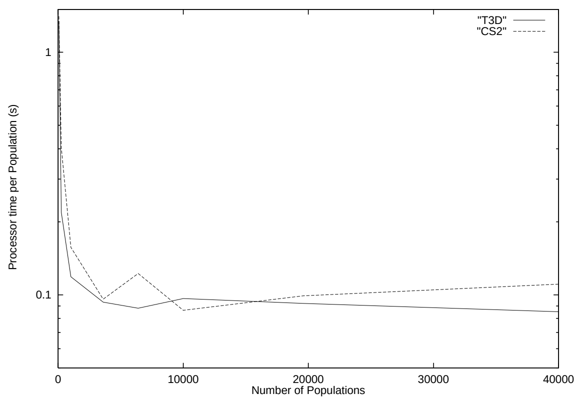


Figure 7: Processing time per population for a one week simulation using 16 processors. The populations are of uniform size 1 million.

this feature indicate that, instead of generating four separate random numbers for the values of S, E, I, and R, we should instead link them in some fashion. For example, we used a range of percentages to relate the four values. These values should be roughly in the range of: $S = 2 - 15\%$, $E < 5\%$, $I < 5\%$, $R > 50\%$.

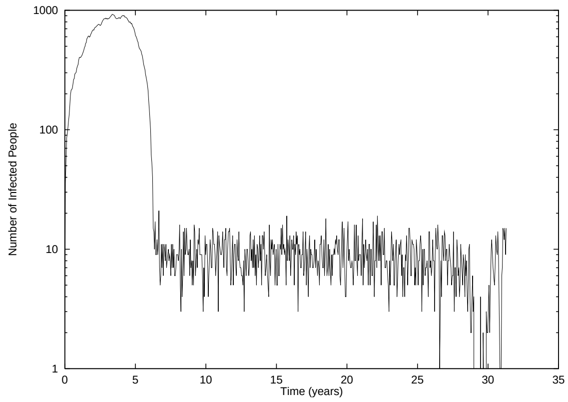


Figure 8: Evolution of a 25,000 Population with highly randomised start values

The coupling constant and population size are linked. For example, for $v = .01$, populations of size 1,000,000 retain their seasonal periodicity for approximately 10 years. For populations of size 25,000, periodicity lasts roughly twice as long. Investigation into this relationship is vital when considering interacting varied size populations, as the coupling constant, $v$, used in the calculation must vary with population size.

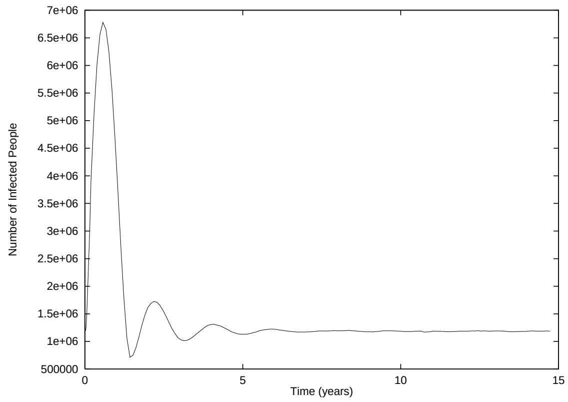


Figure 9: Evolution of total number of infected people, for 10,000 populations of 1 million each

AVS ( Advanced Visual Systems, Inc.) was used to produce visualisation of the percentage of infected people in populations. The results are shown if figures 11 and 12.

## 5 Conclusion

Spatial modelling of epidemics, for large numbers of populations is suited for data parallel programming. Beginning with a serial version of the SEIR model we have parallelised it for tens of thousands of interacting populations, using C/MPI.

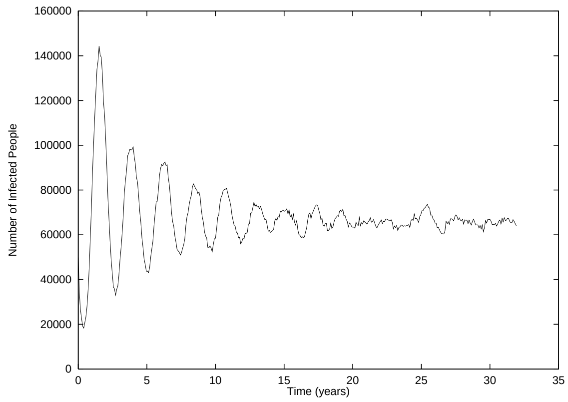


Figure 10: Evolution of total number of infected people, for 10,000 populations of 25,000 each

Disease persistence was readily obtained both for thousands of uniform-sized populations and for random start values. Retaining the periodicity of the disease by varying such parameters as $v$, the coupling constant, and $b_1$, the seasonal parameter are still under investigation. The best values for these parameters have yet to be determined, due to the constraints of time.

Both nearest-neighbour and global interaction versions of the code were produced, with the majority of the time spent optimising the former. In the future, it would be interesting to use actual migration data and, indeed, real health department data to test the code and compare against reality particularly for answering the many chaos questions. Studies of the effect of vaccinations on the evolution of the disease could be very useful as well.

Additional future work might also include making the code more flexible, so that it runs on any number of nodes, rather than just perfect squares. Determination of the relationship between the coupling constant, $v$, and the population size is essential to the further development of the project.

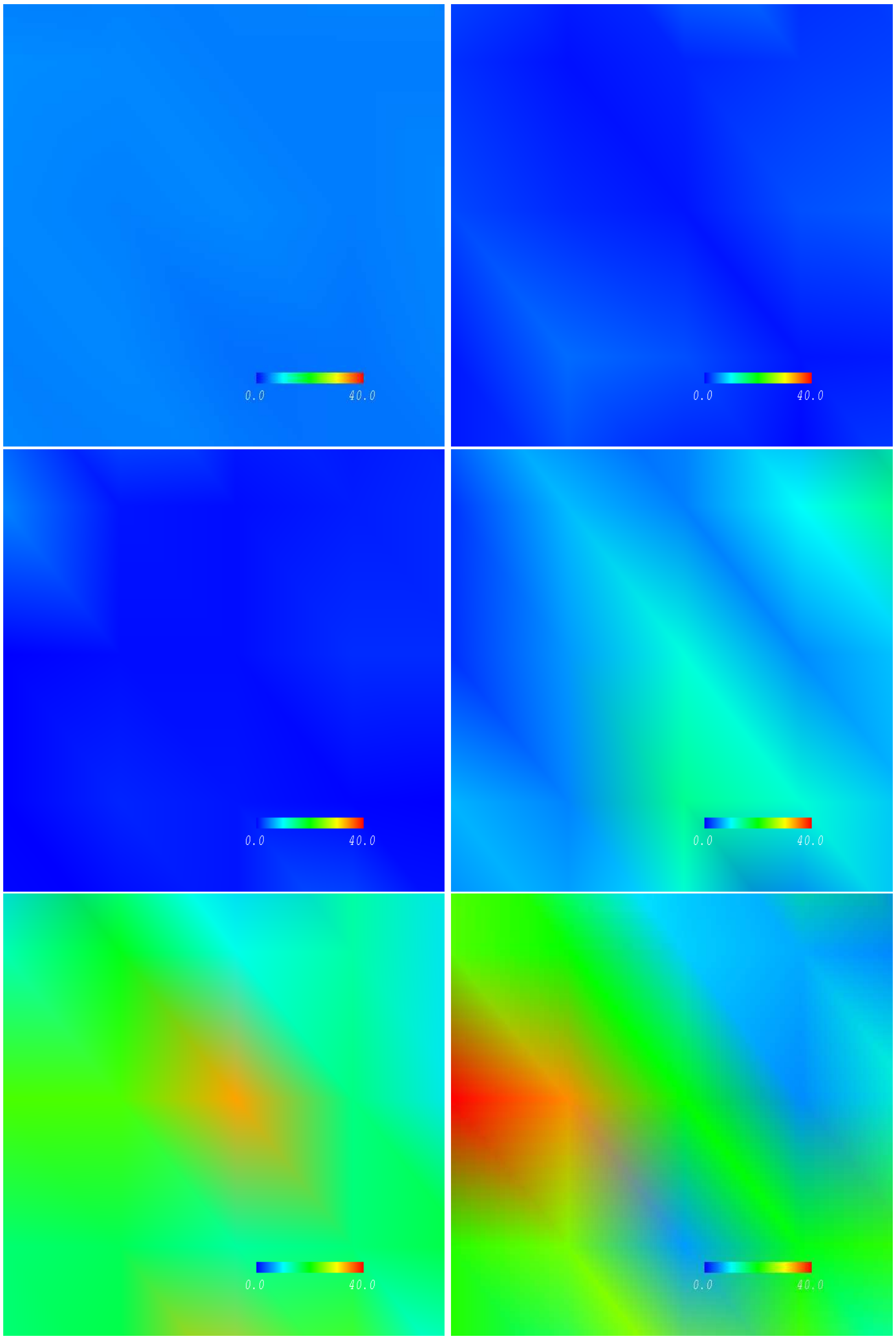


Figure 11: Snap-shots of a 100 x 100 grid at 0,3,6,12,18, and 24 months. The sub-population size is fixed at 25,000, $b_1$ is .25 and v is .01

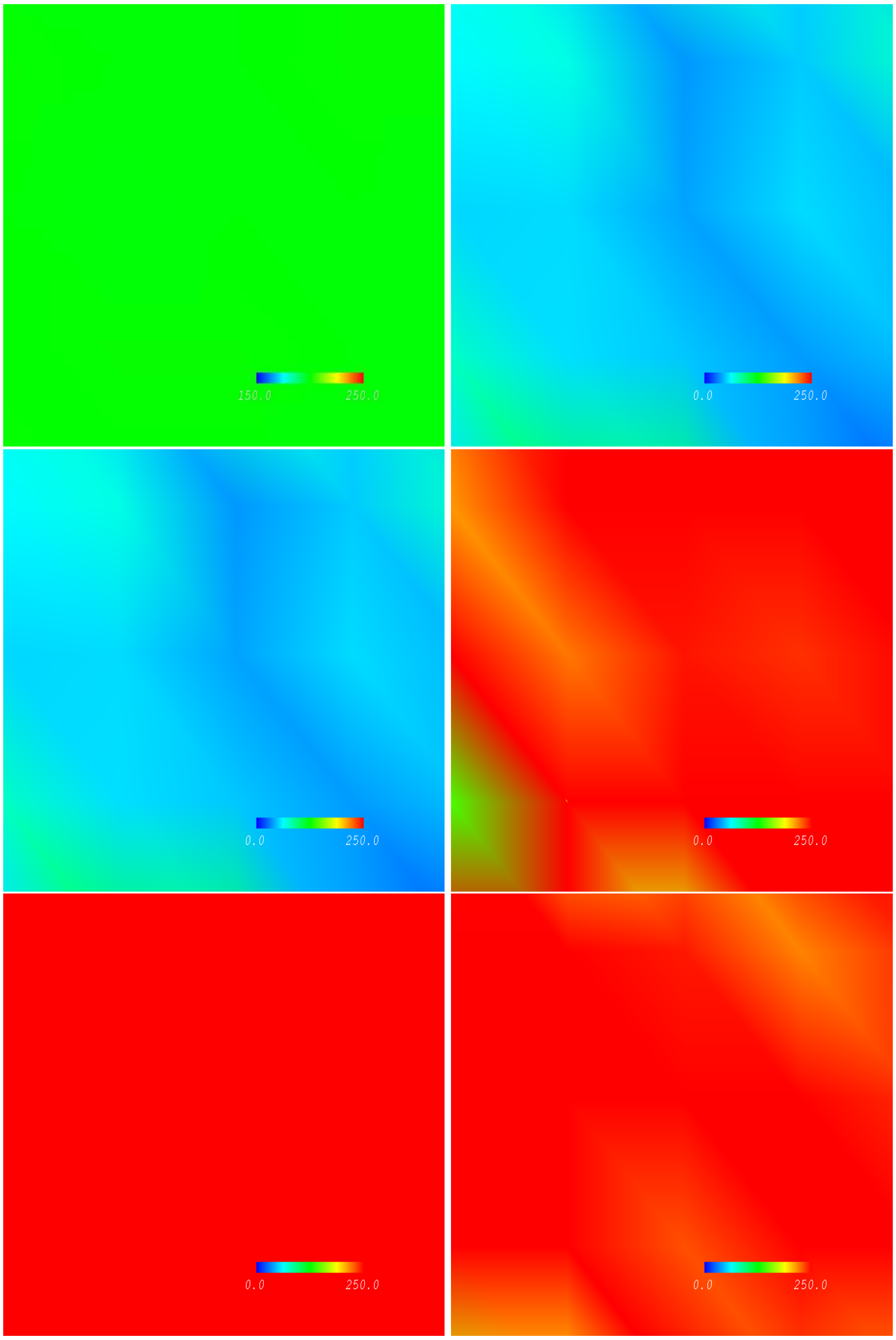


Figure 12: Snap-shots of a 100 x 100 grid at 0,3,6,12,18, and 24 months. The sub-population size is fixed at 1,000,000, $b_1$ is .25 and v is .01

## References


[1] R. Engbert and F.R. Drepper: Chance and Chao in Population Biology - Models in Recurrent Epidemics and Food Chain Dynamics, *Choas, Solitons, and Fractals*, 4(7):1147-1169,1994

[2] E. Frederiksen: *Spatial Aspects of Epidemics*, EPCC-SS95-05, 1995.

[3] `http://www.kidshealth.org`

[4] Oxford Univeristy Dictionary of Epidemology, Centre for the Epidemiology of Infectious Disease.

`http://www.zoo.ox.ac.uk/~js/course-glossary.html.`

[5] Discussions with Mr. Mario Antonioletti.

[6] Human Systems Modelling Consortium.

[7] Stephen Booth, John Fisher, Neil MacDonald, Peter Maccallum, Joel Malard, Alistair Ewing, Alan Simpson, Stuart Paton, and Stephen Breur,"Introduction to the Cray T3D". Edinburgh Parallel Computing Centre, 1996.

[8] Message Passing Interface Forum. MPI:A Massage-Passing Interface Standard, 1994-5 (Postscript or HTML versions available at

`http://www.epcc.ed.ac.uk/epcc-tec/document.html`

)

[9] Complexity International,

`http://www.csu.edu.au/ci`

.

[10] C.W. Tidd L.F. Olsen and W.M. Schaffer: The case for chaos in childhood epidemics:II. Predicting historical Epidemics from mathematical models, *Proc. R. Soc. Lond.* **B254, 257-273 ,1993.**


Jennifer Market will graduate with a BSc in Physics from Texas A&M University in December. She plans to earn her PhD in Physics from Edinburgh University. An unlikely computer nerd, she prefers to spend her time hanging from the side of a mountain or on the back of a horse.

This project was supervised by Elspeth Minty (EPCC).