PAPER

Phys. Educ. **55** (2020) 015021 (21pp) iopscience.org/ped

The pedagogical power of context: extending the Epidemiology of Eyam

J P Cullerne[1,2], A French[1], D Poon, A Baxter and R N Thompson[2]

[1] Science School, Winchester College, Winchester, Hampshire, SO23 9JE, United Kingdom
[2] Mathematical Institute, University of Oxford, Oxford, Oxfordshire, OX2 6GG, United Kingdom

E-mail: jpc@wincoll.ac.uk, af@wincoll.ac.uk and robin.thompson@chch.ox.ac.uk

Abstract

This paper describes the further development of the *Epidemiology of Eyam* project at Winchester College since the publication of our previous paper (French *et al* 2019 Phys. Educ. 54 045008). A much wider cadre of students have been involved with this recent phase of the study, and we have also benefitted from a sabbatical collaboration with Oxford University Mathematical Institute. It is hoped that the research presented here can be used as a case study for an Extended Project Qualification (EPQ) or equivalent in a range of schools around the UK and worldwide. We feel that a project such as this, with a strong humanities and public health rooted context, could incentivise students of mathematics and sciences to participate in inter-disciplinary teams over an educationally significant period, and offer an opportunity to develop vital independent research skills that are required at University, but are often difficult to experience in a School context. The main educational goal of our previous paper was to provide an suitable context to incentivise the introduction of Calculus ideas. In this paper we assume a slightly higher level of mathematical skill, and aim to present a more comprehensive analysis of the epidemiological model that we will refer to as the 'Eyam Equations.' We describe a 'semi-analytic' solution, and make the connection to the approximation of Kermack and McKendrick (Kermack and McKendrick 1927 *Proc. R. Soc. Lond.* A 115 700–21) which held sway for much of the early twentieth century. We revisit the Eyam 1660s plague data of William Mompesson, and also apply the model to Ebola data collected in Liberia during the 2014–16 epidemic in West Africa. Motivated by uncertainty in the size of the 'at-risk' Susceptible population, we re-parameterize the model in terms of alternative inputs which enable a curve-fitting mechanism to be conducted more efficiently, with a much more tightly bounded range of possible parameter values. In addition to a spreadsheet model, we have created a software application in the MATLAB environment which has dynamic tools that could potentially enable a sensible curve fit to be calculated very rapidly in response to new data. From a time series describing the Infective population of the 2014–16 Liberia Ebola epidemic, we can predict Infective

1361-6552/20/015021+21$33.00

$I(t)$, Susceptible $S(t)$ and Dead $D(t)$ populations, calculate the associated total population size N, calculate the Kendall Susceptible threshold ρ, and hence the Basic Reproduction Number $R_0 = N/\rho$. For the Liberia Ebola data, these are: $N = 2542,\ \rho = 1373,\ R_0 = 1.85$. Note this suggests that out of an 'at risk' population of $N = 2542$, about 75% may ultimately have been infected by the Ebola virus. In the WHO Ebola Response Team report (WHO Ebola Response Roadmap Situation Report), $R_0 = 1.83 \pm 0.11$ for the 2014–16 Liberia outbreak.

1. Introduction

The transition from school to university represents a significant change in educational approach, with independent study skills being particularly important in higher education. It is therefore beneficial for students preparing for university study to experience extended project work in which a less directed way of studying is nurtured and developed. There is also the important difference in the use of multiple sources of ideas, which undergraduates new to the approach can find bewildering. In extreme cases, a lack of experience in paraphrasing or acknowledging a range of sources may even inadvertently land the inexperienced into trouble for plagiarism.

How best to provide a research experience at school is therefore of utmost importance. Students entering university are not initially provided with school-like delivery. Instead, they are immediately met with the lecture format, which often provides only background knowledge. To succeed, students must complete independent study around the topics introduced in lectures, fill in the finer details, and to develop and critique the present thinking within their disciplines.

Extended project work at school level can provide this experience in an otherwise very directed environment. To be effective, however, there must be a body of readily accessible material for students to research, and some aid with direction.

In this paper, we illustrate the exchange between students and supervisors arising out of our *Epidemiology of Eyam* project, which was first reported in our previous article [4]. The work of early twentieth century researchers in this field like Kermack, McKendrick, Kendall and Gani [1–3], are replete with examples of applications of mathematics accessible to a pre-university student. However, their original form, with a target audience of professional mathematicians, is perhaps somewhat formidable to all but the keenest students, and therefore can be a barrier to learning. In this paper we have endeavoured to distil the essence of what we shall deem the 'Eyam equations' and communicate the key ideas in a form that is more accessible to a target audience of pre-university students of mathematics and the physical sciences. Once again, the key educational aspect here is that with the right combination of technical and moral support, the students *discover for themselves* how the mathematics they have learnt at school can be applied to something as important as predicting the dynamics of infectious disease outbreaks. Students are then highly incentivised to engage with the project, deriving the major results and applying them to the Eyam plague outbreak as the prototype problem for their newly acquired skills. Guidance by a knowledgeable supervisor is crucial, so that students do not become lost in the large volume of material, but in our experience, students are able to work through this project with only limited direction if the initial briefing and associated resources are appropriate to their prior knowledge.

Since the introduction of the Extended Project Qualification (EPQ) in 2006, the number of students taking up the qualification has increased dramatically. The research aspect of the activity allows students to develop independent study skills, which are beneficial for subsequent work at university, especially if the EPQ involves rigorous academic content developed over an extended period. Our *Epidemiology of Eyam* project has provided a rich seam of material for extended study, which has so far inspired a great many of the students at Winchester College. We hope that it will continue to inspire students at Winchester, and elsewhere, as the study develops. Importantly, this project generates opportunities for inter-disciplinary research. There is scope for collaborative work across subjects including the physical and biological sciences, mathematics and economics. Where our previous paper [4] set the contextual scene and provided an introduction to calculus and associated numerical techniques, in this article we demonstrate that the elegant work in infectious disease epidemiology from the first half of the twentieth century [1–3] can be presented in an accessible way to pre-university students, providing them with a powerful context to delve into interesting problems and incentivising the acquisition of more advanced methods and skills. The techniques we employ to analyse data from the Eyam plague outbreaks of the 1660s can also be used with data from modern day epidemics, as we show in the context of the 2014–16 Ebola epidemic in West Africa [6, 11–13].

To permit our results to be replicated and extended by students, the computational work can be achieved using nothing more than a standard issue calculator and a spreadsheet, though as in previous papers [4, 5] we show this project can be more fully explored if a student implements their mathematical recipes in a coding environment such as MATLAB, a programming language that is widely used during undergraduate courses in the UK. To this end we have developed an intuitive software tool that can enable an epidemiological model to be rapidly fitted to field data. From a time series of Infective population for the 2014–16 Liberia Ebola epidemic [6] we can predict Infective $I(t)$, Susceptible $S(t)$ and Dead[3] $D(t)$ populations, calculate associated population N, calculate the Kendall Susceptible threshold ρ, and hence the Basic Reproduction Number $R_0 = N/\rho$. For the Liberia data, these are:

$$N = 2542, \;\; \rho = 1373, \;\; R_0 = 1.85.$$

Our paper follows a similar structure to that of Rachah and Torres (2015) (Discrete Dynamics in Nature and Society) [7], in the sense that we apply the same classic trio of differential equations to model *S,I,D* populations. However, in this paper we echo the work of Kendall *et al* to develop a 'semi-analytic' solution scheme which we feel enables a student, or perhaps even a field epidemiologist, to determine optimal parameters in a more intuitive manner.

2. The Eyam equations and how to solve them

2.1. The Eyam equations

In our original paper [4] we defined a trio of first-order differential equations to describe the time t variation of Susceptible S, Infective I and Dead D subsets of a fixed (or 'closed') population. The model assumes a flow from Susceptible, to Infective, to Dead, with no possibility of reversion from Infective to Susceptible, or indeed recovery from Infective.

In fact, it is better thought of the model as a *pair* of differential equations for D and S, plus the *constraint* of fixed population:

$$\boxed{\begin{array}{l}\frac{\mathrm{d}D}{\mathrm{d}t} = \alpha I, \quad \frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI \\ S + I + D = \text{constant}\end{array}}.$$

Differentiating the constraint yields the third differential equation in I:

$$\begin{aligned} &\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}D}{\mathrm{d}t} = 0 \\ \therefore &\frac{\mathrm{d}I}{\mathrm{d}t} = -\frac{\mathrm{d}S}{\mathrm{d}t} - \frac{\mathrm{d}D}{\mathrm{d}t} \\ \therefore &\boxed{\frac{\mathrm{d}I}{\mathrm{d}t} = (\beta S - \alpha)\, I}. \end{aligned}$$

We shall refer to this trio as the *Eyam equations* in the subsequent discussion. Rather than use the equations as motivation for a student making their *initial* forays into Calculus, (as we did in 'The Epidemiology of Eyam' [4], and 'Numerical Methods as an Introduction to Calculus' [5]) let us now assume a greater prowess with mathematical technique; i.e. the tools a first year undergraduate in Physics should hopefully be familiar with. In other words, we shall not go straight to the Euler method[4] and a spreadsheet to evaluate the Eyam equations. Before attempting to apply any numerical method, let us see how far we can go towards

[3] In most epidemiological literature the Dead population is known as R 'Removed.' From a modelling point of view, Dead and Removed are the same, with the key characteristic is that they no longer contribute to spread of disease. Interestingly, only 45% of infected actually went on to die in the Ebola 2014–16 outbreak in Liberia, and also, Ebola dead hosts often continue to contribute to transmission. Hence the meaning of 'Dead' may be somewhat ambiguous here. However, to avoid confusion with 'Recovered', (although none are permitted in the Eyam model) and to continue the narrative of [4], we will stick to $D(t)$ meaning 'those who have died due to the disease.'

[4] i.e. approximate the Eyam equations as finite differences using a fixed timestep Δt and then solve iteratively, starting from: $t = 0, I = I_0, S = S_0, D = 0$.

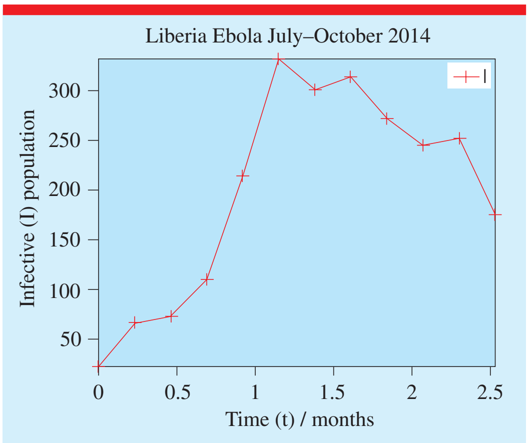


Figure 1. Confirmed infectives versus time for Liberia Ebola outbreak July–October 2014 [6].

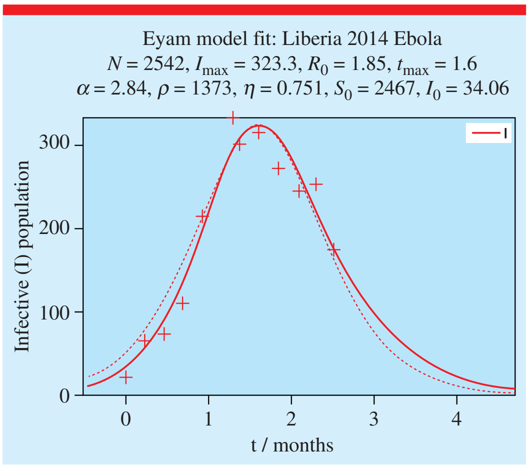


Figure 2. Confirmed infectives versus time for Liberia Ebola outbreak July–October 2014 [6], underlaid with a curve fit. The solid line is the 'semi-analytic' model described in section 4, and the dashed curve is the K&K approximate curve described in section 3. $(t_{\max}, I_{\max})$ are the peak coordinates of the fitted curve. N is the total population $S + I + D$, The Basic Reproduction Number $R_0 = N/\rho$, and η is the ratio between the total cumulative death toll and the total population.

an *analytic* solution. In other words, can we devise a mathematical function which describes the time variation of the S, I, D curves, based upon parameters which we can readily obtain from epidemiological field data?

2.2. The ups and downs of I

The Eyam equation for Infectives I is:

$$\frac{\mathrm{d}I}{\mathrm{d}t} = (\beta S - \alpha)\, I.$$

It is immediately apparent that $\frac{\mathrm{d}I}{\mathrm{d}t} = 0$ if $I = 0$ or $S = \frac{\alpha}{\beta}$. By performing a further time derivative, one can see that I is *maximized* when $S = \frac{\alpha}{\beta}$. This is the Susceptible population at the peak of the infection.

$$\begin{gathered}
\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} = (\beta S - \alpha)\,\frac{\mathrm{d}I}{\mathrm{d}t} + I\beta\frac{\mathrm{d}S}{\mathrm{d}t} \\
\therefore \frac{\mathrm{d}^2 I}{\mathrm{d}t^2} = (\beta S - \alpha)^2 I - I^2\beta^2 S \\
\therefore \left.\frac{\mathrm{d}^2 I}{\mathrm{d}t^2}\right|_{S=\frac{\alpha}{\beta}} = \left(\beta\frac{\alpha}{\beta} - \alpha\right)^2 I - I^2\beta^2\frac{\alpha}{\beta} = -I^2\beta\alpha \\
\therefore \left.\frac{\mathrm{d}^2 I}{\mathrm{d}t^2}\right|_{S=\frac{\alpha}{\beta}} < 0.
\end{gathered}$$

Note we use $\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI$ in the second step.

Now since $\frac{\alpha}{\beta}$ is clearly a useful parameter, define:

$$\boxed{\rho = \frac{\alpha}{\beta}}.$$

If ρ can be found for a given infection, then this represents a *threshold* for the epidemic. If the 'local susceptible' population[5] is less than this, then the Eyam model predicts the number of infectives will *reduce* rather than grow.

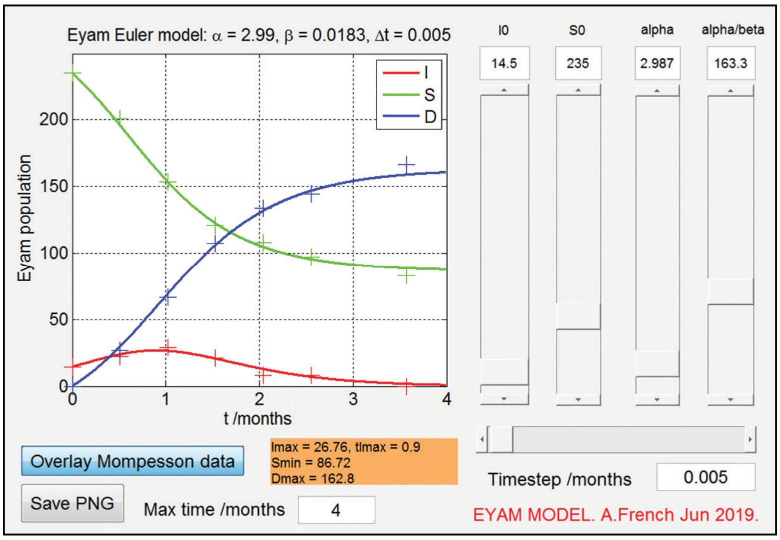
Without any further investigation, it is clear that $I(t)$ must form a single peaked curve, if the initial susceptible population is $> \rho$. Figures 1 and 2 below illustrate an $I(t)$ variation, and a curve fit based upon the recipe that we shall describe in this paper.

2.3. A problem of initial conditions, and dynamic exploration of the Eyam equations using a graphical user interface (GUI)

The Eyam model described in [4] has four key parameters S_0, I_0, α, β. In addition, the numerical evaluation[6] requires a time-step Δt, and a time range $[0, t_{\max}]$. It is very instructive for students to gain an *intuition* how the overall shape of the curves $S(t), I(t), D(t)$ varies with these parameters. An effective mechanism for achieving this is to code the Eyam model into a GUI, and use 'slider bars' (or an equivalent GUI mouse or pointing device driven interface) to enable the student to explore the change of $S(t), I(t), D(t)$ *dynamically* in response to a change to the input parameters.

[5] As suggested in the previous section, a modern epidemic might perhaps be modelled as a 'locally closed' population over a suitably truncated time period.

[6] The Eyam solver used in the GUI is the Euler method described in [4].


Eyam Euler model: α = 2.99, β = 0.0183, Δt = 0.005

I0: 14.5

S0: 235

alpha: 2.987

alpha/beta: 163.3

Imax = 26.76, tImax = 0.9
Smin = 86.72
Dmax = 162.8

Timestep /months: 0.005

Max time /months: 4

EYAM MODEL. A.French Jun 2019.

Figure 3. The Eyam model coded in the form of a GUI, implemented in MATLAB.

The vertical slider bars, a bit like graphical equalizers in an audio mixing desk, enable the four input parameters to be varied. The model in figure 3 is evaluated dynamically. A toggle option is available to overlay (or not) the Mompesson Eyam 1666 plague data. The maximum Infective, minimum Susceptible and maximum Dead values are calculated from the Eyam equations and displayed in the (orange) text box.

The GUI approach can be a very tangible goal for students keen to develop computer programming skills. Below is a screenshot of an equivalent Eyam model GUI implemented in the Godot Game Engine environment by a Sixth Form student, Alfie Baxter.

Unfortunately, this method of modelling is not very applicable to most modern epidemics. The Eyam model requires us to know both the Infective and Susceptible populations at $t = 0$ and the model itself depends on parameters α and β.

The system of equations also assumes a closed population in the sense that $I + S + D = I_0 + S_0$.

For a modern epidemic such as Ebola in Liberia [6], we may only have a record of 'confirmed infectives' versus time, and the assumption of an isolated population with an associated limited susceptible population is likely to be flawed.

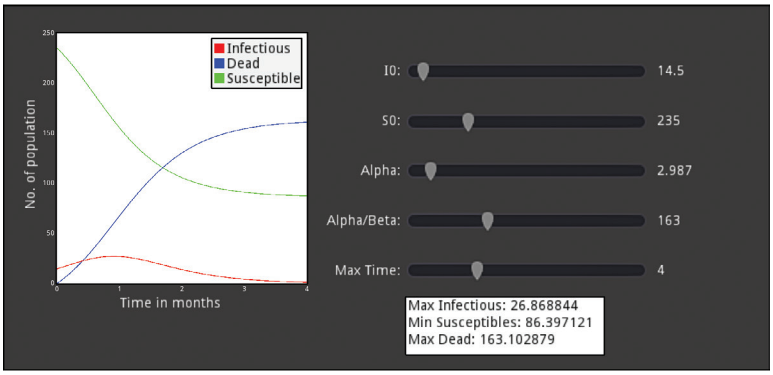
However, infective versus time curves do nonetheless appear to follow similar trends as per the Eyam model. We can therefore flip the logic of the problem: can we fit an Eyam model to work out what the dynamics of $D(t)$ and $S(t)$ would be, given the $I(t)$ data? If we can do this, then perhaps we can gain some insight into the size of the susceptible population that would catalyse the epidemic, and then perhaps use this insight to postulate whether certain social practices (e.g. mass social gatherings during funerals following Ebola deaths) are key accelerants for the epidemic.

Since S_0 and β are very unlikely to be known *a priori*[7], we shall adopt a more sensible set of initial conditions for the Eyam conditions, those that were employed by Kendall [2] in his criticism of the seminal (1927) work of Kermack and McKendrick ('K&K') in their *Contribution to the mathematical theory of epidemics* [1].

We shall consider time to begin at the peak of the infection and extend time over range $[-\infty, \infty]$ rather than be constrained by some artificial starting point of data collection, as per the Mompesson Eyam 1666 data. We shall also set our Dead count to be *zero* at the infection peak, which means a *negative* value prior to this. This might seem a rather perverse thing to do, but it shall turn out to be very useful in simplifying the mathematics of the solution to the Eyam equations. Particularly as $S = \rho$ at the peak of the infection.

Additionally, we shall scale our variables; time by α and S, I, D by ρ. In a similar way that

[7] α might be easier to guess, given it seems to lie within the range $2 < \alpha < 4$ [8,15,16]



I0: 14.5
S0: 235
Alpha: 2.987
Alpha/Beta: 163
Max Time: 4

```
Max Infectious: 26.868844
Min Susceptibles: 86.397121
Max Dead: 163.102879
```

Figure 4. The Eyam model coded in the form of a GUI, implemented in the Godot Game Engine environment by Alfie Baxter.

the equations of fluid dynamics can be solved elegantly if characterized by dimensionless parameters groups such as a Reynolds Number[8], we can use this approach to solve the Eyam equations in a more universal fashion.

Our new variables shall be:

$$\tau = \alpha\left(t - t_{\max}\right)$$
$$x = \tfrac{I}{\rho}, y = \tfrac{S}{\rho}, z = \tfrac{D + D(t=-\infty)}{\rho}$$

and our initial conditions shall be:

$$\tau = 0; \quad x = x_{\max}; \quad y = 1; \quad z = 0$$

where $x_{\max} = \mathrm{I}_{\max}/\rho$.

Note $-D\left(t = -\infty\right)$ is the cumulative dead from the beginning of the model time, to the peak of the infection. So, for $z = 0$ at the infection peak, we must subtract $-D\left(t = -\infty\right)$ from D. Why the minus sign? well $D\left(t = -\infty\right)$ will be *negative*.

2.4. Solving the Eyam equations using a 'semi-analytic approach'

Using our scaled variables, we can re-write our Eyam equations:

$$\tfrac{\mathrm{d}D}{\mathrm{d}t} = \alpha I \Rightarrow \rho\alpha\tfrac{\mathrm{d}z}{\mathrm{d}\tau} = \alpha\rho x$$
$$\therefore \boxed{\frac{\mathrm{d}z}{\mathrm{d}\tau} = x}$$

$$\tfrac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI \Rightarrow \rho\alpha\tfrac{\mathrm{d}y}{\mathrm{d}\tau} = -\beta\rho^2 yx$$
$$\therefore \tfrac{\alpha}{\beta}\tfrac{\mathrm{d}y}{\mathrm{d}\tau} = -\rho yx$$
$$\therefore \boxed{\frac{\mathrm{d}y}{\mathrm{d}\tau} = -yx}$$
$$\tfrac{\mathrm{d}I}{\mathrm{d}t} = \left(\beta S - \alpha\right) I \Rightarrow \rho\alpha\tfrac{\mathrm{d}x}{\mathrm{d}\tau} = \left(\beta\rho y - \alpha\right)\rho x$$
$$\therefore \tfrac{\mathrm{d}x}{\mathrm{d}\tau} = \left(\tfrac{\beta}{\alpha}\rho y - 1\right) x$$
$$\therefore \boxed{\frac{\mathrm{d}x}{\mathrm{d}\tau} = \left(y - 1\right) x}.$$

By this procedure we have distilled the Eyam equations into their simplest form.

To solve, firstly divide the first two to find $z(y)$ (which means $D(S)$ if we know ρ)

$$\tfrac{\mathrm{d}z}{\mathrm{d}\tau} = x, \tfrac{\mathrm{d}y}{\mathrm{d}\tau} = -yx$$
$$\therefore \tfrac{\mathrm{d}y}{\mathrm{d}z} = -y$$
$$\therefore \textstyle\int_1^y \tfrac{1}{y'}\mathrm{d}y' = -\int_0^z \mathrm{d}z'$$
$$\therefore \ln y = -z$$
$$\therefore \boxed{y = e^{-z}}.$$

Since the fixed population constraint means:

$$I + S + D = I_{\max} + \rho + 0$$
$$\Rightarrow x + y + z = x_{\max} + 1.$$

Hence:

$$\boxed{x = x_{\max} + 1 - e^{-z} - z}.$$

[8] When a Reynolds Number is low enough, a fluid will not behave in a turbulent fashion, regardless of the type of fluid or indeed the geometry and kinematics of the flow. All these factors are packaged into the Reynolds number. Re $= \frac{uL}{\nu}$ where u is the fluid velocity, L is a characteristic linear dimension and ν is the kinematic viscosity of the fluid. For a pipe of diameter D, the transition from laminar to turbulent flow typically occurs for Reynolds numbers above about 2300 [14].

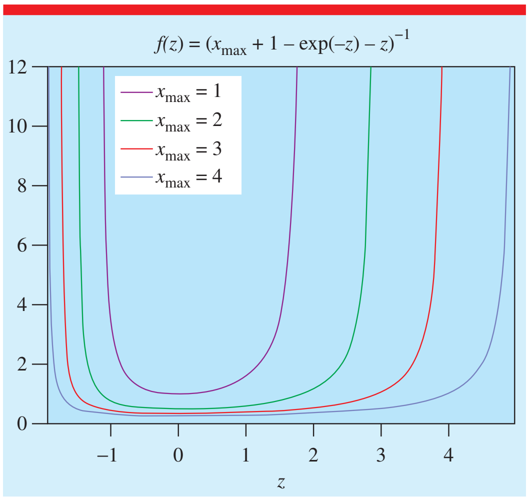


Figure 5. Plots of the integrand $f(z) = \frac{1}{x_{\max}+1-e^{-z}-z}$ of $\tau(z)$ for various values of $x_{\max}$. The limits of z are determined by the Newton Raphson iterative method.

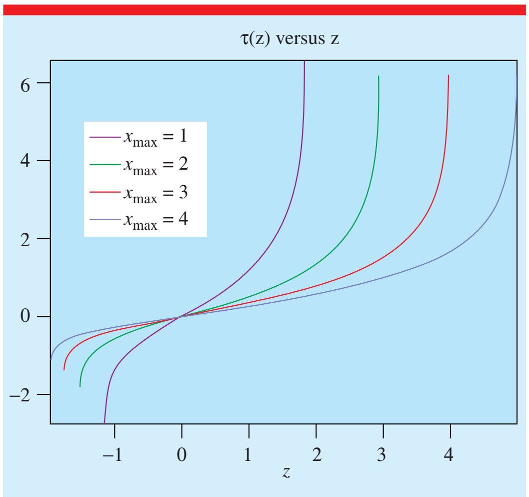


Figure 6. Plots of $\tau(z) = \int_0^z \frac{dz'}{x_{\max}+1-e^{-z'}-z'}$ for various values of $x_{\max}$. The integral has been evaluated numerically via the use of a *cubic spline fit* (see appendix C) between 1000 equally spaced z values between limits $[z_-, z_+]$.

Note that with $\tau = 0, z = 0$, this means that the total population N does *not* equal $\rho\,(x + y + z)$. In fact, it is:

$$N = \rho\,(x + y + z - z_-)$$

which means we have a strong incentive to figure out what $z_- = D\,(t = -\infty)$ is in terms of model inputs. As alluded to in the caption of figure 2, $R_0 = N/\rho$ is called the *Basic Reproduction Number*. A more detailed explanation of its epidemiological relevance is provided in appendix A.

We can use $x = x_{\max} + 1 - e^{-z} - z$ and $\frac{dz}{d\tau} = x$ to express an equation for $\tau(z)$. This is an *integral* expression, which unfortunately cannot be evaluated in a closed form sense; it requires a numerical approach[9]. Hence the 'semi-analytic' description of this solution scheme.

$$\frac{dz}{d\tau} = x_{\max} + 1 - e^{-z} - z$$
$$\therefore \int_0^z \frac{dz'}{x_{\max}+1-e^{-z'}-z'} = \int_0^\tau d\tau'$$
$$\therefore \boxed{\tau(z) = \int_0^z \frac{dz'}{x_{\max} + 1 - e^{-z'} - z'}}.$$

Example graphs of $\tau(z)$ and its integrand $f(z) = \frac{1}{x_{\max}+1-e^{-z}-z}$ are given in figures 5 and 6.

It is clear from the form of $f(z)$, that zero values of the denominator will cause an infinite asymptote. The meaning of this limit is more than pure mathematical consequence, since $\tau(z) = \int_0^z \frac{dz'}{x(z')}$. $x = 0$ means $I = 0$, which represents the boundary values of the infection at times $[-\infty, \infty]$. Since $I \geqslant 0$, this means the limits of z are the solutions to:

$$x_{\max} + 1 - e^{-z_-} - z_- = 0$$
$$x_{\max} + 1 - e^{-z_+} - z_+ = 0.$$

Since $x(z)$ has a single maximum at $z = 0$, an iterative numeric solution can readily be found by using the efficiently converging Newton Raphson[10] method, starting with initial values $z_\pm = \pm 1$.

$$z_\pm^{(n+1)} = z_\pm^{(n)} - \frac{x_{\max} + 1 - e^{-z_\pm^{(n)}} - z_\pm^{(n)}}{e^{-z_\pm^{(n)}} - 1}.$$

[9] E.g. a method which divides up the area under the integrand, approximately, into strips of precisely known size (e.g. trapeziums, parabolae, cubics etc) and then sums these. The thinner the strip width, the more accurate the numeric integral. For MATLAB implementation, see appendix B.

[10] $f(z) = 0$ can be solved by performing the iteration $z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}$. This can converge very rapidly as long as you do not encounter a stationary point where $f'(z) = 0$.

2.5. Eyam equation solution summary

The method above defines a ‘recipe’ for the solution of the Eyam equations. In summary:
Inputs parameters:

$$\rho, I_{\max}, \alpha, t_{\max}.$$

Mathematical model recipe:

1. Define $x_{\max} = \frac{I_{\max}}{\rho}$
2. Find limits of z, $[z_-, z_+]$ by solving the following Newton-Raphson iteration:

$$z_{n+1} = z_n - \frac{x_{\max} + 1 - e^{-z_n} - z_n}{e^{-z_n} - 1}.$$

   Start with $z_0 = \pm 1$, which avoids any stationary points.
3. Evaluate numerically the $\tau(z)$ integral, over an equally spaced range between $[z_-, z_+]$ Methods could be the Trapezium rule, ‘cubic spline fit and integrate’ etc. (See appendices B and C).

$$\tau(z) = \int_0^z \frac{dz'}{x_{\max} + 1 - e^{-z'} - z'}.$$

4. Determine x, y

$$x = x_{\max} + 1 - e^{-z} - z$$
$$y = e^{-z}.$$

5. Determine S, I, D, t

$$t = \frac{\tau}{\alpha} + t_{\max}, \quad I = \rho x \quad S = \rho y, \quad D = \rho\left(z - z_-\right).$$

6. Calculate population size N and hence Basic Reproduction Number R_0

$$N = I_{\max} + \rho - \rho z_-$$
$$R_0 = \tfrac{N}{\rho}.$$

3. The K&K approximate solution to the Eyam equations

3.1. The K&K approximation of $\tau(z)$

Until the criticism of Kendall [2], Kermack and McKendrick [1] was orthodoxy on the shape of $I(t)$ curves, which they showed could take a sech^2 functional form. As Kendall pointed out, and we will demonstrate using both the Liberia Ebola 2014–16 and the Mompesson Eyam 1966 plague data, a symmetric form is not an accurate portrayal of the curve, and indeed can lead to additional errors in the prediction of $D(t)$ and $S(t)$ curves.

The K&K approximation is to assume:

$$e^{-z} \approx 1 - z + \tfrac{1}{2}z^2$$

which means:

$$x \approx x_{\max} + 1 - \left(1 - z + \tfrac{1}{2}z^2\right) - z$$
$$\therefore \quad x \approx x_{\max} - \tfrac{1}{2}z^2.$$

For this to be valid, $z \ll 1$, which means the ratio of the dead D to the threshold ρ must be smaller than unity. We can immediately see from figure 5 that the z limits can be much larger than unity as we pass the infection peak, so this approximation will become increasingly crude as time beyond peak infection increases.

However, if only for the educational and historical value, let us persist with the analysis.

Under the K&K approximation, the $\tau(z)$ integral becomes:

$$\begin{aligned}
\tau(z) &= \textstyle\int_0^z \frac{\mathrm{d}z'}{x_{\max}+1-e^{-z'}-z'}\\
\tau(z) &\approx \textstyle\int_0^z \frac{\mathrm{d}z'}{x_{\max}+1-\left(1-z'+\frac{1}{2}z'^2\right)-z'}\\
&= \textstyle\int_0^z \frac{\mathrm{d}z'}{x_{\max}-\frac{1}{2}z'^2}\\
&= 2\textstyle\int_0^z \frac{\mathrm{d}z'}{2x_{\max}-z'^2}\\
&= 2\textstyle\int_0^z \frac{\mathrm{d}z'}{\left(\sqrt{2x_{\max}}\right)^2-z'^2}\\
&= \tfrac{2}{\sqrt{2x_{\max}}}\tanh^{-1}\left(\tfrac{z}{\sqrt{2x_{\max}}}\right)
\end{aligned}$$

where the final step makes use of the standard integral:

$$\int \frac{\mathrm{d}x}{a^2 - x^2} = \tfrac{1}{a}\tanh^{-1}\left(\tfrac{x}{a}\right) + c.$$

Hence:

$$\boxed{z \approx \sqrt{2x_{\max}}\,\tanh\left(\sqrt{\tfrac{1}{2}x_{\max}}\tau\right)}$$

and therefore:

$$\begin{aligned}
x &\approx x_{\max} - \tfrac{1}{2}z^2\\
&= x_{\max} - \tfrac{1}{2}2x_{\max}\tanh^2\left(\sqrt{\tfrac{1}{2}x_{\max}}\tau\right)\\
&= x_{\max}\left\{1 - \tanh^2\left(\sqrt{\tfrac{1}{2}x_{\max}}\tau\right)\right\}\\
\therefore \quad &\boxed{x \approx x_{\max}\mathrm{sech}^2\left(\sqrt{\tfrac{1}{2}x_{\max}}\tau\right)}.
\end{aligned}$$

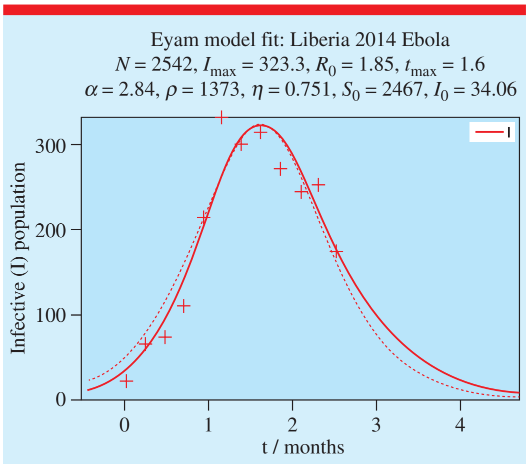


Figure 7. Liberia 2014–16 Ebola data underlaid with the semi-analytic, and K&K $I(t)$ curve. The dashed line indicates the K&K sech^2 approximation. The 'exact' (i.e. the semi-analytic approach defined in section 2.5) is the much better fit.

In summary, the K&K approximate solution to the Eyam equations is:

$$\boxed{\begin{aligned} &x \approx x_{\max}\mathrm{sech}^2\left(\sqrt{\tfrac{1}{2}x_{\max}}\tau\right)\\ &z \approx \sqrt{2x_{\max}}\tanh\left(\sqrt{\tfrac{1}{2}x_{\max}}\tau\right)\\ &y = e^{-z}. \end{aligned}}$$

Now $\lim_{x\to\pm\infty} \tanh(x) = \pm 1$

So therefore: $z_{\pm} \approx \pm\sqrt{2x_{\max}}$
Hence:

$$\boxed{\begin{aligned} &t = \tfrac{\tau}{\alpha} + t_{\max}, I = \rho x, S = \rho y\\ &D = \rho\left(z + \sqrt{2x_{\max}}\right)\\ &N = I_{\max} + \rho - \rho\sqrt{2x_{\max}}\\ &R_0 = \tfrac{N}{\rho} \end{aligned}}.$$

3.2. Evaluation of the K&K model

To assess the efficacy of the K&K approximation, let us firstly apply the model to the Liberia Ebola data illustrated in figure 2. The same input parameters are used:

$$I_{\max} = 323,\ t_{\max} = 1.60,\ \rho = 1373,\ \alpha = 2.84,$$

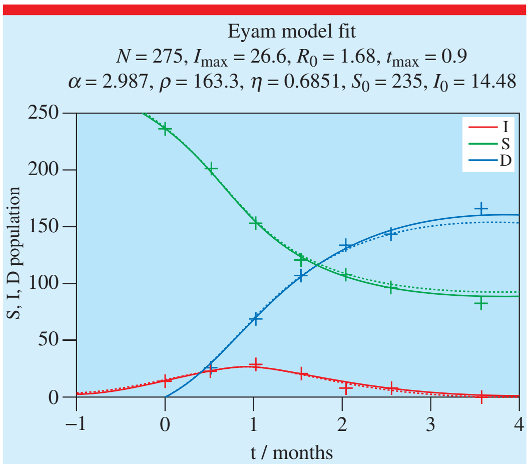


Figure 8. Mompesson 1966 data underlaid with the semi-analytic, and K&K $I(t), S(t), D(t)$ curves. The dashed lines correspond to the K&K approximation.

It is clear from figure 7 that the K&K approximation overestimates $I(t)$ below the peak, and underestimates beyond the peak. A similar discrepancy between semi-analytic and K&K approaches can be seen for the Mompesson 1666 plague data. Both models use the same parameters:

$$I_{\max} = 26.76,\ t_{\max} = 0.900,\ \rho = 163.3,\ \alpha = 2.987.$$

It is worth noting that the Mompesson data is based upon $D = 0$ at $t = 0$. Initial S, I values are $S(0) = 235$, $I(0) = 14.5$, giving a total population of 250. Although when to start the dead count is arbitrary, the total associated population N is not, as we have shown in previous sections. It is likely there would have been deaths due to plague at Eyam *before* Mompesson started collecting data.

Calculating N using the semi-analytic procedure yields $N = 275.42$, so about 25 people would have died of plague at Eyam before Mompesson's parish records began, had the infection obeyed the idealized Kendall curve from $t = -\infty$[11].

However, to directly compare to Mompesson's data, we shall shift both exact and K&K $D(t)$ curves by $D(0)$, so all pass through the origin.

3.3. Problems with the K&K approach

The time symmetric K&K $I(t)$ is not universally true, and the skew of the curve can lead to a discrepancy in predictions, as illustrated in figures 7 and 8.

[11] Actually there was a previous Plague outbreak at Eyam, so the $I(t)$ curve is only valid for the time range associated with Mompesson's data.

In addition, we also have the problem of what values to assign ρ and α. This is a problem for our semi-analytic method too. Based on field data such as a World Health Organization (WHO) Situational Report [6], all we have is a set of $\{t, I\}$ data. For the Mompesson records, we showed in [4] how α, β (and hence $\rho = \frac{\alpha}{\beta}$) could be obtained from the $S(t), I(t), D(t)$ dataset. For most epidemiological data, we do not have this luxury.

These problems shall be addressed in the next section.

4. A semi-analytic solution to the Eyam equations without ρ

4.1. The problem with ρ

The Eyam equation solution summary in 2.5 requires the inputs:

$$\rho, I_{\max}, \alpha, t_{\max}.$$

If the only data available is $\{t, I\}$, then $(t_{\max}, I_{\max})$ can be readily guessed directly from the data[12].

The range of α is relatively well bounded in epidemiological literature. For many epidemics, it is typically within the range of $2 < \alpha < 4$. (Units of months^{-1}). [8, 15, 16].

The problem parameter is ρ, which can vary significantly with the populations involved. For Mompesson 1966, $\rho = 163.3$, for Liberia Ebola 2014–16, $\rho = 1538$.

4.2. Replacing ρ with η

We cannot avoid having four inputs to our Eyam model, but we can replace ρ with a different variable that has a much more well-defined range of values. In fact, it is possible to find a parameter η which has bounds of $[0, 1]$.

The total dead is given by:

$$D_{\text{tot}} = \rho\,(z_+ - z_-)\,.$$

Clearly this cannot exceed the total population N, or indeed be less than zero.

$$0 < \tfrac{D_{\text{tot}}}{N} < 1$$
$$\therefore 0 < \tfrac{\rho(z_+ - z_-)}{N} < 1$$
$$\therefore 0 < \tfrac{z_+ - z_-}{N/\rho} < 1.$$

Define $\boxed{\eta = \dfrac{z_+ - z_-}{N/\rho}}$, which *must* be in the range [0,1].

Now since $x = x_{\max} + 1 - e^{-z} - z$,
when $x = 0$:
$x_{\max} + 1 - e^{-z_-} - z_- = 0$ and
$x_{\max} + 1 - e^{-z_+} - z_+ = 0$
Now: $\frac{N}{\rho} = x_{\max} + 1 - z_-$
Therefore using $x_{\max} + 1 - e^{-z_-} - z_- = 0$

$$\frac{N}{\rho} - e^{-z_-} = 0$$
$$\therefore \boxed{z_- = -\ln\left(\frac{N}{\rho}\right)}.$$

From the definition of η:

$$z_+ = \eta\frac{N}{\rho} + z_-$$
$$\therefore \boxed{z_+ = \eta\frac{N}{\rho} - \ln\left(\frac{N}{\rho}\right)}.$$

We can now substitute this in the equation for $x = 0$ for positive z:

$$x_{\max} + 1 - e^{-z_+} - z_+ = 0$$
$$x_{\max} + 1 - z_- - e^{-z_+} - z_+ = -z_-$$
$$\frac{N}{\rho} - e^{-z_+} = z_+ - z_-$$
$$\frac{N}{\rho} - e^{-\eta\frac{N}{\rho} + \ln\left(\frac{N}{\rho}\right)} = \frac{N}{\rho}\eta.$$

With the last step noting:

$$\eta = \frac{z_+ - z_-}{N/\rho}$$
$$\therefore \frac{\eta N}{\rho} = z_+ - z_-.$$

Hence:

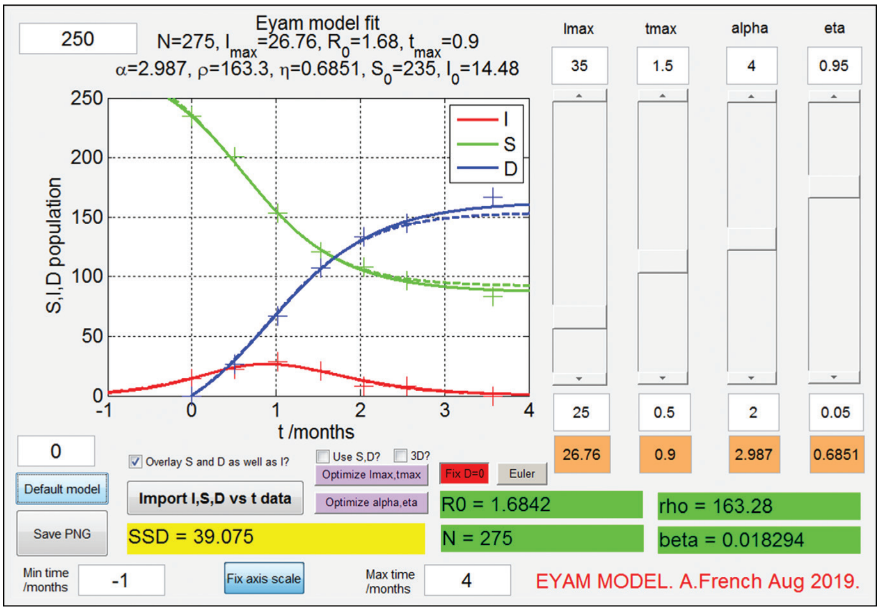
$$\frac{N}{\rho} - \frac{N}{\rho}e^{-\eta\frac{N}{\rho}} = \frac{N}{\rho}\eta$$
$$\therefore 1 - e^{-\eta\frac{N}{\rho}} = \eta$$
$$\therefore \boxed{\frac{N}{\rho} = -\frac{\ln(1-\eta)}{\eta}}.$$

If we specify η we can therefore define:

$$z_+ = \eta\frac{N}{\rho} - \ln\left(\frac{N}{\rho}\right)$$
$$\therefore \boxed{z_+ = -\ln(1-\eta) - \ln\left(-\frac{\ln(1-\eta)}{\eta}\right)}$$
$$z_- = -\ln\left(\frac{N}{\rho}\right)$$
$$\therefore \boxed{z_- = -\ln\left(-\frac{\ln(1-\eta)}{\eta}\right)}.$$

If we start with η we do not need to solve an iterative equation for $z_\pm$, so this method will significantly reduce computation time.

[12] Or indeed a good first guess. We will show in section 4.5 how an initial guess based upon the largest I value can be refined automatically.



Eyam model fit

N=275, I_{max}=26.76, R_0=1.68, t_{max}=0.9

α=2.987, ρ=163.3, η=0.6851, S_0=235, I_0=14.48

Imax	tmax	alpha	eta
35	1.5	4	0.95
25	0.5	2	0.05
26.76	0.9	2.987	0.6851

☑ Overlay S and D as well as I?

□ Use S,D? □ 3D?

SSD = 39.075

R0 = 1.6842

rho = 163.28

N = 275

beta = 0.018294

Min time /months: -1

Max time /months: 4

EYAM MODEL. A.French Aug 2019.

Figure 9. MATLAB GUI which enables Eyam model solver input parameters to be varied until a visual fit to data is achieved. The slider bars on the right-hand side allow $I_{\max}, t_{\max}, \alpha, \eta$ to be changed dynamically. In the graph, solid lines represent the semi-analytic model, and the dashed lines correspond to the K&K approximation.

Also:

$$\frac{N}{\rho} = x_{\max} + 1 - z_-$$
$$\Rightarrow x_{\max} = \frac{N}{\rho} - 1 + z_-$$
$$\therefore \boxed{x_{\max} = -\frac{\ln(1-\eta)}{\eta} - 1 - \ln\left(-\frac{\ln(1-\eta)}{\eta}\right)}.$$

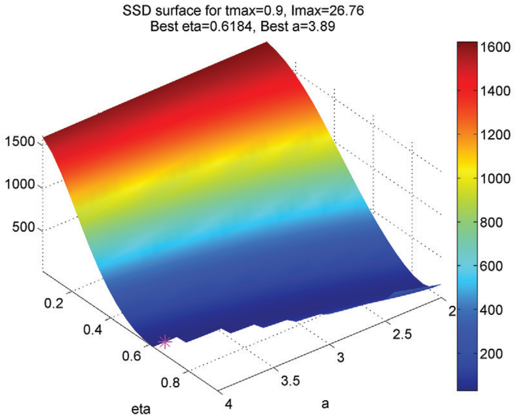
So, if we know $I_{\max}$ we can find:

$$\boxed{\rho = \frac{I_{\max}}{x_{\max}}}.$$

4.3. Eyam equation solver via η and a GUI

Let us summarize our Eyam equation solver using the $\eta = \frac{D_{\text{tot}}}{N}$ parameter:

$$\boxed{\begin{array}{l}
z_+ = -\ln(1-\eta) - \ln\left(-\frac{\ln(1-\eta)}{\eta}\right) \\
z_- = -\ln\left(-\frac{\ln(1-\eta)}{\eta}\right) \\
x_{\max} = -\frac{\ln(1-\eta)}{\eta} - 1 - \ln\left(-\frac{\ln(1-\eta)}{\eta}\right) \\
\rho = \frac{I_{\max}}{x_{\max}} \\
\tau(z) = \int_0^z \frac{\mathrm{d}z'}{x_{\max}+1-e^{-z'}-z'} \\
x = x_{\max} + 1 - e^{-z} - z \\
y = e^{-z} \\
t = \frac{\tau}{\alpha} + t_{\max}, \ \ I = \rho x \ \ S = \rho y, \ \ D = \rho(z - z_-) \\
N = I_{\max} + \rho - \rho z_- \\
R_0 = \frac{N}{\rho}
\end{array}}.$$



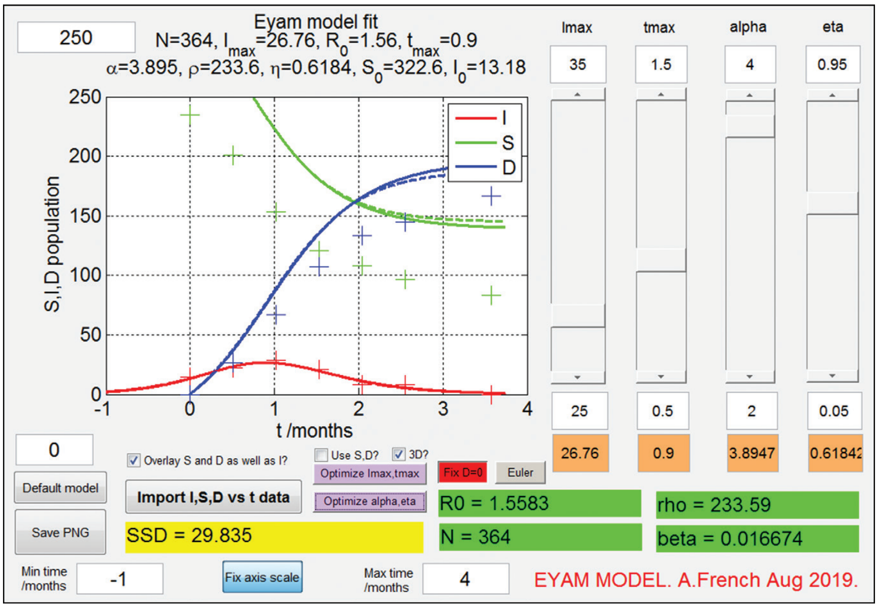
SSD surface for tmax=0.9, Imax=26.76

Best eta=0.6184, Best a=3.89

Figure 10. An SSD surface is formed by varying α and η parameters over the range $2 \leqslant \alpha \leqslant 4$ and $0.05 \leqslant \eta \leqslant 0.95$.

Although we know $0 < \eta < 1$, we still need a practical mechanism to determine what best fits a $\{t, I\}$ data set.

To achieve this, we have developed the GUI concept described in section 2.3. A screenshot is provided in figure 9, with the Mompesson plague data as an example.



250

Eyam model fit

N=364, I_{max}=26.76, R_0=1.56, t_{max}=0.9

α=3.895, ρ=233.6, η=0.6184, S_0=322.6, I_0=13.18

Imax	tmax	alpha	eta
35	1.5	4	0.95
25	0.5	2	0.05
26.76	0.9	3.8947	0.61842

0

☑ Overlay S and D as well as I?

□ Use S,D? ☑ 3D?

SSD = 29.835

R0 = 1.5583

N = 364

rho = 233.59

beta = 0.016674

Min time /months -1

Max time /months 4

EYAM MODEL. A.French Aug 2019.

Figure 11. Eyam solver MATLAB GUI with $t_{\max} = 0.900$, $I_{\max} = 26.76$. Data corresponds the Mompesson 1666 plague. The curves formed based upon the 'optimal' values of α, η are plotted in figure. Although SSD improves from 39.1 (see figure 9) to 29.8, it is clear that the S and D curves are not well matched to the Mompesson data.

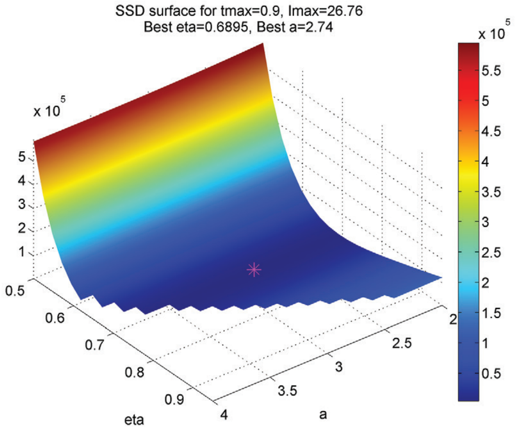


Figure 12. An SSD surface is formed by varying α and η parameters over the range $2 \leqslant \alpha \leqslant 4$ and $0.5 \leqslant \eta \leqslant 0.95$.

Inputs of $I_{\max} = 26.76$, $t_{\max} = 0.900$, $\alpha = 2.987$ were set as per the calculation performed in [4], and η was found to be $\boxed{\eta = 0.6851}$ when $\rho = 163.3$.

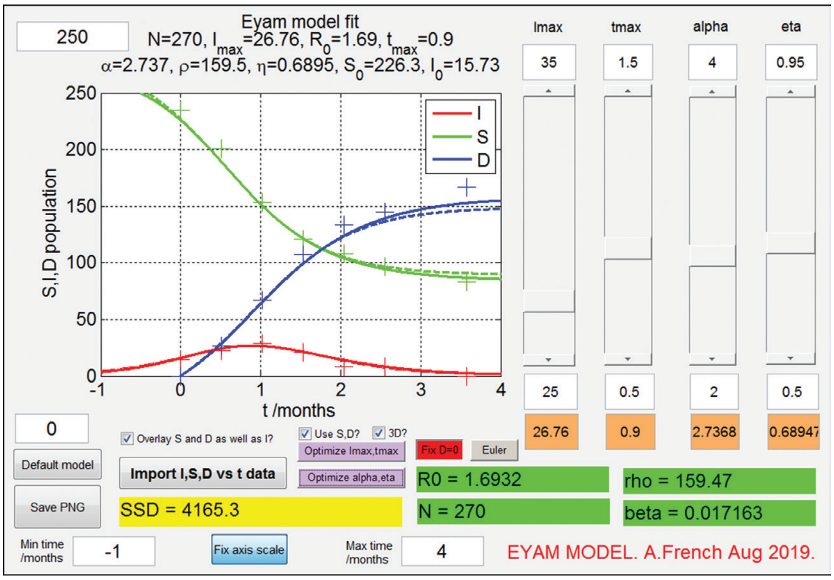
The correct values of $I_0 = 14.5, S_0 = 235$ are predicted, as is the total population of $N = 275$. Note in figure 9, the $D(t)$ curves have been shifted by 25 such that $D(0) = 0$, to enable comparison with the Mompesson data where the dead count starts at zero when time is zero.

4.4. Sum of squared differences

To offer a quantitative measure of curve fit, beyond visual correlation, a 'sum of squared differences' (SSD) is computed. This is the sum of the squares of the difference of model $I(t)$ values, interpolated at $\{t, I\}$ data times, to the corresponding I data values. In the above example SSD $= 39.075$. Alternatively, SSD can also include the sum of squared differences between model S and D predictions and their data counterparts, if the latter exist. The '`Use S, D?`' GUI checkbox allows for both possibilities.

4.5. Optimising $t_{\max}, I_{\max}, \alpha, \eta$

The SSD metric can be used as the basis of an automated optimisation of Eyam model parameters. The principle is to create a *surface* based upon two of the four Eyam model input parameters $t_{\max}, I_{\max}, \alpha, \eta$, and determine the parameter pair coordinate associated with the minimum of this surface. An ideal situation would be to explore all


Eyam model fit
N=270, I_{max}=26.76, R_0=1.69, t_{max}=0.9
α=2.737, ρ=159.5, η=0.6895, S_0=226.3, I_0=15.73

250

0

Imax 35 25 26.76

tmax 1.5 0.5 0.9

alpha 4 2 2.7368

eta 0.95 0.5 0.68947

☑ Overlay S and D as well as I?

☑ Use S,D? ☑ 3D?

R0 = 1.6932

rho = 159.47

N = 270

beta = 0.017163

SSD = 4165.3

Min time /months -1

Max time /months 4

EYAM MODEL. A.French Aug 2019.

Figure 13. Eyam solver MATLAB GUI with $t_{\max} = 0.900, I_{\max} = 26.76$, and optimized for α, η. Data corresponds the Mompesson 1666 plague. SSD improves from 4637 to 4165, if calculated to include S, D data as well as I.

four parameters in some form of 4D space, but this would be somewhat difficult to visualize!

Which pair of inputs from $t_{\max}, I_{\max}, \alpha, \eta$ should be chosen? The natural grouping appears to be $t_{\max}, I_{\max}$ i.e. visual characteristics of the $I(t)$ curve peak, and α, η, which relate to the width and skew of the $I(t)$ curve. The Eyam solver GUI has two 'Optimize' buttons which construct a mesh of parameters values, based upon the limits set by the sliders, and run the model over the full range. In figures 10–15 below (which are based upon the Mompesson data defaults), a grid of 20 × 20 parameters are used.

Automatic optimization based upon α and η parameters actually makes for a worse fit, if only I data is used in the computation of SSD. Figures 12 and 13 show the effect of comparing I, S, D in the SSD. This results in a much better fit. (Although note the SSD numbers are now much higher than if just I is compared).

Figure 12 indicates a somewhat flat SSD surface with α and η parameters, so it is perhaps harder to judge whether this method of automatic optimization confers much benefit over a manual movement of the GUI sliders until a visual fit is obtained.

An alternative automated optimization based upon fixing α and η and varying $t_{\max}, I_{\max}$ offers a much more obvious optimum, and is therefore

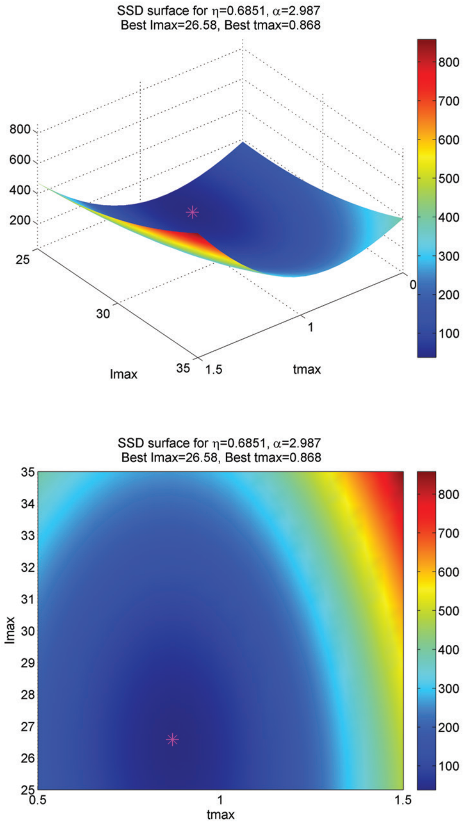
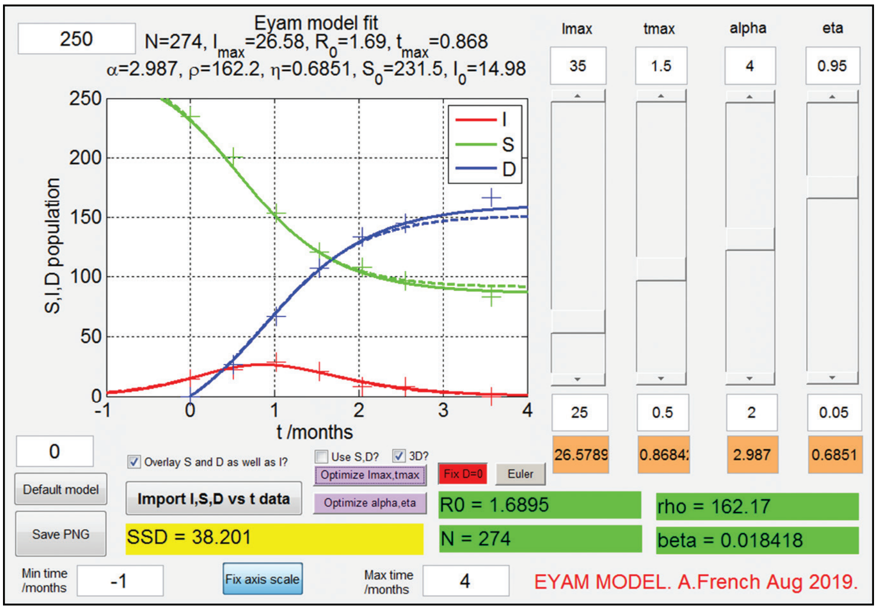

SSD surface for η=0.6851, α=2.987
Best Imax=26.58, Best tmax=0.868

Figure 14. An SSD surface is formed (3D and 2D views) using the Mompesson data, by varying $t_{\max}$ and $I_{\max}$ parameters, and fixing α, η.



Eyam model fit

N=274, I_{max}=26.58, R_0=1.69, t_{max}=0.868

α=2.987, ρ=162.2, η=0.6851, S_0=231.5, I_0=14.98

R0 = 1.6895

N = 274

rho = 162.17

beta = 0.018418

SSD = 38.201

EYAM MODEL. A.French Aug 2019.

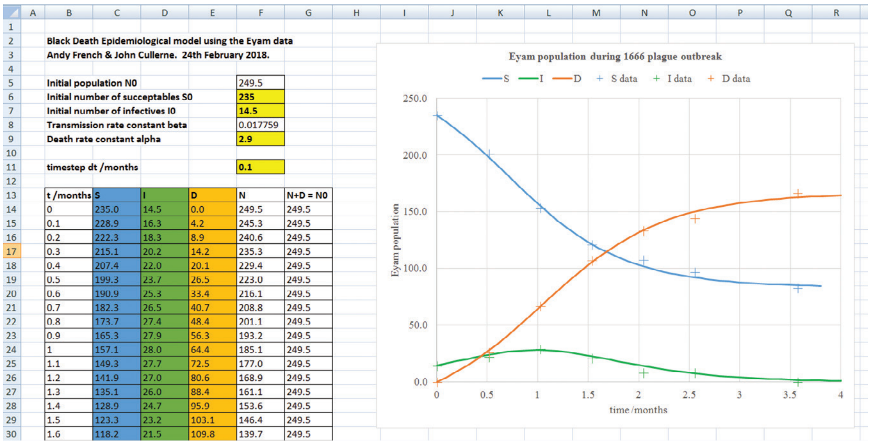
Figure 15. S, I, D curves using optimized $t_{\max}, I_{\max}$ parameters. There is a marginal SSD improvement from 39.1 to 38.2.

suggested as the preferred method. Figures 14 and 15 below illustrate this for the Mompesson data.

5. A story of students and spreadsheets

The narrative presented in sections 2 till 4 represents the authors most up-to-date understanding of the scenario posed and using what we hope are the most appropriate language and computational tools. The actual modelling process, which has been done over the past two years, was of course a much more meandering narrative; a slow process of gradual revelation and multiple iterations. Although this paper describes a mathematical model applied to epidemiology, the main aim of the study is not to contribute to epidemiological knowledge. We would prefer to leave that to the professionals! Our primary goal as educators is to develop a rich context to incentivise our students to engage in research and apply the skills they learn at school or university. A small, but growing, group of students have made significant contributions to this paper. It is important that some of the details of their involvement is recorded, as it will hopefully illustrate that the Eyam story can be re-explored by students with a wide range of skills.

Dexter Poon has been working on the *Epidemiology of Eyam* for about a year and began by making a study of the K&K model using his Pre-U Further Mathematics Calculus skills. He was able to derive the tanh and sech^2 forms of the $D(t)$ and $I(t)$ curves, starting with the idea that $t = 0$ at the peak of the infection. To help work through the derivation without being overwhelmed, both the K&K [1] and the Kendall [2] papers were literally chopped up and provided in manageable sections. Dexter then worked with JPC in developing in spreadsheet form what is described in section 4 as a 'semi-analytic' Eyam solver, which uses the η parameter rather than ρ. This was initially done without a MATLAB numeric integration routine, and instead each sample point of the $\tau(z)$ curve was evaluated using the integration function of a Casio FX-991EX calculator, which is standard issue at Winchester College and relatively inexpensive. Although time consuming, there was a certain excitement in seeing the curves develop. A spreadsheet was also used to solve for $z_{\pm}$ (and hence the limits of D) as described in section 2.4. Finally, the SSD surface concept was used, initially with a small mesh of α, η about what was deemed to be a sensible guess of optimal parameters. This led to a modest optimisation of the Liberia Ebola $I(t)$ fit, although the 'crumpled



Black Death Epidemiological model using the Eyam data
Andy French & John Cullerne. 24th February 2018.

Initial population N0	249.5
Initial number of succeptables S0	235
Initial number of infectives I0	14.5
Transmission rate constant beta	0.017759
Death rate constant alpha	2.9
timestep dt /months	0.1

t /months	S	I	D	N	N+D = N0
0	235.0	14.5	0.0	249.5	249.5
0.1	228.9	16.3	4.2	245.3	249.5
0.2	222.3	18.3	8.9	240.6	249.5
0.3	215.1	20.2	14.2	235.3	249.5
0.4	207.4	22.0	20.1	229.4	249.5
0.5	199.3	23.7	26.5	223.0	249.5
0.6	190.9	25.3	33.4	216.1	249.5
0.7	182.3	26.5	40.7	208.8	249.5
0.8	173.7	27.4	48.4	201.1	249.5
0.9	165.3	27.9	56.3	193.2	249.5
1	157.1	28.0	64.4	185.1	249.5
1.1	149.3	27.7	72.5	177.0	249.5
1.2	141.9	27.0	80.6	168.9	249.5
1.3	135.1	26.0	88.4	161.1	249.5
1.4	128.9	24.7	95.9	153.6	249.5
1.5	123.3	23.2	103.1	146.4	249.5
1.6	118.2	21.5	109.8	139.7	249.5

Figure 16. Solving the Eyam equations using a spreadsheet is a recommended first step for students. In this initial situation, they should use the Euler method, and Mompesosn's 1666 data.

paper' shape of the surface suggested that this pair of parameters may not be the most appropriate. This initial investigation was the spur to the more comprehensive analysis afforded by our MATLAB tools described in the previous section.

We have also involved several other students, perhaps the keenest being Alfie Baxter, who created an alternative Eyam Euler solver-based GUI using a Game Engine based programming environment (see figure 4).

From a pedagogical perspective, our recommended educational path through the Eyam story is as follows:

1. Start from the original Euler model of the Mompesson 1666 plague data described in [4]. Once the context is set via the educator (via reading the introduction to the paper and/or our associated slide pack), students should solve the Eyam equations via a spreadsheet (see figure 16).
2. Students that are more familiar with calculus should repeat the steps in sections 2 and 3. Based upon recommended input parameters for $t_{\max}, I_{\max}, \alpha, \rho$, they should attempt to generate the curves for the Mompesson and Liberia Ebola epidemics. Using the K&K approximation would be relatively straightforward in a spreadsheet such as Excel. Numeric evaluation of the $\tau(z)$ is more technically difficult, although a heroic semi-manual approach with a calculator will work eventually—but only for a very persistent student!
3. The $\tau(z)$ evaluation should be a strong incentive to use a programmatic tool such as MATLAB, Python or equivalent. As we discovered with Alfie Baxter, some students already have experience in a coding environment, and they should be free to choose to solve the problem in the language that they feel most comfortable. However, if students are starting out, MATLAB would be our recommendation.
4. At this point, students should explore the Eyam scenario using our MATLAB GUI, starting with the Mompesson and Liberia Ebola data, and then given the challenge to fit curves to other epidemics based upon WHO situational reports or equivalent. It would also be instructive for them to create their own Eyam solver from first principles. Rather than the complexity of a GUI, a simple MATLAB program to calculate and plot the $S(t), I(t), D(t)$ curves, overlaid with actual data, is suggested. The teacher could provide them with model inputs $t_{\max}, I_{\max}, \alpha, \eta$ and associated data. An even better situation would be for the students to determine optimal parameters themselves from the MATLAB Eyam GUI, and then check using their own code that they can regenerate the same curves.

6. Conclusions

6.1. A rich context for cross-curricular study

Our *Epidemiology of Eyam* project has proved to be a rich seam of material for sixth form students interested in extended project work in mathematical modelling. With encouragement and a little technical guidance, students can develop the techniques and approaches that allow them to make sense of epidemiological data, both historical and very much in the present. The amazing human story of the Eyam plague outbreaks in 1660s Derbyshire provides, in addition, a universal cross-curricular pedagogical context, which can then enable a linkage of this work to many other academic disciplines. Students of mathematics can contribute creatively with biologists, physicists, chemists, computer scientists, historians and, given this work all began with a study of a play *The Roses of Eyam*, the arts too.

6.2. A neat, holistic problem of applied mathematics

In this paper we present many opportunities to demonstrate the application of pre-university and early undergraduate mathematics, to something as important as modelling the spread of infectious disease. Each stage requires the gathering of several methods and techniques. Since the introduction of modularisation of A-level mathematics in the UK, there has been a growing risk of a lack of synoptic overview, for which this form of holistic project work may provide a remedy. The new linear A-levels will hopefully deal with this shortfall as well, though there will always be need for neat, classroom appropriate examples that can bring techniques together and thereby bring them to life.

6.3. Summary of our epidemiological research

The work here extends our previous paper's Euler approach [4] by calculating the Kendall 'exact' curve that would be equivalent. This is what we refer to as a 'semi-analytic solution' in sections 2 and 4. We apply the methods developed to the Mompesson 1666 Eyam plague outbreak, as well as the 2014–16 Ebola epidemic in Liberia. Using the optimization procedure described in section 4.5, our optimal parameters for the Liberia Ebola data are:

$$t_{\max} = 1.605,\ I_{\max} = 323.3,\ \ \alpha = 2.84,\ \ \eta = 0.751$$

and with derived parameters:

$$\rho = 1373, N = 2542, R_0 = 1.85.$$

Note this suggests that out of an 'at risk' population of $N = 2542$, about 75% may ultimately be infected by the Ebola virus.

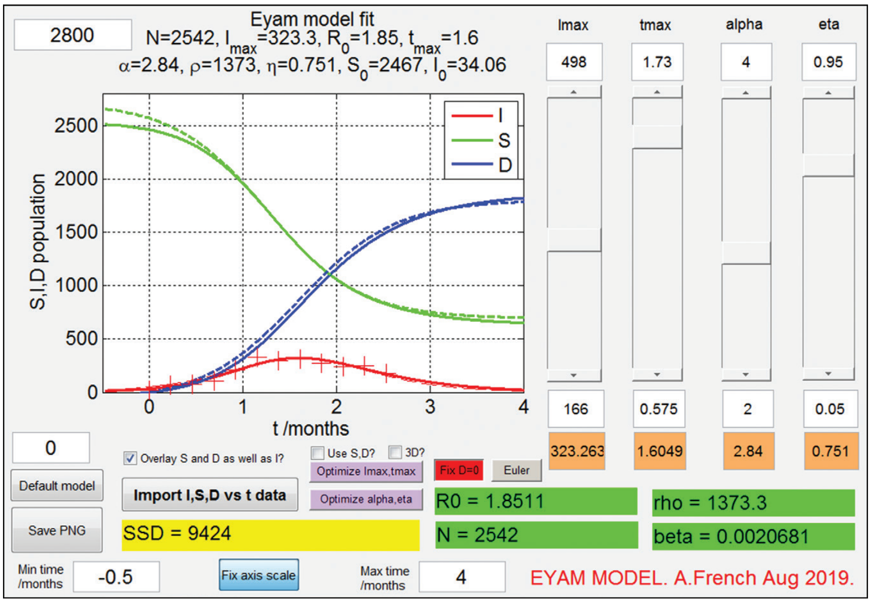
To achieve the result above, we performed *several* iterations using both the $(t_{\max}, I_{\max})$ and (α, η) optimizations. The $I(t)$ in figure 17 is indeed the same graph in figure 2. Figure 18 is the 2D rendering of the SSD surface used in the final automated $(t_{\max}, I_{\max})$ optimization.

We think that having an easy-to-use software tool may possibly be of use to modern epidemiologists as well, in addition to its educational value. To determine the predicted time series of an infective population, and to be able to suggest characteristic parameters of the epidemic such as the associated population N and Basic Reproduction number, could be useful in the field. Our tools could form part of an initial analysis of epidemiological data, and perhaps provide a sanity check of more detailed models.

In the WHO Ebola Response Team report [15], $R_0 = 1.83 \pm 0.11$ for the 2014-16 Liberia outbreak, so it is encouraging that our value of $R_0 = 1.85$ is within the same range.

6.4. Future projects

We are currently pursuing two parallel projects with our Eyam research group at Winchester, now also linked to the Mathematical Institute at Oxford University. The first is to re-imagine the Eyam equations in *stochastic* form, using statistical methods to predict how a given infection may spread, and comparing this to the *deterministic* methods of this and our previous paper. The second is to update a physical model of the Eyam system. We first reported in our previous paper [4] a Phillips MONIAC style system based upon a series of connected water vessels, pipes and pulleys. We are in the process of developing an improved version with our students and laboratory technicians. We are also in the process of developing a chemical analogue of the Eyam system via an auto-catalytic reaction. In addition, we have been exploring a stochastic variant ('Reed Frost') in using a physical implementation that is used in the teaching of epidemiology to medical students at Oxford University.


Eyam model fit

N=2542, I_{max}=323.3, R_0=1.85, t_{max}=1.6

α=2.84, ρ=1373, η=0.751, S_0=2467, I_0=34.06

2800

Imax: 498, 166, 323.263

tmax: 1.73, 0.575, 1.6049

alpha: 4, 2, 2.84

eta: 0.95, 0.05, 0.751

☑ Overlay S and D as well as I?

□ Use S,D?

□ 3D?

R0 = 1.8511

rho = 1373.3

N = 2542

beta = 0.0020681

SSD = 9424

Min time /months: -0.5

Max time /months: 4

0

EYAM MODEL. A.French Aug 2019.

Figure 17. An optimized $I(t)$ curve fit to the Liberia 2014–16 Ebola data using our Eyam model MATLAB GUI.

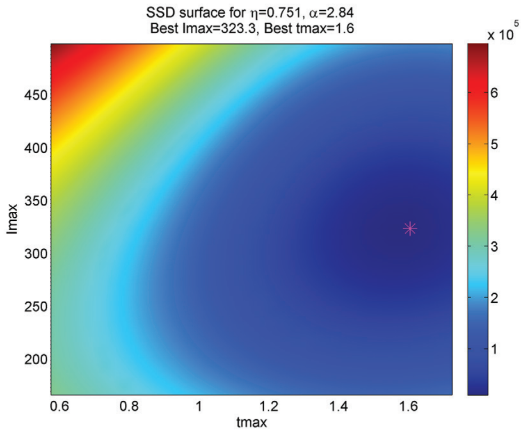


Figure 18. An SSD surface is formed, using the Liberia Ebola data, by varying $t_{\max}$ and $I_{\max}$ parameters, and fixing α, η.

Acknowledgments

JPC and AF would like to thank our employers at Winchester for encouraging this cross-curricular enterprise, and most importantly, our students. Significant contributions to this current paper have been from Dexter Poon and Alfie Baxter, both sixth form students at the time of writing.

JPC would like to thank all those responsible for enabling his recent sabbatical at the Mathematical Institute of the University of Oxford, particularly Dr Robin Thompson and Professor Dame Angela McLean, and indeed the Governing Body of Winchester College. RNT would like to thank Christ Church, Oxford, for funding this work via a Junior Research Fellowship.

Appendix A. Basic reproduction numbers

The pair of differential equations in section 2.1 for Susceptible $S(t)$ and Infectives $I(t)$ may be considered near the disease-free equilibrium position $(S_0, 0)$ by expanding with a Taylor expansion. Let:

$$\begin{aligned}\frac{\mathrm{d}s}{\mathrm{d}t} &= f(S, I)\\ \frac{\mathrm{d}I}{\mathrm{d}t} &= g(S, I).\end{aligned}$$

To first order in small displacements from the disease-free equilibrium:

$$\begin{aligned}\frac{\mathrm{d}s}{\mathrm{d}t} &= f(S_0, 0) + (S - S_0)\frac{\partial f}{\partial S} + (I - 0)\frac{\partial f}{\partial I}\\ \frac{\mathrm{d}I}{\mathrm{d}t} &= g(S_0, 0) + (S - S_0)\frac{\partial g}{\partial S} + (I - 0)\frac{\partial g}{\partial I}\end{aligned}$$

where the partial derivatives are evaluated at the equilibrium.

Now, $f(S_0, 0) = g(S_0, 0) = 0$, at the disease-free equilibrium, since dS/dt and dI/dt both equal zero at this point. If we represent the small displacements from the equilibrium by x_1 and x_2 we have:

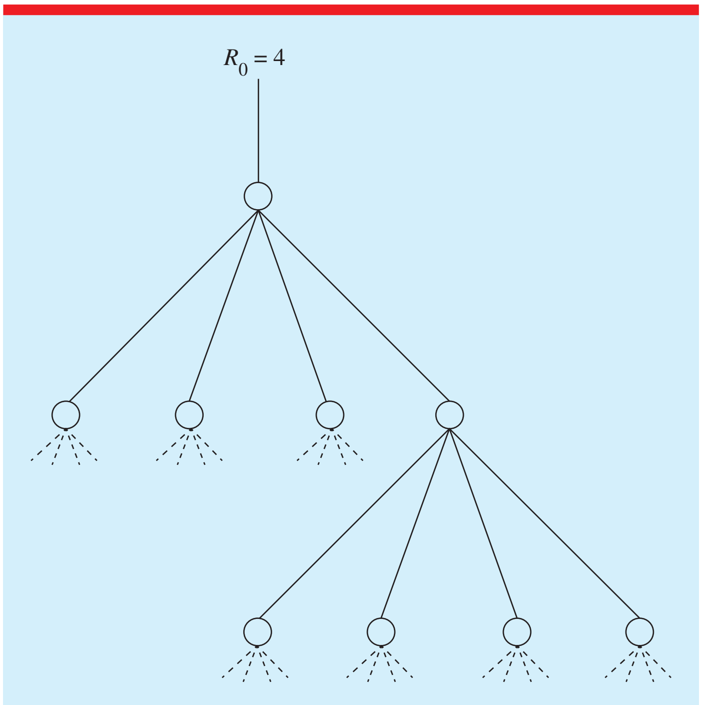


Figure 19. A Basic Reproduction Number of $R_0 = 4$ implies one infective infects four susceptibles. Each of these becomes infectives and then go on to inject four more susceptibles, etc. Until you run out of susceptibles!

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= \frac{\partial f}{\partial S}x_1 + \frac{\partial f}{\partial I}x_2 \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= \frac{\partial g}{\partial S}x_1 + \frac{\partial g}{\partial I}x_2 \end{aligned}$$

which, in matrix form with the values of the partial derivatives evaluated at the equilibrium inserted, becomes:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & -\beta S_0 \\ 0 & (\beta S_0 - \alpha) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

At this point it is informative for students to look at what this 'linearization' has achieved by considering the 1D case, noting the well-known differential equation $\dot{x} = kx$ has the *exponential* solution $x(t) = ce^{kt}$ where c is the initial value of $x = x(0)$. As time progresses, we see that the value, and the sign of k tells us that as t increases, x either moves towards $x = 0$ for $k < 0$, away from $x = 0$ for $k > 0$, or remains a constant for $k = 0$. This is a statement on the stability of the point $x = 0$: we see that we could classify $x = 0$ as a stable equilibrium in the case of $k < 0$ because of the tendency for the solution to return to the point. The opposite is of course true for the $k > 0$ case. So, extending this to our 2D situation and writing the components x_1 and x_2 as follows:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} c_1 e^{kt} \\ c_2 e^{kt} \end{pmatrix}$$

leads to the *eigenvalue* equation:

$$\begin{pmatrix} 0 & -\beta S_0 \\ 0 & (\beta S_0 - \alpha) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = k \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

with $\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ as the eigenvector and k the eigenvalue. This expression only has a non-trivial solution[13], if the determinant below is equal to zero:

$$\begin{vmatrix} -k & -\beta S_0 \\ 0 & (\beta S_0 - \alpha) - k \end{vmatrix} = 0$$

and so, we have the quadratic:

$$k\{k - (\beta S_0 - \alpha)\} = 0$$

So, $k = 0$ or $k = (\beta S_0 - \alpha)$. Here our 1D example gives us insight: if k is greater than zero, then the disease-free equilibrium is unstable, and there will be an increase in I, which of course corresponds

[13] The trivial solution being $\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

to an epidemic situation. We rearrange to see that the threshold condition may be represented as $\frac{\beta S_0}{\alpha} = 1$ with the quantity $\frac{\beta S_0}{\alpha} > 1$ representing the condition for an epidemic. This quantity is called the *Basic Reproduction Number* and denoted by the symbol R_0. A quick reminder of the meanings of the quantities here:

- α behaves like a decay constant for the death process, so $1/\alpha$ is the mean lifetime of an infective.
- β is an infection rate or probability of an individual infective infecting susceptibles per unit time.
- S_0 is the initial number of susceptibles available to be infected at the beginning.

The quantity $\frac{\beta S_0}{\alpha}$ is essentially:

$$\text{probability per time} \times \text{time} \times \text{number}$$

or the expected number of cases in that time. The diagram in figure 19 shows how this idea might be illustrated. Here $R_0 = 4$, i.e. each infective infects four susceptibles, and each of them go on to infect four more.

One should note that the meaning of R_0 described here is not unique. Certain epidemiological systems, e.g. those involving an additional vector population, may result in the 'branching ratio' illustrated in figure 19 depending on R_0^2 rather than R_0. As systems become more complicated and the fine details of the biology become important, there is a better chance of defining R_0 in terms of the details specific to the disease and environmental conditions.

Appendix B. MATLAB code for numerical evaluation of $\tau(z)$ integral

Two methods have been used to evaluate the integral:

$$\tau(z) = \int_0^z \frac{\mathrm{d}z'}{x_{\max} + 1 - e^{-z'} - z'}.$$

For students attempting this for the first time, and with access to MATLAB, we recommend the *Trapezium Rule* as this can be easily coded. The idea is to define an equally spaced range of z between its limits $z_\pm$ (1000 steps are used in this paper) and then evaluate:

$$f(z) = \frac{1}{x_{\max} + 1 - e^{-z} - z}.$$

We assume the area under $f(z)$ is the summed area of trapeziums formed by piecewise straight lines joining the $\{z, f(z)\}$ sample points.

The trapezium rule may be implemented in MATLAB via the following code:

```
%trapezium_rule
% Integration via the trapezium rule.
x is an equally spaced vector, and y
% are the corresponding integrand
values. The integration is from x(1)
to
% the values in x. Output vector I has
the same dimensions as x.
function I = trapezium_rule(x, y)
dx = x(2)-x(1);
I = dx*(cumsum(y) - y(1)/2 - y(end)/
2);
```

The built-in function `cumsum(y)` yields the cumulative sum of elements of a vector `y`. For example:

```
y = [1,2,3,4,5];
cumsum(y) = [1,3,6,10,15]
```

This can then be used by another function, which evaluates the integral $\tau(z)$. Note the negative and positive values of z must be evaluated separately.

```
%tor_of_z
% Evaluates numeric integral of
tor(z).
function [z,tor] = tor_
of_z(zminus,zplus, Ni, xmax)
dz = 0.01;
z_pos = linspace(0, zplus - dz, Ni);
z_neg = linspace(0, zminus + dz, Ni);
f_pos = 1./(xmax + 1 - exp(-z_pos)
- z_pos);
f_neg = 1./(xmax + 1 - exp(-z_neg)
- z_neg);
tor_pos = trapezium_rule(z_pos,
f_pos);
tor_neg = trapezium_rule(z_neg,
f_neg);
```

```
z = [fliplr(z_neg),z_pos];
tor = [fliplr(tor_neg),tor_pos];
```

Appendix C. Calculus with cubic-splines

Although much more of a ‘black-box’ approach, our preferred method of performing numerical integration (or indeed differentiation) from a set of $\{x, y\}$ data points is to firstly determine a *cubic spline* fit. This is the method used to create all the ‘semi-analytic’ curves in this paper.

In MATLAB, we have implemented the procedure into a higher-level function `integrator`. Although this looks simpler than the trapezium rule implementation, it masks the fact that plenty of code (and mathematics) is being done behind the scenes and therefore invisible to the student.

```
%tor_of_z_cspline
% Evaluates numeric integral of
tor(z) using cubic cpline fit
function [z,tor] = tor_of_z_
cspline(zminus,zplus, Ni, xmax)
dz = 0.01;
z = linspace(zminus + dz, zplus - dz,
Ni);
f = 1./(xmax + one - exp(-z) - z);
tor = integrator(z, f, 0, z);
```

MATLAB has several in-built functions to achieve integration via cubic splines, and the methods are readily applied using the code described in Hanselman & Littlefield’s *Mastering MATAB*. [10]. In the latter, spline coefficients associated with each data point in $\{x, y\}$ are used to determine the integral or differential of the spline, since this can be achieved trivially for a cubic.

A cubic spline itself, illustrated in figure 20, is a set of piecewise cubic curve segments, defined to pass through a ‘buffer’ of contiguous data points, such that the curves and their first derivatives are continuous at the boundaries between the segments. Splines also require a boundary condition relating to the second derivatives at the ends of the spline. A *natural spline* sets the second derivatives to be zero at these points [9] pp 116.

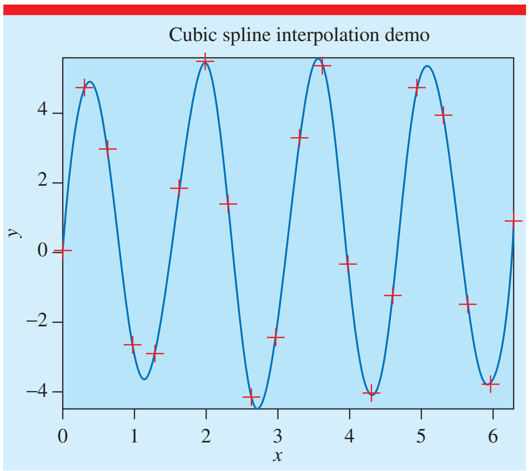


Figure 20. Demonstration of interpolation via cubic splines using MATLAB. The fitting of a piecewise cubic (with continuous first derivatives between cubics) is not just useful for smoothing data. The cubics can also be readily differentiated or integrated, and hence enable a calculus method to be applied to a set of x, y data points. The code to generate the graph above is:

```
%Fontsize for graphs
fsize = 22;

%Define x,y coordinates
N = 20; x = linspace(0, 2*pi,N);
y = 5*sin(4*x) + rand(size(x));

%Define spline interpolation
xx = linspace(0, 2*pi,1000);
yy = interp1(x,y,xx, 'spline');

%Plot spline and print a 300dpi PNG
file
plot(xx,yy,'b',x,y,'r+',
'markersize',...
18,'linewidth',2);
set(gca,'fontsize',fsize); grid on;
axis tight
xlabel('x','fontsize',fsize);
ylabel('y','fontsize',fsize);
title('Cubic spline interpolation
demo','fontsize',fsize);
print(gcf,'cspline demo.
png','-dpng','-r300');
```

ORCID iDs

J P Cullerne https://orcid.org/0000-0001-5263-0715

A French https://orcid.org/0000-0002-0902-9729

R N Thompson https://orcid.org/0000-0001-8545-5212

Received 30 August 2019
Accepted for publication 2 October 2019
https://doi.org/10.1088/1361-6552/ab4a59

References

[1] Kermack W O and McKendrick A G 1927 A contribution to the mathematical theory of epidemics *Proc. R. Soc. Lond.* A **115** 700–21

[2] Kendall D G 1956 Deterministic and stochastic epidemics in closed populations *Proc. of the 3rd Berkeley Symp. on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Health* (Berkeley, CA: University of California Press) pp 149–65

[3] Gani J 1967 On the general stochastic epidemic *Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability, Volume 4: Biology and Problems of Health* (Berkeley, CA: University of California Press) pp 271–9

[4] French A, Kanchanasakdichai O and Cullerne J P 2019 The pedagogical power of context: iterative calculus methods and the Epidemiology of Eyam *Phys. Educ.* **54** 045008

[5] French A, Cullerne J P and Kanchanasakdichai O 2019 Numerical methods as an introduction to calculus *Phys. Educ.* **54** 045009

[6] World Health Organization 2014 *Ebola Response Roadmap Situation Report: 29 October 2014*

[7] Rachah A and Torres D F M 2015 Mathematical modelling, simulation and optimal control of 2014 Ebola Outbreak in West Africa *Discrete Dyn. Nat. Soc.* **2015** 842797

[8] BMJ 2014;349:g7348 (*Inforgrahpic: Ebola: A clinical guide*) (https://doi.org/10.1136/bmj.g7348)

[9] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 2003 *Numerical Recipes in C++. The Art of Scientific Computing* 2nd edn (Cambridge: Cambridge University Press)

[10] Hanselman D and Littlefield B 2001 *Mastering MATLAB* 6 (Upper Saddle River, NJ: Prentice Hall) pp 292–4

[11] Thompson R N, Gilligan C A and Cunniffe N J 2016 Detecting pre-symptomatic infection is necessary to forecast major epidemics in the earliest stages of infectious disease outbreaks *PLoS Comput. Biol.* **12** e1004836

[12] Thompson R N and Hart W S 2018 Effect of confusing symptoms and infectiousness on forecasting and control of Ebola outbreaks *Clin. Infect. Dis.* **67** 1472–4

[13] Hart W, Hochfilzer L, Cunniffe N, Lee H, Nishiura H and Thompson R N 2019 Accurate forecasts of the effectiveness of interventions against Ebola may require models that account for variations in symptoms during infection *Epidemics* (https://doi.org/10.1016/j.epidem.2019.100371)

[14] Faber T E 1997 *Fluid Dynamics for Physicists* (Cambridge: Cambridge University Press) p 33

[15] WHO Ebola Response Team 2014 Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections *New Engl. J. Med.* **371** 1481

[16] Sullivan N *et al* 2003 Ebola virus pathogenesis: implications for vaccines and therapies *J. Virol.* **77** 9733–7

Dr J P Cullerne is currently the Undermaster at Winchester College, and is a former Housemaster and Head of Physics. His is the co-author of *The Language of Physics*.

Dr A French has taught Physics, Mathematics and Computer Programming at Winchester College since 2011. He was previously a Radar Systems Engineer and was educated at Cambridge University and University College London.

Dr R N Thompson is a Junior Research Fellow at Christ Church, University of Oxford. His research involves mathematical modelling of infectious disease epidemics in populations of humans, animals and plants.