Variational data assimilation with epidemic models

C.J. Rhodes, T.D. Hollingsworth

www.elsevier.com/locate/yjtbi

Variational data assimilation with epidemic models

C.J. Rhodes, T.D. Hollingsworth

PII: S0022-5193(09)00079-4
DOI: doi:10.1016/j.jtbi.2009.02.017
Reference: YJTBI 5473

To appear in: *Journal of Theoretical Biology*

Received date: 24 September 2008
Revised date: 28 January 2009
Accepted date: 19 February 2009

Cite this article as: C.J. Rhodes and T.D. Hollingsworth, Variational data assimilation with epidemic models, *Journal of Theoretical Biology* (2009), doi:10.1016/j.jtbi.2009.02.017

Variational Data Assimilation with Epidemic Models

C. J. Rhodes[*1, 2] and T. D. Hollingsworth[2]

[1]*Institute for Mathematical Sciences, Imperial College London, 53 Prince's Gate, Exhibition Road, South Kensington, London, SW7 2PG, United Kingdom.*
[2]*MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London, W2 1PG, United Kingdom.*

c.rhodes@imperial.ac.uk Telephone: +44(0)2075941753
d.hollingsworth@imperial.ac.uk Telephone: +44(0)2075943217

* Author for correspondence:

Abstract

Mathematical modelling is playing an increasing role in developing an understanding of the dynamics of communicable disease and assisting the construction and implementation of intervention strategies. The threat of novel emergent pathogens in human and animal hosts implies the requirement for methods that can robustly estimate epidemiological parameters and provide forecasts. Here, a technique called variational data assimilation is introduced as a means of optimally melding dynamic epidemic models with epidemiological observations and data to provide forecasts and parameter estimates. Using data from a simulated epidemic process the method is used to estimate the start time of an epidemic, to provide a forecast of future epidemic behaviour and estimate the basic reproductive ratio. A feature of the method is that it uses a basic continuous-time *SIR* model, which is often the first point of departure for epidemiological modelling during the early stages of an outbreak. The method is illustrated by application to data gathered during an outbreak of influenza in a school environment.

Keywords: epidemic model; parameter estimation; data assimilation, forecasting.

1. Introduction

Mathematical models of infectious disease epidemics are important tools for assessing the impact of communicable disease in both human and animal (wild and domesticated) populations (Anderson and May, 1991). At the most basic level they are a concise means of quantitatively representing the essential epidemiological and biological factors that relate to a particular disease pathogen in a given population. More sophisticated implementations can be used to assist in the design or evaluation of actual or prospective intervention programmes, such has been done for childhood disease immunisation (Grenfell *et al*., 2001; Jansen *et al*., 2003), analysing the spread and containment of BSE and foot-and-mouth disease in the United Kingdom (Anderson *et al*., 1996; Keeling *et al*., 2003), investigating the impact of HIV(Anderson and May, 1991), and in assessing the role of badgers in the on-going bovine TB epidemic in south-west England for example (Donnelly *et al*., 2006).

In common with mathematical modelling in other disciplines there is a research imperative to develop ever more realistic epidemic models and simulations that include higher-fidelity representations of underlying biological or population processes. However, more realistic models tend to be more complex and they are frequently populated by a proliferation of parameters that need to be robustly estimated (Riley *et al*., 2003; Ferguson *et al*., 2005; Longini *et al*., 2005). In practice, the degree of model sophistication that is used in a given situation often reflects a judicious balance between the questions the model is required to address and the ability to reliably estimate parameters.

When an unknown or poorly understood pathogen suddenly emerges in a population the challenges of epidemic modelling are exacerbated because much of the quantitative work has to be performed on a fast time-scale, usually with limited or poor quality data. There is a risk that the epidemiological dynamics might outpace the modelling response if too sophisticated an approach is adopted. One of the immediate tasks faced by policy advisors and modellers is determining the epidemic curve; specifically, the initial requirement is to estimate model parameters, produce a forecast for the duration of the epidemic and provide an estimate of the expected proportion of the population to be infected by the disease. In these circumstances the

simplest epidemic models are generally used because it is pointless to compound epidemiological uncertainties with those generated by inappropriate use of over-complex and poorly parameterised models. Fitting epidemic models to data is a topic of ongoing research. Anderson and May (1991) and Daley and Gani (1999) give accounts of commonly used methods. Wearing *et al*. (2005), Ferrari *et al*. (2005) and Fraser (2005) provide a discussion of many of the challenges faced when attempting to estimate the basic reproductive ratio from epidemiological data.

Motivated by the requirement for robust methods for parameter estimation and the need to derive full predictive benefit from the most basic of epidemic models in the early stages of an outbreak we present a new method for the determination of epidemiological parameters and for the subsequent production of epidemic forecasts. This method is one that has hitherto been used in weather and climate modelling, and it is known as variational data assimilation (VDA) (Bouttier and Courtier, 1999; Huang and Yang, 1996). A version has also been developed for predator-prey systems (Lawson *et al.*, 1995). VDA has been developed to optimally combine a dynamical model with observations of the system to produce accurate forecasts, and it can be readily adapted to epidemiological applications.

In this paper we introduce the concept of data assimilation and show how it can be adapted and applied to epidemic modelling and infectious disease management. In Section 2 we review the variational data assimilation procedure, and in Section 3 it is applied to a continuous-time *SIR* epidemic model. In Section 4 we introduce some simulated data (with errors) for a basic underlying epidemic process and show how the assimilation can be carried out using a model and data. Section 5 illustrates the use of assimilation to estimate the basic reproductive ratio R_0 of the epidemic that gave rise to the data. In Section 6 the method is applied to an outbreak of influenza in a school setting.

2. Variational Data Assimilation

Exploiting dynamical models that are constrained by observations or measurements to produce a prediction of future system behaviour is an issue that is of on-going interest to weather and climate forecasters. The challenges faced when attempting to predict the future behaviour of an epidemic and when forecasting the weather have much in common. Observations and data (with their associated uncertainties) need to be combined with (often non-linear) dynamical models in order to produce an estimate of future behaviour, i.e. a forecast. In weather forecasting the dynamical models are based on the underlying fluid mechanics of the atmosphere. In epidemic modelling there is clearly no underlying physics governing the dynamics, so simpler models that are known to reflect typical epidemic behaviours have to be used.

Variational data assimilation is an iterative technique in which the difference between observations of a system at given time points and the initial states of the model is minimised. Assuming that all the model parameters are known, the result of VDA is the initial condition that that generate a best-fit of the model to the observations. In the next Section we shall show how VDA can also be used to estimate model parameters, but in what follows we assume that all model parameters are known to make the presentation straightforward. Our presentation is based on that given in Huang and Yang (1996).

2.1 Model states, observations and cost function

Typically, a dynamical model integrates an initial state of a number of field variables, $w(i)$, forward in time. In an epidemic case these fields $w(i)$ could represent the numbers of Susceptibles (S), Infectives (I) and Recovered (R) individuals in an *SIR* model, for example. Let us define the initial state of the fields at time $t = 0$ as $\overline{w}_0$, and the fields at some future time t to be $\overline{w}_t$. In most dynamical models there are several fields of interest, so, assuming there are three fields:

$$\overline{w} = \begin{pmatrix} w(1) \\ w(2) \\ w(3) \end{pmatrix} \tag{2.1}$$

The dynamical model, which henceforth we shall call the forward model, connects the state of the system at time *t* with the state at time *t+1*, i.e.

$$\overline{w}^{f}_{t+1} = M\left(\overline{w}^{f}_{t}\right) \tag{2.2}$$

Where *M* represents the forward model that takes the current state to the state at the next time step, and the superscript *f* refers to the forward model.

Starting at the initial state $\overline{w}^{f}_{0}$, the state of the model at any time *t* is given by repeated application of the forward model, so

$$\overline{w}^{f}_{t} = M.....M\left(\overline{w}^{f}_{0}\right) \tag{2.3}$$

Usually we also have some observations (usually with uncertainty attached) of the system at certain time points, so the objective of the assimilation process is to determine the initial conditions that are consistent with these observations. To do this we define a cost function given by

$$J = \frac{1}{2}\sum_{t}\left(\overline{w}^{f}_{t} - \overline{w}^{o}_{t}\right)^{T}\left(\overline{w}^{f}_{t} - \overline{w}^{o}_{t}\right) \tag{2.4}$$

Where $(...)^{T}$ is the transpose operator and $\overline{w}^{o}_{t}$ are the observations of the fields at a given time, *t*. The summation is over only those time-points where data is available. The cost function can be defined in a variety of ways, depending upon the application, and the elaborations are discussed in more detail in Huang and Yang (1996) and Daley (1991).

We now seek a minimisation of J with respect to the initial conditions, i.e. we require

$$\nabla J\left(\overline{w}_0^f\right) \to 0 \tag{2.5}$$

We will do this using the adjoint method. The advantage of this method is that it generates an exact expression for the cost function gradient which can then be used in a computationally efficient minimisation procedure. A simpler alternative would be to estimate the cost function gradient by perturbing each of the fields in turn. In practice this is computationally intensive (particularly for models with many fields) and it often yields an approximation to the gradient that fails to converge to a minimum. Whichever method is used, the result is a set of initial conditions for the field variables that minimises the cost function, equation 2.4.

The detailed calculations for how the cost function gradient can be calculated from the adjoint method, and then minimised, are presented in Appendix A.

3. Application of VDA to an Epidemic Model

3.1 A basic epidemic model

The foundation of the majority of epidemic models is the Susceptible-Infectious-Recovered (*SIR*) compartmental model (Bailey, 1957; Anderson and May 1991). Despite its simplicity this basic formulation, cast in either deterministic or stochastic form, has provided a wealth of insight into the dynamics of many different transmissible diseases in a variety of population types. The structural simplicity and ease of parameterisation make the continuous-time *SIR* model the first point of departure when modelling epidemic outbreaks of communicable disease. It is straightforward to adapt this model to accommodate refinements, such as adding exposed-but-not-infectious compartments, seasonality in contact rate or age structure for example. For a population of size N ($=S+I+R$), the basic *SIR* model is defined as follows:

$$\frac{dS}{dt} = -\beta SI \tag{3.1}$$

$$\frac{dI}{dt} = \beta SI - \sigma I \tag{3.2}$$

$$\frac{dR}{dt} = \sigma I \tag{3.3}$$

where β is the contact rate and σ^{-1} is the duration of infectiousness. It is straightforward to show that the basic reproductive ratio of the pathogen R_0 is given by $\beta S_0 / \sigma$.

In Section 2 (and Appendix A) the framework of the variational data assimilation method was assembled with no reference to any particular model or system of interest. Moving from the general to the specific, here we show how the method can be used to derive expressions for the cost function gradient for a basic *SIR* epidemic model. In so doing the means by which VDA can be applied to real systems becomes clearer.

3.2 Derivation of the tangent linear model and adjoint model

The first step is to write down the adjoint model, which is derived via the intermediate tangent linear model of the forward model. Equations 3.1-3.3 represent the forward model for the epidemic process. From equation 1, we define the *S*, *I* and *R* fields to be $w(1)$, $w(2)$ and $w(3)$.

To formulate the tangent linear model, we apply the Jacobian in equation A5, giving

$$\frac{dw(1)^{tl}}{dt} = -\beta w(2) w(1)^{tl} - \beta w(1) w(2)^{tl} \tag{3.4}$$

$$\frac{dw(2)^{tl}}{dt} = \beta w(2) w(1)^{tl} + \left(\beta w(1) - \sigma\right) w(2)^{tl} \tag{3.5}$$

$$\frac{dw(3)^{tl}}{dt} = \sigma w(2)^{tl} \tag{3.6}$$

We can now write the equation for the tangent linear model in matrix form

$$\begin{pmatrix} w(1)^{tl} \\ w(2)^{tl} \\ w(3)^{tl} \\ \frac{dw(1)^{tl}}{dt} \\ \frac{dw(2)^{tl}}{dt} \\ \frac{dw(3)^{tl}}{dt} \end{pmatrix} = L \begin{pmatrix} w(1)^{tl} \\ w(2)^{tl} \\ w(3)^{tl} \\ \frac{dw(1)^{tl}}{dt} \\ \frac{dw(2)^{tl}}{dt} \\ \frac{dw(3)^{tl}}{dt} \end{pmatrix} \tag{3.7}$$

where the tangent linear operator is given by

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -\beta w(2) & \beta w(1) & 0 & 0 & 0 & 0 \\ \beta w(2) & (\beta w(1) - \sigma) & 0 & 0 & 0 & 0 \\ 0 & \sigma & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.8}$$

Recalling equation A11, the tangent linear model can be used to derive the adjoint model by the transpose operation. Denoting the variable of the adjoint model by $w(i)^{ad}$ for $i = 1,2,3$ and $dw(i)^{ad}/dt$ for $i = 1,2,3$ the matrix form of the adjoint model can be written as

$$\begin{pmatrix} w(1)^{ad} \\ w(2)^{ad} \\ w(3)^{ad} \\ \dfrac{dw(1)^{ad}}{dt} \\ \dfrac{dw(2)^{ad}}{dt} \\ \dfrac{dw(3)^{ad}}{dt} \end{pmatrix} = L^T \begin{pmatrix} w(1)^{ad} \\ w(2)^{ad} \\ w(3)^{ad} \\ \dfrac{dw(1)^{ad}}{dt} \\ \dfrac{dw(2)^{ad}}{dt} \\ \dfrac{dw(3)^{ad}}{dt} \end{pmatrix} \tag{3.9}$$

Where

$$L^T = \begin{bmatrix} 1 & 0 & 0 & -\beta w(2) & \beta w(2) & 0 \\ 0 & 1 & 0 & \beta w(1) & (\beta w(1) - \sigma) & \sigma \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.10}$$

The adjoint model can therefore be written as

$$w(1)^{ad} = w(1)^{ad} - \beta w(2)\, dw(1)^{ad} / dt + \beta w(2)\, dw(2)^{ad} / dt \tag{3.11}$$

$$w(2)^{ad} = w(2)^{ad} - \beta w(1)\, dw(1)^{ad} / dt + (\beta w(1) - \sigma)\, dw(2)^{ad} / dt + \sigma w(2)\, dw(3)^{ad} / dt \tag{3.12}$$

$$w(3)^{ad} = w(3)^{ad} \tag{3.13}$$

By integration of equations 3.11-3.13, this model is used to derive the gradient of the cost function using equation A10. Recalling from equation A11 that the time ordering of the adjoint model is reversed, it is necessary to integrate the adjoint model from $t = t_{\max}$ to $t = 0$ and initialise the model with the difference between the observation recorded at a given time *t* and the value of the forward model at that time i.e. $\left(\overline{w}_t^{f} - \overline{w}_t^{o} \right)$. The gradient of the cost function is the result at time *t*=0 of a single integration of the adjoint model.

Using the adjoint method to perform the assimilation is straightforward. The objective of the calculation is to use the model and observational data to calculate a plausible set of initial conditions for the epidemic model fields *S*, *I* and *R* at some determined point prior to the first observational data point (t=0). In practice it is therefore necessary to first make an estimate of the initial conditions of *S*, *I* and *R*, i.e. $\left(\overline{w}_0^f\right)$ at t=0. Using these estimated initial conditions the forward model (equations 3.1 – 3.3) is integrated forward in time from the time of the estimated initial conditions to a desired time-point beyond the latest data observation (i.e. to the point of furthest forecast), say t_{max}. The adjoint model (equations 3.11 – 3.13) is now integrated backward in time from the forecast point t_{max} to the time when the initial conditions were estimated, t=0. (Note: from equations 3.11-3.13 the forward model fields are used in the backwards integration of the adjoint model and the values of the observational data are used to adjust the adjoint model fields at the time of each observation). The values of the three adjoint model fields reported at the end of the adjoint model integration are then used to incrementally adjust the initial conditions of the forward model (using equation A12). The cost function (equation 2.4) relating the forward model to the observations is then calculated. The forward model is then integrated forward in time using the adjusted initial conditions. Once again, the adjoint model is integrated backward and the process is repeated until there is a minimisation of the cost function, *J*. At this point the assimilation process stops. The result is a set of initial conditions at t=0 that are consistent with the choice of forward model and the observational data points. Code that can be used to perform a basic VDA on an *SIR* model is provided in the Supplementary Online Material.

In what follows we will show how VDA can be applied to realistic data. In Section 4 we use VDA to assimilate a forward (*SIR*) model (with known epidemiological parameters) with noisy observations to find an appropriate set of initial conditions, and then use those to produce a forecast. In Section 5 we relax the assumption of known epidemiological parameters and extend the assimilation method to estimate one forward model parameter (in this case R_0) and initial conditions. Finally, in Section 6, we demonstrate the method by applying it to a well-known influenza epidemic curve.

4. Assimilation Using Model Data

One of the principal motivations for modelling epidemic outbreaks is to use simple models to provide insight into the epidemiological characteristics of the outbreak, and this can be particularly challenging if the outbreak is due to a novel or poorly understood infectious pathogen. In this section we will show how the VDA method can be used to meld observations with a dynamic model to provide estimates of initial conditions and then to provide predictions of future epidemic behaviour. To illustrate the method we will use data generated from a modelled epidemic. Additional errors of known magnitude were added to the epidemic curve to simulate fluctuations or errors in reporting.

4.1 Data

We assume that we have a communicable disease that confers lifetime immunity following infection. Using the *SIR* equations 3.1-3.3 we define a population size $N = 3000$, a mean period of infectiousness σ^{-1} of 5 days and a basic reproductive ratio $R_0 = 4$. The resulting epidemic curve is shown in Figure 1, and this assumes there is no previous exposure to the pathogen in the population and that the infection is triggered by the arrival of a single index case.

To simulate stochasticity of transmission and errors in daily reporting we add (or subtract) a uniformly distributed random deviate (Press *et al.*, 1992; pp267-273) with a fixed maximum percentage error. The raw data that might therefore be available to epidemiologists as a consequence of this outbreak are presented in Figure 2. Often the infectious cases only begin to be reported some time after the epidemic has taken hold and it becomes apparent to public health services that there is an epidemic occurring. The time at which the epidemic began will probably be unknown and the case reporting may be quite infrequent and almost certainly have errors (particularly in the early epidemic phase). In Figure 2 we have assumed that the epidemic was first noticed on Day 10 with first case reports on Day 12, followed by reports on Days 14 and 16. A random error of $\pm 16\%$ is assumed for the case reporting to replicate stochastic transmission and reporting error. Note that the data used here represents

prevalence of infection, whereas often it is disease incidence that is reported. In Section 6 an example of an epidemic where data on disease prevalence was reported is analysed.

4.2 Estimating initial conditions

The first task is to estimate the values of the model fields around the time of the start of the reported epidemic on Day 10. For the purposes of the calculation in this section we assume that we know what disease we are dealing with, and that we know the values of the epidemiological parameters R_0 and σ. Looking at the observed data for the number of Infectives in Figure 2 (triangles only), on the basis of a backwards linear extrapolation, it might be expected that there are likely to be around 700 Infectives on Day 10. In most epidemic outbreaks the extent of previous exposure to the pathogen is unknown, so we generally don't have detailed information on the precise proportion of Susceptible and Recovered individuals in a population, so we have to estimate these. Given that we are in the early stages of the epidemic it is plausible to suggest that we might have several hundred Recovered, so we assume 200. Given that the total population is 3000, we therefore have 3000-700-200 = 2100 Susceptibles on Day 10.

Using the assimilation scheme described in Section 3 we can now refine these estimates to find the set of conditions on Day 10 that are consistent with the *SIR* model and the data observations on Days 12, 14 and 16. The assimilation method gives the *S*, *I*, and *R* initial conditions for Day 10 to be 2633, 278 and 187. These initial conditions are shown in Figure 3a and are good estimates of the epidemiological state of the population on Day 10. Using assimilation we have gone from $(2100;700;200)_{estimate} \rightarrow (2633;278;187)_{assimilated}$. Note that the sum of all the epidemiological classes is not 3000 for the assimilated state. This is expected as we have used data that has error associated with it to make the calculation. The data in Figure 2 is just one realisation of a noisy epidemic process. To check that the assimilation is robust against different realisations, and ensure that the quality of agreement in Figure 3a is not just fortuitous, we repeat the assimilation for three

further realisations of the epidemic process also at the 16% error level. This generates three different observational data sets on Days 12, 14 and 16 that can then be assimilated. Repeating the assimilation for these data sets generates the initial conditions shown in Figure 3b-3d. It can be seen that the assimilation process performs well for each realisation and is robust to noise.

In Figure 3, there was an implicit assumption that there was little previous exposure to the pathogen, so the numbers in the Recovered class were low at the start of the epidemic. This assumption matched the epidemiological dynamics that were represented in the underlying data. For an infection such as chickenpox in humans, for example, it is possible that the population may already contain significant numbers in the Recovered class and it might be expected that there are significant numbers of Recovered individuals in the population. If we now estimate that there are, say, 1500 Recovered, 700 Infected and 800 Susceptibles on Day 10, the VDA initial conditions again identify the state $(800;700;1500)_{estimate} \rightarrow (2633;278;187)_{assimilated}$. This result suggests that the assimilation is robust to the estimate of the initial conditions. To test the sensitivity of the assimilation to this estimate the assimilation was repeated for several different initial conditions for each of the observational data sets underlying Figures 3a-3b. The results are show in Table 1. For each data set the result of the assimilation using the initial estimate in the first column is shown and it can be seen that the assimilation iterates to the same set of initial conditions regardless of the choice of the first estimate of the initial conditions. More exhaustive testing (not shown) has not revealed any counter-examples. There are four separate realisations of the underlying epidemic process represented in Table 1, so each realisation has a slightly different assimilated set of initial conditions. This is inevitable as each data set has different numbers of Infectives at Days 12, 14 and 16 due to the noise; however, for each of the realisations the results are comparable and they also compare favourably with the (noise free) state of the epidemic at Day 10. Table 2 shows how the assimilation from a given initial estimate performs under varying levels of noise. As the noise increases, the assimilation becomes inevitably less precise, but nevertheless within the error bounds of the true epidemic state on Day 10.

So far we have used three observations of the epidemic on Days 12, 14 and 16, to estimate the initial epidemiological conditions on Day 10. It is possible attempt the same type of calculation using just a single observation of the epidemic. Let us assume that we have a single observation of the number of Infectives on Day 14 to be the 964 recorded in Figure 2. For two very different estimated initial conditions, using this single observation, we obtain:

$(2100;700;200)_{estimate} \rightarrow (2391;294;245)_{assimilated}$ and

$(1500;700;800)_{estimate} \rightarrow (2363;297;250)_{assimilated}$.

The assimilated initial conditions are comparable with those in the calculation for Figure 3a where, using three days data we found,

$(2100;700;200)_{estimate} \rightarrow (2633;278;187)_{assimilated}$. Therefore, using only a single observation of the number of infectives (and very different estimates for the prior exposure of the population) we see that it is possible to assimilate to plausible initial conditions for Day 10.

A closely related issue to that of estimating initial conditions is estimating when an epidemic started. Returning to the three observations of the number of infectives in Figure 2 it is not obvious when the epidemic began, as we simply have three data points each separated by two days. Estimating when the epidemic began is equivalent to finding the time at which the assimilated calculation for initial number of infectives is close to unity. The assimilation calculation for Day 10 can be repeated for Days 8, 6, 4 and 2. The result for the initial number of Infectives at each of these days is shown in Figure 4. It is clear that the assimilation method is able to determine the start of the epidemic to within a day or so. It should be remembered that the continuous model used here for the forward model (i.e. that permits non-integer numbers of individuals in the *S*, *I* and *R* compartments) may not be good approximations to epidemic dynamics when the number of infectives is small or the host population is small. Stochastic effects can play a significant role in such circumstances, and this can be an issue when estimating the start-time of an epidemic.

4.3 Epidemic forecasts

In most epidemiological analyses of disease outbreaks a particularly valuable insight that modelling can deliver is the ability to make projections or forecasts of the future development of the epidemic. These predictions can be useful in assisting the formulation of disease intervention strategies.

Returning to the data set in Figure 2, the task is to use the dynamical model and the available data in order to provide a forecast. There are a number of different ways in which assimilation can be used continuously in real-time to provide continually updated forecasts as data continues to become available (Bouttier and Courtier, 1999). Here, for the purposes of illustrating the application, we show a basic single-step forecast.

By assimilating the data and model from Day 12 when the first observation is available and onwards, the output of VDA will be the initial conditions that are consistent with the choice of forward model (i.e. the *SIR* model in this case) and the observed data. The initial conditions for Day 12 go as $(2000;800;200)_{estimate} \rightarrow (2066;663;368)_{assimilated}$, and the forecast the results is shown in Figure 5, and the agreement between the forecast and the underlying epidemic process that generated the data is good.

For a more prolonged epidemic outbreak, a continual process of assimilation could be used whereby recent data is assimilated with the forward model to continually update the forecast in a sequential fashion (Bouttier and Courtier, 1999). Work is currently underway to demonstrate this using real data.

5. Parameter Estimation

In the development and application of VDA to an epidemiological context we have heretofore assumed that all the parameters of the forward model are known. In the *SIR* forward model this amounts to knowing the basic reproductive ratio, R_0 of the

pathogen and the average period of infectiousness, σ. Often these quantities are known or can be estimated. This is particularly true when the outbreak is triggered by a known pathogen, such as measles or chickenpox in humans, for example (Fraser, 2007). However, in many cases of interest, when an unknown or novel pathogen is implicated (e.g. SARS, pandemic influenza, smallpox) whilst it is often possible to get estimates for the within-host parameters (such as period of infectiousness or incubation period) it can be much more difficult to estimate the reproductive ratio (Ferrari *et al*., 2005; Gani and Leach, 2001; Ferguson *et al*., 2003; Anderson *et al*., 2004). In this Section we show how VDA can be used to provide an estimate of the basic reproductive ratio R_0.

Looking at the data in Figure 2, there are two difficulties with estimating R_0; there are large uncertainties in the recorded data *and* the epidemic was well under way by the time data started being recorded. A widely used method to estimate R_0 is to look at the phase of exponential increase in the number of Infectives during the early stage of the outbreak and calculate the basic reproductive ratio from $R_0 = 1 + r\sigma^{-1}$, where r is the rate of increase of infectives (Appendix B). Applying this exponential analysis to our raw data (*i.e.* the triangles on Figure 2, and recalling that it was based on an underlying epidemic process with $R_0 = 4$) we find $R_0 = 1.8$, i.e. a significant underestimate. Repeating this but now using data from the underlying model epidemic process (*i.e.* fitting an exponential to the data points, but now without any added noise) we find $R_0 = 2.3$, indicating that there are too many Infectives around to allow this approach to be mathematically valid - in this case we have been alerted to the epidemic too late.

An alternative strategy is to use VDA to compare the quality of fits of models with different values of R_0. The assimilation process produces a set of initial conditions for the forward model that minimises the cost function given by equation 2.4. It is the quality of this fit, as reflected by the asymptotic value of the cost function after much iteration, which can be used to estimate R_0. Within the forward model (equations 3.1-3.3) we are free to specify the values of the parameters. Assuming that observations of infectious cases has provided an estimate for the mean duration of infectiousness

(σ^{-1}) we can undertake assimilation of the forward model and the data using models with different contact rates (β), i.e. different values of R_0.

Let us assume that our initial estimate for $R_0 = 6$. It is straightforward to use the forward model and the data in Figure 2 to assimilate to a set of initial conditions. As before, we assimilate to Day 10 from an initial condition estimate of $(2100; 700; 200)_{estimate}$. Figure 6 shows the value of the cost function for an increasing number of iteration cycles (n). It can be seen that the cost function iterates to a stable asymptotic value. This asymptotic value reflects the quality of the assimilation of the model and the data, and in this case it is 139324. It is straightforward to repeat this for a number of different values of R_0 in the forward model. A plot of the asymptotic value of the cost function as a function of R_0 in the forward model is shown in Figure 7a. There is a minimum in the vicinity of $R_0 \simeq 3.5$, suggesting that this value gives the best quality assimilation of data and forward model. Given the paucity of available data points and the error levels in recording the number of Infectives, the estimate for R_0 is good. In practice it would be appropriate to use this value of R_0 for assimilation and forecasting purposes.

For the purposes of comparison, an estimate of R_0 was made using the same data from Figure 2 (i.e. the sequence of three observations on alternate days of the numbers of infectives, as denoted by the triangles), but this time using a standard least squares fitting method for fitting dynamical models and data. This was done using the Berkeley Madonna analysis package, which is a widely-used data analysis and graphing package for dynamical systems modelling. The details of the procedure that was used are presented in Appendix C. Fitting the three data points this way (using a partially constrained fit, see Appendix C) yielded a value of $R_0 \simeq 2.3$, which is a poor estimate. Figure 7b shows the comparison between the fit using data assimilation (black line) with the estimated R_0 of 3.5 and the fit using the standard least squares method (grey line). The dots on Day 10 and Day 22 are the number of infectives of the underlying epidemic process. It can be seen that as well as giving a better estimate for R_0 the assimilation method gives a better estimate of the initial conditions (Day

10) and a better forecast of the future number of infectives (Day 22). In order to present a detailed comparison of the performance of the two methods we have concentrated on analysing one realisation of the noisy process underlying the epidemic data (data from Figure 2). However, the method performs equally well for other realisations (not shown) and, as discussed in Section 4, the assimilation result shown here is insensitive to the estimate of initial conditions.

To test the robustness of the estimate of R_0 as the error in the prevalence reporting changes, the fitting exercise was repeated for a number of different error levels. The results are shown in Figure 8. Even up to 50% error in reporting, there is still a discernible minimum giving an estimate of R_0 of ~3.

There are other, more refined, methods for estimating R_0 from data and this comparison does not imply a definitive advantage in using assimilation for parameter estimation. However, it does indicate that VDA is capable of estimating parameters with some accuracy, and is a useful way of combining observations with a basic model to provide such estimates.

6. Application to an Outbreak of Influenza

In order to illustrate how assimilation might be used in a real application we apply the method to data recorded during an outbreak of influenza in a boarding school (Anonymous, 1978). Following initiation of the epidemic, disease prevalence (as reflected by the number of cases confined to hospital) was recorded over a two week period in a closed population of 763 individuals. Here we will use data assimilation with an *SIR* model to estimate the basic reproductive ratio. This analysis is not intended to be detailed or exhaustive; rather, the motivation is to show how the method could be used in practice and to show that plausible results are obtained. Future work will address applications of assimilation to empirical data sets.

On Days 2, 3 and 4 of the epidemic the reported number of infectives was 6, 25 and 76. The numbers of recovered individuals are not recorded, so we assume that on

these Days the numbers were 3, 12 and 40. The resulting numbers of susceptibles are, therefore, 754, 726 and 647. For the forward model we use an *SIR* model (equations 3.1 – 3.3) with $\sigma = 0.5$ days^{-1} (i.e. assuming a 2 day infectious period). Given the small scale of the outbreak in a closed community we do not ascribe any errors to the data reporting.

Using the techniques described in Section 5 we can assimilate from an initial estimate to set of initial conditions that are consistent with the model and data, whilst using the asymptotics of the cost function as a means of estimating R_0. Our estimate for the initial conditions for (*S*; *I*; *R*) are (755; 5; 3) and we assume that the epidemic begins two days before Day 2. Assimilation using the model and assumptions just stated gives a basic reproductive ratio $R_0 \simeq 3.5$ with initial conditions on Day 0 of (762; 1; 2). Figure 9 shows a comparison of the forecast for the epidemic based on this assimilation using the data for Days 2, 3 and 4. In this example, because the errors in reporting are minimal, due to small reported case numbers, a least squares approach gives comparable results. In practice the assimilation would continue as more data arrived. Each day a new forecast would be produced using the most recent data.

7. Conclusions

Some of the most challenging epidemiological modelling applications take place in real-time with a need to meld data and simple epidemic model structures to provide forecasts and parameter estimations. When intervention measures are applied during an outbreak the underlying dynamics of the epidemic inevitably changes. Behavioural changes that are a response to an epidemic in the population can also generate significant dynamical changes. As a consequence, only the limited amount of relatively recent data is relevant when attempting to estimate epidemiological parameters or provide forecasts. Here we have shown how VDA can be used to efficiently combine models and imperfectly reported data to provide robust single-step forecasts and parameter estimates. It is a technique that has not, hitherto, been used in epidemic modelling and the results demonstrated here are encouraging and

indicate that it is worthy of further investigation. Though we have not shown it here, it is straightforward to weight the cost function to allow for variation in the observation error as the epidemic progresses.

The purpose of this paper is to introduce this method and show how it can be used to perform some basic modelling tasks that are motivated by real-time outbreak analysis. It should be noted that recently Wearing *et al*. (2005) noted the influence of specific model assumptions when estimating epidemic parameters. Therefore, in any application of assimilation to real scenarios it would be advisable to pay attention to the form of the forward model and its impact on parameter estimation and forecasting.

The present discussion has been confined to simple non-spatial epidemic models. A basic *SIR* model with two epidemiological parameters is assumed to govern the dynamics, and when trying to estimate one parameter (R_0) the other (σ) is assumed to be known. In many cases of interest more complex epidemic models are required and it will be of interest to investigate to what extent assimilation can estimate a parameter such as R_0 in those cases. Also in many applications of assimilation to forecasting there is acknowledgment of the spatial component to modelling and data gathering. Future work will investigate more detailed application of VDA to real data sets also using more complex epidemiological models.

Appendix A

1. Cost function gradient from the adjoint model

We now show how the cost function gradient can be derived from the adjoint model. This is done by considering perturbations to the cost function with respect to the initial state and with respect to a future state.

Taking the initial state first, denote a small change to the initial state as $\overline{w}_0^{f,tl}$. This results in a change to the cost function

$$\delta J\left(\overline{w}_0^f\right) = J\left(\overline{w}_0^f + \overline{w}_0^{f,tl}\right) - J\left(\overline{w}_0^f\right) \tag{A1}$$

Taking the limit $\left|\overline{w}_0^{f,tl}\right| \to 0$, gives

$$\delta J\left(\overline{w}_0^f\right) = \left[\nabla J\left(\overline{w}_0^f\right)\right]^T \overline{w}_0^{f,tl} \tag{A2}$$

Using the definition of the cost function in equation 2.4, we can now differentiate the cost function with respect to a future state $\overline{w}_t^f$ giving

$$\delta J\left(\overline{w}_0^f\right) = \sum_t \left(\overline{w}_t^f - \overline{w}_t^o\right)^T \overline{w}_t^{f,tl} \tag{A3}$$

Equating equation A2 and A3 gives

$$\left[\nabla J\left(\overline{w}_0^f\right)\right]^T \overline{w}_0^{f,tl} = \sum_t \left(\overline{w}_t^f - \overline{w}_t^o\right)^T \overline{w}_t^{f,tl} \tag{A4}$$

Equation A4 gives us the gradient of the cost function, but we need to know how the perturbation of the initial state $\left(\overline{w}_0^{f,tl}\right)$ is related to the perturbation at the future

state $\left(\overline{w}_t^{f,tl}\right)$. Recall from equation 2.2 that the forward model connects the fields at adjacent time steps, so perturbations to those fields are connected as follows

$$\overline{w}_{t+1}^{f,tl} = \frac{\partial M\left(\overline{w}_t^{f}\right)}{\partial \overline{w}_t^{f}} \overline{w}_t^{f,tl} \tag{A5}$$

Equation A5 can be written

$$\overline{w}_{t+1}^{f} = L\left(\overline{w}_t^{f}\right)\overline{w}_t^{f,tl} \tag{A6}$$

Where $L\left(\overline{w}_t^{f}\right)$ is called the tangent linear operator of the forward model. In the same way as equations 2.1 and 2.2, we can make the desired connection between the perturbation of the initial state and the perturbation at the future state

$$\overline{w}_t^{f,tl} = L\left(\overline{w}_{t-1}^{f}\right)L\left(\overline{w}_{t-2}^{f}\right).....L\left(\overline{w}_1^{f}\right)L\left(\overline{w}_0^{f}\right)\overline{w}_0^{f,tl} \tag{A7}$$

which can be more economically written as

$$\overline{w}_t^{f,tl} = L_t \overline{w}_0^{f,tl} \tag{A8}$$

Substituting this in equation A4 gives

$$\left[\nabla J\left(\overline{w}_0^{f}\right)\right]^T \overline{w}_0^{f,tl} = \sum_t \left(\overline{w}_t^{f} - \overline{w}_t^{o}\right)^T L_t \overline{w}_0^{f,tl} \tag{A9}$$

that is

$$\nabla J\left(\overline{w}_0^{f}\right) = \sum_t L_t^T \left(\overline{w}_t^{f} - \overline{w}_t^{o}\right) \tag{A10}$$

We have now reached our objective, namely, an exact expression for the gradient of the cost function with respect to the initial conditions in terms of the transpose of the tangent linear model and the system observations, $\overline{w}_t^o$.

The transpose of the tangent linear operator, L_t^T is known as the adjoint operator and can be represented

$$L_t^T = L^T\left(\overline{w}_0^f\right)L^T\left(\overline{w}_1^f\right).....L^T\left(\overline{w}_{t-2}^f\right)L^T\left(\overline{w}_{t-1}^f\right) \tag{A11}$$

Note that the ordering of the time index is reversed in the adjoint operator. To calculate the gradient of the cost function using the r.h.s. of equation A10 it is necessary to integrate the adjoint model from $t = t_{\max}$ to $t = 0$ using the difference between the forward model and the observation fields as initial conditions.

2. A minimisation algorithm

Now that we have an expression for the gradient of the cost function with respect to the initial conditions, it can be used to iterate an initial estimate for $\overline{w}_0^f$ to the initial conditions that minimise the cost function. This can be done in the following straightforward way

$$\overline{w}_0^{f,n+1} = \overline{w}_0^{f,n} - \alpha\nabla J\left(\overline{w}_0^{f,n}\right) \tag{A12}$$

Where n is the number of iterations and α is a parameter selected to achieve an efficient convergence. For large n the gradient of the cost function $\nabla J\left(\overline{w}_0^{f,n}\right) \to 0$ and the optimum initial condition is reached. In this simple minimisation scheme the choice of α determines the speed (and possibility) of convergence to the optimum

initial conditions, and care must be taken to inspect the cost function to ensure that convergence has occurred.

Other, more sophisticated, minimisation algorithms could be used and some possibilities are discussed in Huang and Yang (1996).

Appendix B

From equation 3.2, $\frac{dI}{dt} = \beta SI - \sigma I$. At the start of an epidemic the number of susceptibles $S \simeq S_0$ so we can write the time evolution of the number of infectives as $\frac{dI}{dt} = (\beta S_0 - \sigma) I$. This has the solution $I = I_0 e^{(\beta S_0 - \sigma)t}$. So the rate of growth of the number of infectives, *r*, is $r = \beta S_0 - \sigma$. From Section 3.1 we saw $R_0 = \beta S_0 / \sigma$, so $r = R_0 \sigma - \sigma$. Re-arranging this gives $R_0 = 1 + r\sigma^{-1}$.

Appendix C

The off-the-shelf package used a least-squares fitting algorithm. The package was set up to do least-squares fitting of the three data points in Figure 7b. It is possible to constrain the values of the initial conditions of the search, i.e. the user can set limits on the likely values of *S*, *I* and *R* at the beginning of the epidemic (i.e. *t*=0 is Day 10).

Three separate constraint regimes were tested for the least-squares fit..

i) Unconstrained: Here we allowed the initial conditions of S, I and R to be any value between 0 and 3000. This resulted in an estimate of $R_0 \simeq 2.3$. The method gave an estimate for the initial conditions of $S_{t=0} = 3000, I_{t=0} = 535, R_{t=0} = 141$.

ii) Partially Constrained: We set the following constraint on the initial conditions $0 < S_{t=0} < 3000, 0 < I_{t=0} < 3000, 0 < R_{t=0} < 1000$, and this gave an

estimate of $R_0 \simeq 2.3$. The estimate of the initial conditions was calculated to be $S_{t=0} = 3000, I_{t=0} = 526, R_{t=0} = 170$.

iii) Constrained: We set the following constraint on the initial conditions to be $1500 < S_{t=0} < 3000, 200 < I_{t=0} < 1000, 0 < R_{t=0} < 500$, and this gave an estimate of $R_0 \simeq 2.3$. The estimate of the initial conditions was calculated to be $S_{t=0} = 3000, I_{t=0} = 525, R_{t=0} = 166$.

VDA gave an estimate of $R_0 \simeq 3.5$ (the underlying epidemic process that generated the data had $R_0 = 4$). The estimate of the initial conditions from assimilation was $S_{t=0} = 2653, I_{t=0} = 370, R_{t=0} = 135$.

Acknowledgements

CJR is supported by the Research Councils of the United Kingdom and Imperial College. TDH was supported by the European Community (SARSTRANS, contract SP22-CT-2004-511066). The authors thank two referees for useful comments that significantly improved the presentation of the manuscript.

Tables

Table 1: The final assimilated state is shown for a variety of different starting conditions for the noisy (16%) observational data on Days 12, 14 and 16 from Figures 3a-3d. The assimilated state is that for Day 10. For comparison, the epidemic state from the underlying model on Day 10 is $(2555;325;120)$.

	Starting State	Assimilated State
3a	(2100;700;200)	(2633;278;187)
	(800;700;1500)	(2633;278;187)
	(100;1000;1900)	(2633;278;187)
	(500;2000;950)	(2633;278;187)
3b	(100;700;1200)	(2558;338;67)
	(1800;700;500)	(2558;338;67)
	(100;1900;1000)	(2558;338;67)
	(150;1200;850)	(2558;338;67)
3c	(2500;300;200)	(2798;312;71)
	(1000;400;1600)	(2798;312;71)
	(200;200;2600)	(2798;312;71)
	(30;1500;1470)	(2798;312;71)
3d	(2000;500;500)	(2530;317;106)
	(400;1200;1400)	(2530;317;106)
	(100;300;2600)	(2530;317;106)
	(70;2900;30)	(2530;317;106)

Table 2: as the noise level on the observational data decreases the assimilated state is closer to the underlying model state at Day 10.

Noise Level	Starting State	Assimilated State
1%	(2100;700;200)	(2557;323;125)
4%	(2100;700;200)	(2572;314;137)
9%	(2100;700;200)	(2598;299;158)
16%	(2100;700;200)	(2633;278;187)
25%	(2100;700;200)	(2677;253;223)

Figure Captions

Figure 1: Time series for the first 40 days of the simulated epidemic using an *SIR* model. The pathogen (with $R_0 = 4$) is introduced by a single Infective at Day 0, and the peak number of Infectives occurs around Day 15. Susceptibles (squares), Infectives (triangles) and Recovered (circles).

Figure 2: To simulate realistic epidemic conditions, it is assumed that the outbreak is first noticed on Day 10 with data on the numbers of Infectives recorded on Days 12, 14 and 16 (triangle). The data is generated by introducing a random error $(\pm 16\%)$ on the underlying epidemic process (circles).

Figure 3a: The dots represent the results of the assimilation to estimate the values of *S*, *I* and *R* (square, triangle, circle) at Day 10. The result is close to the underlying epidemic process (as shown by the lines; long dashed (*S*), solid (*I*), short dashed (*R*)). Figure 3b-3d: same as Figure 3a for three different realisation of the noisy epidemic process showing the performance of the assimilation for different observational data sets (not shown) for Days 12 to 16.

Figure 4: Using the data points for Days 12, 14 and 16 (triangles) an estimate is made of when the epidemic was initiated. Assimilation indicates (grey dots) that the epidemic began around just over 10 days before the first recorded data point, i.e. around Day 1.

Figure 5: Using the three recorded data points (triangles) a forecast is made of the future number of infectives (thick line). The time series for the underlying model is shown for comparison (thin grey line).

Figure 6: Cost function (log scale) over 3000 iteration cycles for an assumed $R_0 = 6$.

Figure 7a: Asymptotic value of the cost functions as a function of R_0 obtained using the three data points on Day 12, 14 and 16 with $\pm 16\%$ error in prevalence reporting. This suggests that the basic reproductive ratio for the epidemic is between 3.5 and 4.

Figure 7b: Comparison of the results from assimilation (black line) and Berkeley Madonna package least squares fit (grey line). Also shown (dots) are the number of infectives from the underlying epidemic process (with $\pm 16\%$ error) that is used to generate the data.

Figure 8: Repeat of Figure 7a, though in this case the prevalence reporting error is increasing from $\pm 1\%$ to $\pm 50\%$. As the reporting error increases the estimate of R_0 starts to gradually decrease, but nevertheless continues to provide a reasonable estimate.

Figure 9: Result of data assimilation (solid line) using and *SIR* model and three data points (Day 2, 3 and 4; triangles) from an outbreak of influenza.

References

Anderson, R. M. and May, R. M., 1992. Infectious diseases of humans: dynamics and control. Oxford University Press, Oxford.

Anderson, R. M., Donnelly, C. A., Ferguson, N. M., Woolhouse, M. E. J. Watt, C. J., Udy, H. J., Mawhinney, S., Dunstan, S. P., Southwood, T. R. E., Wilesmith, J. W., Ryan, J. B. M., Hoinville, L. J., Hillerton, J. E., Austin, A. R. and Wells, G. A. H., 1996. Transmission dynamics and epidemiology of BSE in British cattle. Nature 382, 779-788.

Anderson, R. M., Fraser, C., Ghani, A. C., Donnelly, C. A., Riley, S., Ferguson, N. M., Leung, G. M., Lam, T. H. and Hedley, A. J., 2004. Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic. Phil. Trans. Roy. Soc. B359, 1091-1105.

Anonymous. 4th March 1978. Influenza in a boarding school. Brit. Med. J.

Bailey, N. J. T., 1957. The mathematical theory of epidemics. Griffin. London.

Bouttier, F. and Courtier, F., 1999. Data assimilation concepts and methods. European Centre for Medium Range Weather Forecasting. http://ecmwf.net/newsevents/training/rcourse_notes/DATA_ASSIMILATION/ASSIM_CONCEPTS/Assim_concepts2.html

Daley, R., 1991. Atmospheric data analysis. Cambridge University Press, Cambridge.

Daley, D. J. and Gani, J., 1999. Epidemic modelling: an introduction. Cambridge University Press, Cambridge.

Donnelly, C. A., Woodroffe, R., Cox, D. R., Bourne, F. J., Cheeseman, C. L., Clifton-Hadley, R. S., Wei, G., Gettinby, G., Gilks, P., Jenkins, H., Johnston, W. T., LeFevre, A. M., McInderney, J. P. and Morrison, W. I., 2006. Positive and negative effects of widespread badger culling on tuberculosis in cattle. Nature 439, 843-846.

Ferguson, N. M, Keeling, M. J., Edmunds, W. J., Gani, R., Grenfell, B. T., Anderson, R. M. and Leach, S., 2003. Planning for smallpox outbreaks. Nature 425, 681-685.

Ferguson, N. M., Cummings, D. A. T., Cauchemez, S., Fraser, C. and Riley, S., Iamsirithaworn A. M. S. and Burke, D. S., 2005 Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437, 209-214.

Ferrari, M. J., Bjornstad, O. N. and Dobson, A. P., 2005. Estimation and inference of R0 of an infectious pathogen by a removal method. Math. Biosc. 198, 14-26.

Fraser, C., 2007. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS One, 2(8), doi: 10.1371/journal.pone.0000758.

Gani, R and Leach, S., 2001. Transmission potential of smallpox in contemporary populations. Nature 414, 748-751.

Grenfell, B. T., Bjornstad, O. N. and Kappey, J., 2001. Travelling waves and spatial hierarchies in measles epidemics. Nature 414, 716-723.

Huang, X. Y. and Yang, X., 1996. Variational data assimilation with the Lorenz model. www.hirlam.org/open/publications/TechReports/TR26.ps.gz.

Jansen, V. A. A., Stollenwerk, N., Jensen, H. J., Ramsay, M. E., Edmunds, W. J. and Rhodes, C. J., Measles outbreaks in a population with declining vaccine uptake. 2003. *Science*, 301, 804.

Keeling, M. J., Woolhouse, M. E. J., May, R. M., Davies, G. and Grenfell, B. T. 2003. Modelling vaccination strategies against foot and mouth disease. Nature 421, 136-142.

Lawson, L. M., Spitz, Y. H., Hofmann, E. and Long, R. B. 1995. A data assimilation technique applied to a predator-prey model. Bull. Math. Biol. 57, 593-617.

Longini, I. M., Nizam, A., Xu, S. F., Ungchusak, K., Hanshaoworakul, W., Cummings, D. A. T. and Halloran, M. E., 2005. Containing pandemic influenza at the source. Science 309, 1083-1087.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. Numerical Recipes in Fortran (2nd ed.). 1992. Cambridge University Press, Cambridge.

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L-M, Lam, T-H., Thach, T. Q., Chau, P., Chan, K-P., Lo, S. V., Leung, P-Y., Tsang. T., Ho, W., Lee, K-H., Lau, E. M. C., Ferguson, N. M. and Anderson, R. M., (2003) Transmission dynamics of the etiological agent of SARS in Hong Kong: impact pf public health interventions. Science 300, 1961-1966.

Wearing, H. J., Rohani, P. and Keeling, M. J., 2005. Appropriate models for the management of infectious diseases. PloS Med. 2, 0621-0627.