Using the stochastic Galerkin method as a predictive tool during an epidemic

D. B. Harman[1] P. R. Johnston[2]

17 November 2017; revised 10 February 2019

Abstract

The ability to accurately predict the course of an epidemic is extremely important. This article looks at an influenza outbreak that spread through a small boarding school. Predictions are made on multiple days throughout the epidemic using the stochastic Galerkin method to consider a range of plausible values for the parameters. These predictions are then compared to known data points. Predictions made before the peak of the epidemic had much larger variances compared to predictions made after the peak of the epidemic.

Subject class: 92D30

Keywords: epidemic modelling; stochastic Galerkin; predictions

DOI:10.21914/anziamj.v59i0.12654, Published July 25, 2019, as part of the Proceedings of the 13th Biennial Engineering Mathematics and Applications Conference. ISSN 1445-8810. (Print two pages per sheet of paper.)

Contents

1 Introduction

Compartment epidemic models are often used for modelling the likely course of an epidemic and have been extensively studied [7]. However, the parameters within these models are often not known with certainty [11]. It is important for this uncertainty to be included into epidemic models in order to obtain accurate predictions.

One way of incorporating uncertainty into an epidemic model is to make the uncertain parameters functions of random variables [2]. The mean and variance can then be determined using a sampling technique such as Monte Carlo sampling. However, this can be computationally expensive depending upon the probability distributions of the random variables and the number of uncertain parameters. Alternatively, the stochastic Galerkin method has been

shown to produce accurate results while being much more computationally efficient [4].

Because of the stochastic Galerkin's advantages over Monte Carlo sampling, it has been extensively studied and applied to a variety of problems. However, only a limited number of articles have applied the stochastic Galerkin method to epidemic modelling [2, 12, 8, 11, 10, 4, 3]. This article extends the work done by Roberts [11] and Harman and Johnston [3] by using the stochastic Galerkin method to make predictions of the epidemic curve on multiple days during an epidemic by considering a range of plausible values for the parameters. Roberts [11] uses the stochastic Galerkin method to predict the likely course of an influenza outbreak in New Zealand using data from 2009. Uncertainty was incorporated into the reproduction number, whereas this article considers uncertainty in multiple parameters of an SIR model. This allows for much greater flexibility in the representation of the uncertainty. While Harman and Johnston [3] consider multiple uncertain parameters, the prediction obtained from the stochastic Galerkin method is only calculated towards the end of the epidemic, well after the peak of the epidemic. This article considers multiple predictions around the peak of the epidemic. These predictions are then compared to known data points.

2 The SIR epidemic compartment model

One of the most common epidemic compartment models is the SIR model [9]. In the SIR model, each person within the population is placed into one of three possible compartments: susceptible (S), infected (I) or recovered (R). The differential equations for the SIR model are given by

$$\frac{dS}{dt} = -\beta SI, \qquad \frac{dI}{dt} = \beta SI - \gamma I, \qquad \frac{dR}{dt} = \gamma I, \tag{1}$$

where β is the 'contact rate' and $1/\gamma$ is the average recovery time [7]. Note that these equations have been normalised such that $S + I + R = 1$. The

parameters β and γ are often assumed to be constants. However, they are rarely known with certainty. To represent the uncertainty in these parameters, they can be considered functions of random variables. For example,

$$\beta = \beta(\xi_1), \quad \gamma = \gamma(\xi_2),$$

where ξ_1 and ξ_2 are random variables with probability density functions $w_1(\xi_1)$ and $w_2(\xi_2)$ respectively and probability spaces $(\Omega_1, \mathcal{F}_1, \mathcal{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathcal{P}_2)$ respectively.

As the SIR model now contains random variables, it can no longer be solved using a single call to a numerical ODE solver. Random sampling techniques such as Monte Carlo sampling can be used to determine the mean solution and its variance. However, a more computationally efficient method is the stochastic Galerkin method [3].

2.1 Applying the stochastic Galerkin method

To apply the stochastic Galerkin method, the solutions for S and I are expanded in the form

$$\begin{aligned} S(t, \xi_1, \xi_2) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} S_{ij}(t) \Psi_i(\xi_1) \Phi_j(\xi_2), \\ I(t, \xi_1, \xi_2) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} I_{ij}(t) \Psi_i(\xi_1) \Phi_j(\xi_2), \end{aligned} \tag{2}$$

where $S_{ij}(t)$ and $I_{ij}(t)$ are deterministic functions that need to be determined, and $\Psi_i(\xi_1)$ and $\Phi_j(\xi_2)$ are appropriately chosen orthogonal polynomials [1]. The orthogonal polynomials $\Psi_i(\xi_1)$ and $\Phi_j(\xi_2)$ form a basis over which the solutions for S and I can be expanded. It is important to note that S and I are now written explicitly as functions of not only time, but the random variables ξ_1 and ξ_2 as well.

Substituting the form of the solutions for S and I (Equations (2)) into the SIR model (Equations (1)) gives

$$\begin{aligned}
\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{dS_{ij}}{dt}\Psi_i\Phi_j &= -\beta\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}S_{ij}I_{mn}\Psi_i\Phi_j\Psi_m\Phi_n,\\
\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{dI_{ij}}{dt}\Psi_i\Phi_j &= \beta\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}S_{ij}I_{mn}\Psi_i\Phi_j\Psi_m\Phi_n\\
&\quad -\gamma\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}I_{ij}\Psi_i\Phi_j.
\end{aligned} \tag{3}$$

Applying a Galerkin projection by multiplying through by $\Psi_u(\xi_1)\Phi_v(\xi_2)$ (where $u, v = 1, 2, \ldots$), integrating over the probability space and truncating the expansions at the Pth order gives

$$\begin{aligned}
\frac{dS_{uv}}{dt} &= -\frac{1}{K_{uv}}\sum_{i=0}^{P}\sum_{j=0}^{P-i}\sum_{m=0}^{P}\sum_{n=0}^{P-m}S_{ij}I_{mn}\langle\beta\Psi_i\Phi_j\Psi_m\Phi_n,\Psi_u\Phi_v\rangle,\\
\frac{dI_{uv}}{dt} &= \frac{1}{K_{uv}}\sum_{i=0}^{P}\sum_{j=0}^{P-i}\sum_{m=0}^{P}\sum_{n=0}^{P-m}S_{ij}I_{mn}\langle\beta\Psi_i\Phi_j\Psi_m\Phi_n,\Psi_u\Phi_v\rangle\\
&\quad -\frac{1}{K_{uv}}\sum_{i=0}^{P}\sum_{j=0}^{P-i}I_{ij}\langle\gamma\Psi_i\Phi_j,\Psi_u\Phi_v\rangle,
\end{aligned} \tag{4}$$

where the inner product $\langle F, G\rangle$ is defined as

$$\langle F, G\rangle = \int_{\Omega_2}\int_{\Omega_1} FGw_1(\xi_1)w_2(\xi_2)d\xi_1 d\xi_2,$$

and

$$K_{uv} = \int_{\Omega_2}\int_{\Omega_1}\left(\Psi_u(\xi_1)\right)^2\left(\Phi_v(\xi_2)\right)^2 w_1(\xi_1)w_2(\xi_2)d\xi_1 d\xi_2.$$

If the orthogonal polynomials $\Psi_i(\xi_1)$ and $\Phi_j(\xi_2)$ are chosen such their weight functions are $w_1(\xi_1)$ and $w_2(\xi_2)$ respectively, many of the inner products trivially evaluate to zero. This gives a system of $2\binom{P+2}{2}$ *deterministic* differential equations for $S_{ij}(t)$ and $I_{ij}(t)$ [6].

For example, assume β has a uniform distribution on $[2.5, 5.5]$ and γ has a uniform distribution on $[0.5, 1.5]$. As β and γ have uniform distributions, the Legendre orthogonal polynomials would be chosen for $\Psi_i(\xi_1)$ and $\Phi_j(\xi_2)$. The inner products in Equations (4) could either be evaluated numerically or symbolically. The first order ($P = 1$) system of equations for $S_{ij}(t)$ and $I_{ij}(t)$ is given by

$$
\begin{aligned}
\frac{dS_{00}}{dt} &= -4S_{00}I_{00} - \frac{4}{3}S_{01}I_{01} - \frac{1}{2}I_{00}S_{10} - \frac{1}{2}I_{10}S_{00} - \frac{4}{3}I_{10}S_{10},\\
\frac{dI_{00}}{dt} &= 4S_{00}I_{00} + \frac{4}{3}S_{01}I_{01} + \frac{1}{2}I_{00}S_{10} + \frac{1}{2}I_{10}S_{00} + \frac{4}{3}I_{10}S_{10} - I_{00} - \frac{4}{3}S_{10}I_{10},\\
\frac{dS_{01}}{dt} &= -4S_{01}I_{00} - 4S_{00}I_{01} - \frac{1}{2}S_{10}I_{01} - \frac{1}{2}S_{01}I_{10},\\
\frac{dI_{01}}{dt} &= 4S_{01}I_{00} + 4S_{00}I_{01} + \frac{1}{2}S_{10}I_{01} + \frac{1}{2}S_{01}I_{10} - I_{01} - \frac{1}{2}I_{00},\\
\frac{dS_{10}}{dt} &= -\frac{3}{2}S_{00}I_{00} - \frac{1}{2}S_{01}I_{01} - 4S_{10}I_{00} - 4S_{00}I_{10} - \frac{9}{10}S_{10}I_{10},\\
\frac{dI_{10}}{dt} &= +\frac{3}{2}S_{00}I_{00} + \frac{1}{2}S_{01}I_{01} + 4S_{10}I_{00} + 4S_{00}I_{10} + \frac{9}{10}S_{10}I_{10} - I_{10}.
\end{aligned}
$$

As the system of differential equations is deterministic, it can easily be solved using a numerical solver such as the MATLAB function `ode45`. While uncertainty was included in the model using random variables, the final system of equations is deterministic and only needs to be solved once.

2.2 Determining mean and variance from stochastic Galerkin solution

Once the stochastic Galerkin solution has been obtained, the mean and variance can easily be determined with very little additional computation. For example, the mean number of infected individuals, $\mathsf{E}[I(t, \xi_1, \xi_2)]$, is given by

$$\mathsf{E}[I(t, \xi_1, \xi_2)] = I_{00}(t),$$

and the variance, $\mathsf{Var}[I(t, \xi_1, \xi_2)]$, is given by

$$\mathsf{Var}[I(t, \xi_1, \xi_2)] = \sum_{i=0}^{P} \sum_{j=0}^{P-i} (I_{ij}(t))^2 \left\langle (\Psi_i(\xi_1))^2, (\Phi_j(\xi_2))^2 \right\rangle - (I_{00}(t))^2.$$

Therefore the mean solution is simply given by the zero order term of the stochastic Galerkin expansion and the variance can quickly be calculated from the higher order terms [15].

3 Influenza spreading through a small boarding school

In this section, an influenza epidemic that spread through a small boarding school in the North of England will be investigated [13]. An SIR model will be used to make predictions on different days during the epidemic using only data that would have been available on that day. These predictions will then be compared to known data points.

3.1 Using a 'best fit' approach

One of the simplest prediction methods is to simply find the 'best fit' using a least squares fitting technique. For chosen values of β and γ, the error for a prediction made D days after the start of the epidemic, $E^D_{\beta,\gamma}$, is given by

$$E^D_{\beta,\gamma} = \sqrt{\sum_{k=1}^{D} [I_{\beta,\gamma}(k) - I_R(k)]^2}, \tag{5}$$

where $I_{\beta,\gamma}(k)$ is the fraction of individuals infected on day k using the SIR model and $I_R(k)$ is the actual fraction of individuals infected on day k.

While this prediction has the lowest error, there is no guarantee that this is an accurate prediction. It also gives no confidence intervals on what *could* happen.

3.2 Using the stochastic Galerkin method

Rather than simply considering the 'best' values for β and γ in the SIR model, a range of plausible values for each of the parameters will be considered. A pair of β and γ values will be considered plausible if it is below a predetermined error threshold.

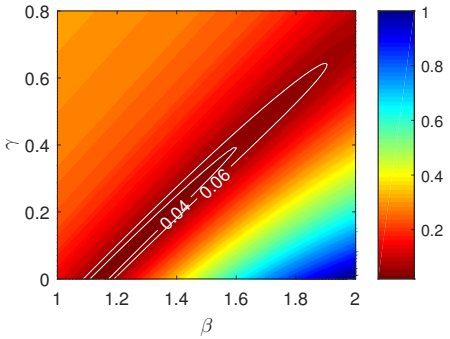
As the outcome of the epidemic is already known, it would be easy to simply choose error thresholds on a given day of the epidemic that resulted in accurate predictions. Therefore, it is important to implement an algorithm for calculating the error thresholds so that knowledge of future data points does not influence the determination of error thresholds. Therefore, the error thresholds for a given day will be based upon the error of the 'best fit' for that day. Two error thresholds will be considered: double and triple the error obtained when using the 'best' values. The stochastic Galerkin method will then be applied to this range of plausible values and the mean prediction and

its variance obtained. For example, on day five of the epidemic, the 'best fit' prediction has an error of approximately 0.02. Therefore the two error thresholds for plausible β and γ values will be 0.04 and 0.06. A heat map of the error as well as the ranges of plausible β and γ values on days five, six and seven of the epidemic can be seen in Figure 1. The error heat maps were calculated using Equation (5).

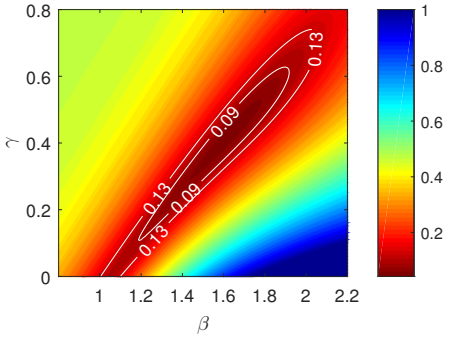
While Figure 1 shows the full error heat maps, it is important to note that it is not necessary to fully calculate the error heat maps in order to obtain probability distributions for β and γ. As the range of plausible values forms a closed shape, a simple algorithm can be used to find the border of plausible values [5]. This significantly decreases the number of parameter pairs that need to be tested. For example, the error for parameter pairs in the top left corner of the heat maps does need to be calculated as it falls well outside the border of plausible values. However, for clarity, the full error heat maps are included in this article.

It is also important to note that the aim of this article is not to compare the predictions obtained from the 'best fit' and stochastic Galerkin methods. These are two very different methods. The 'best fit' can be implemented with a handful of lines of code in MATLAB, whereas the stochastic Galerkin method requires significantly more effort to implement. The 'best fit' prediction is simply calculated so that the error thresholds for the stochastic Galerkin method are chosen without bias.

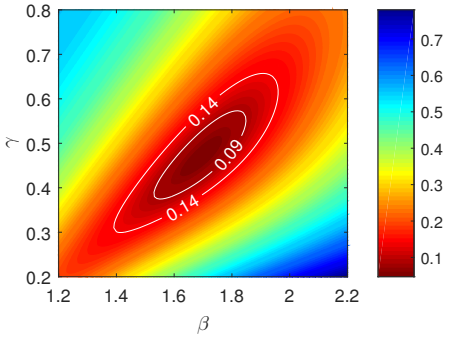
Using the ranges of plausible values for β and γ (Figure 1), the probability density functions for β and γ were calculated in order to apply the stochastic Galerkin method. To determine the probability distribution for β, the number of plausible γ values for each plausible value of β was counted. The results were then plotted as a histogram. By normalising the area under the histogram, the general shape of the probability distribution was found. A similar process was used to find the probability density function for γ. The probability distributions for β and γ on days five, six and seven of the epidemic can be seen in Figure 2. Parameter ranges for β, γ and R_0 are given in Table 1.



(a) Day 5

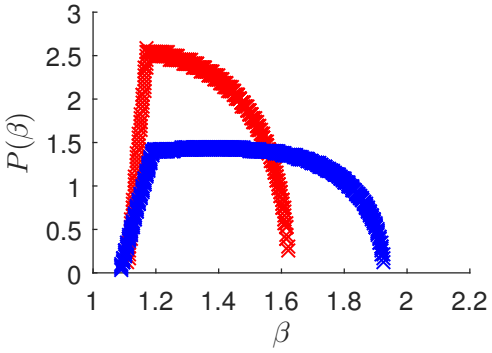


(b) Day 6

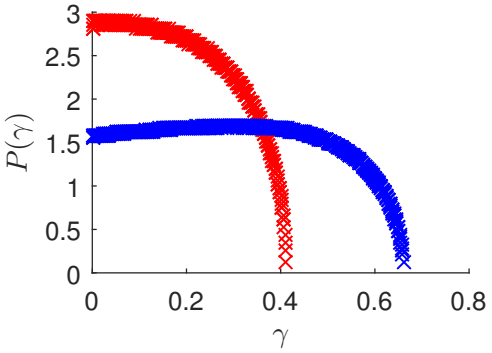


(c) Day 7

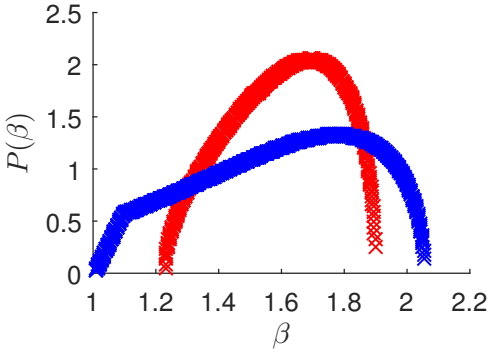
Figure 1: Error heat maps for β and γ values on days five, six and seven of the epidemic. Contours are double and triple the 'best fit' error on that day.



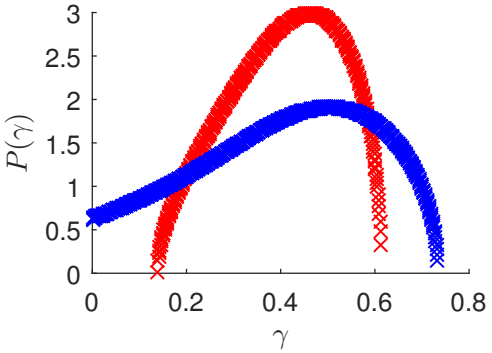
(a) Day 5 β distribution



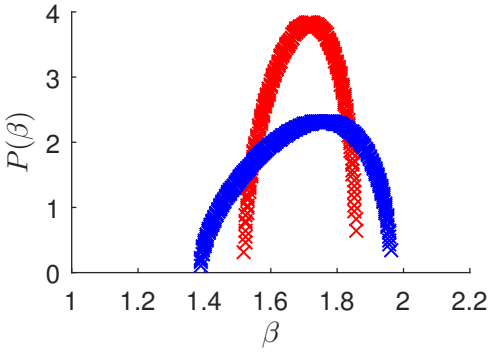
(b) Day 5 γ distribution



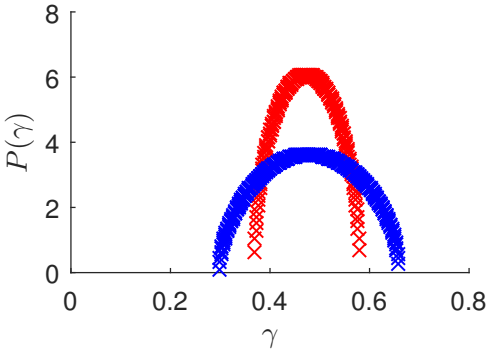
(c) Day 6 β distribution



(d) Day 6 γ distribution



(e) Day 7 β distribution



(f) Day 7 γ distribution

Figure 2: Probability distributions for β and γ on days five, six and seven of the epidemic. Red uses the smaller error threshold while blue uses the larger error threshold.

Day	Error Threshold	β range	γ range	R_0 range
5	0.04	1.11 - 1.62	0 - 0.41	3.93 -
5	0.06	1.09 - 1.92	0 - 0.66	2.90 -
6	0.09	1.23 - 1.90	0.14 - 0.61	3.04 - 8.86
6	0.13	1.01 - 2.05	0 - 0.73	2.72 -
7	0.09	1.52 - 1.86	0.37 - 0.58	3.12 - 4.22
7	0.14	1.39 - 1.96	0.30 - 0.66	2.85 - 4.79
8	0.10	1.53 - 1.80	0.38 - 0.54	3.22 - 4.24
8	0.16	1.43 - 1.89	0.33 - 0.61	2.95 - 4.65

Table 1: Parameter ranges for β, γ and R_0 on days 5-8 of the epidemic. Upper limits on R_0 are not possible on days 5 and 6 as $\gamma = 0$ was a plausible value.

As the probability distributions were non-standard, fifth order polynomials were used to approximate the probability distributions. The associated orthogonal polynomials were then derived using the Gram-Schmidt orthogonalisation method [14]. Finally, the stochastic Galerkin method was applied to find the mean prediction and its variance.

3.3 Predictions

Figure 3 shows the predictions calculated using the stochastic Galerkin method on days five, six and seven of the epidemic. These predictions were calculated using MATLAB version R2016a. The 'best fit' predictions are also shown in Figure 3, as well as the known data points.

On day five, both of the predictions from the stochastic Galerkin method overestimate the peak of the epidemic as well as its tail. While the contours of plausible values for β and γ appear quite narrow (Figure 1(a)), plausible γ values range from 0 to 0.66 for the larger error threshold. This causes the variance in the stochastic Galerkin predictions to be quite large, especially when using the larger error threshold. It is interesting to note that when

using the smaller error threshold, many of the 'future' data points fall outside one standard deviation. However, when using the larger error threshold, most of the data points fall within one standard deviation.

Both stochastic Galerkin predictions calculated on day six of the epidemic are much better than those calculated on day five. The predictions begin very similarly and reach their peaks at approximately the same time. However, their tails are slightly different and overestimate the tail of the epidemic. Even using the smaller error threshold, most of the data points fall within one standard deviation of the mean prediction.

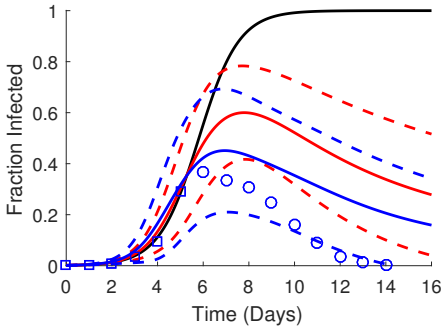
By day seven, the range of plausible γ values has significantly decreased with the larger error threshold having a minimum γ value of 0.3. This is due to the day seven data point showing a decrease in the number of infected students for the first time which greatly helps in the estimation of γ. Because of this, the stochastic Galerkin predictions are very similar, with the prediction calculated using the smaller error threshold having a slightly higher peak. Both predictions underestimate the fraction infected on days eight, nine and ten but overestimate the fraction infected on days twelve through fourteen. The variance is also considerably smaller than the day six predictions.

Predictions calculated after day seven give similar results to the predictions calculated on day seven as the peak of the epidemic has already passed. Parameter ranges for β and γ on day eight of the epidemic can be seen in Table 1.

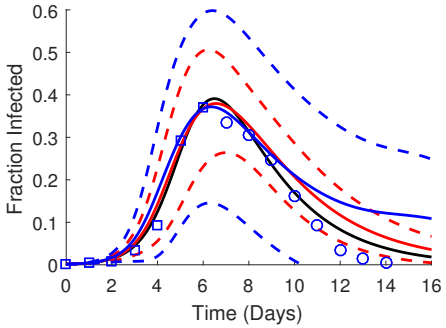
4 Conclusion

This article has looked at using the stochastic Galerkin method to make predictions on different days during an epidemic and comparing the predictions to known data points.

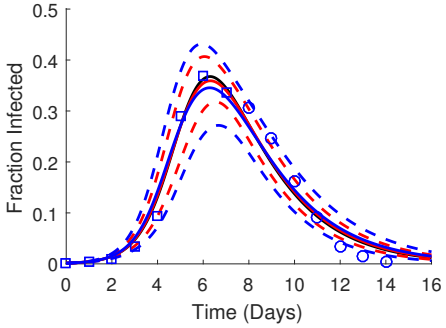
Rather than simply considering the parameter values that result in the



(a) Day 5



(b) Day 6



(c) Day 7

Figure 3: Predictions made on days five, six and seven of the epidemic. Squares are known data points while circles are future data points. Black is the 'best fit' prediction. Red and blue are stochastic Galerkin predictions with double and triple the error of the 'best fit' respectively. Solid lines are mean predictions and dashed lines are one standard deviation from the mean.

smallest error, a range of plausible values for the parameters can instead be considered. From these ranges of plausible values, probability distributions for the parameters can be determined. The stochastic Galerkin method can then be used to determine the mean prediction. The variance can also be determined from the stochastic Galerkin solution which gives confidence intervals for the prediction.

In this article, it was assumed that there was a single student who returned to the boarding school infected with influenza. However, this work could be extended to find plausible ranges of initial conditions and what effect this has on the predictions during the epidemic.

Acknowledgments This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

References

[1] B. M. Chen-Charpentier, J. C. Cortes, J. V. Romero, and M. D. Rosello. Some recommendations for applying gPC (generalized polynomial chaos) to modeling: An analysis through the Airy random differential equation. *Applied Mathematics and Computation*, 219(9):4208 – 4218, 2013. doi:10.1016/j.amc.2012.11.007 C304

[2] B. M. Chen-Charpentier and D. Stanescu. Epidemic models with random coefficients. *Mathematical and Computer Modelling*, 52:1004 – 1010, 2010. doi:10.1016/j.mcm.2010.01.014 C302, C303

[3] D. B. Harman and P. R. Johnston. Applying the stochastic galerkin method to epidemic models with individualised parameter distributions. In *Proceedings of the 12th Biennial Engineering Mathematics and Applications Conference, EMAC-2015*, volume 57 of *ANZIAM J.*, pages C160–C176, August 2016. doi:10.21914/anziamj.v57i0.10394 C303, C304

[4] D. B. Harman and P. R. Johnston. Applying the stochastic galerkin method to epidemic models with uncertainty in the parameters. *Mathematical Biosciences*, 277:25 – 37, 2016. doi:10.1016/j.mbs.2016.03.012 C303

[5] D. B. Harman and P. R. Johnston. Boarding house: find border. 2019. doi:10.6084/m9.figshare.7699844.v1 C309

[6] D. B. Harman and P. R. Johnston. SIR uniform equations. 2 2019. doi:10.6084/m9.figshare.7692392.v1 C306

[7] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000. doi:10.1137/S0036144500371907 C302, C303

[8] R.I. Hickson and M.G. Roberts. How population heterogeneity in susceptibility and infectivity influences epidemic dynamics. *Journal of Theoretical Biology*, 350(0):70 – 80, 2014. doi:10.1016/j.jtbi.2014.01.014 C303

[9] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, August 1927. doi:10.1098/rspa.1927.0118 C303

[10] M. G. Roberts. A two-strain epidemic model with uncertainty in the interaction. *The ANZIAM Journal*, 54:108–115, 10 2012. doi:10.1017/S1446181112000326 C303

[11] M. G. Roberts. Epidemic models with uncertainty in the reproduction number. *Journal of Mathematical Biology*, 66(7):1463–1474, 2013. doi:10.1007/s00285-012-0540-y C302, C303

[12] F. Santonja and B. Chen-Charpentier. Uncertainty quantification in simulations of epidemics using polynomial chaos. *Computational and Mathematical Methods in Medicine*, 2012:742086, 2012. doi:10.1155/2012/742086 C303

[13] Communicable Disease Surveillance Centre (Public Health Laboratory Service) and Communicable Diseases (Scotland) Unit. Influenza in a boarding school. *BMJ*, 1(6112):587, 1978. doi:10.1136/bmj.1.6112.586 C307

[14] G. Strang. *Linear Algebra and Its Applications*. Thomson, Brooks/Cole, 2006. C312

[15] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010. C307

Author addresses

1. **D. B. Harman**, School of Environment and Science, Griffith University, Queensland 4111, Australia.
mailto:david.harman@alumni.griffithuni.edu.au

2. **P. R. Johnston**, School of Environment and Science, Griffith University, Queensland 4111, Australia.
mailto:p.johnston@griffith.edu.au