Characterising model dynamics using sparse grid interpolation: Parameter estimation of cholera

Aditya Sai and Nan Kong

Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA

ABSTRACT
Sparse grid interpolation is a popular numerical discretization technique for the treatment of high dimensional, multivariate problems. We consider the case of using time-series data to calibrate epidemiological models from both phenomenological and mechanistic perspectives using this computational tool. By capturing the dynamics underlying both global and local spaces, our algorithm identifies potentially optimal regions of the parameter space and directs computational effort towards resolving the dynamics and resulting fits of these regions. We demonstrate how sparse grid interpolants can be effectively deployed to fit available data and discriminate between competing hypotheses to explain the current cholera epidemic in Yemen.

ARTICLE HISTORY
Received 6 December 2017
Accepted 29 July 2018

KEYWORDS
Surrogate modelling; parameter estimation; epidemiology; sparse grid interpolation

1. Introduction
Mathematical models of biological phenomena rely on parameters to capture model behaviour [33]. Parameter values must be estimated with accuracy to provide any meaningful insight into critical biological problems. Limited prior knowledge on parameter regimes often prohibits targeted or smart sampling strategies, hindering efforts at successful parameter estimation. In the domain of epidemiology, many parameters are not easily derived from literature, nor directly observable, and yet are indispensable to characterizing the force of infection within a population. Parameter estimation in epidemiology usually relies on approaches like Bayesian [4,9,18], likelihood-based [19,34,35,38], evolutionary computing [1], and least squares methods [7,31]. We propose an alternative parameter estimation strategy that can provide dynamical snapshots of model behaviour within specified parameter spaces. Specifically, we use sparse grid interpolation, a surrogate modelling technique, to estimate relevant model dynamics across a predefined parameter space. By enclosing the estimation problem within a proxy environment, sufficient samples can be taken to obtain a comprehensive assessment of parameter fitness at a fraction of the cost of directly simulating the model.
The purpose of this work is to introduce an effective surrogate modelling technique for parameter inference of two stylistically distinct models describing the current cholera epidemic in Yemen. This paper is organized as follows. We review sparse grid interpolation in Section 2. Section 3 proposes a sparse grid-enabled algorithm to inform parameter values for epidemiological models equipped with available time-series data. In Sections 4 and 5, we introduce and demonstrate our approach on two cholera models. Finally, Section 6 analyzes our findings and offers some perspective.

2. Sparse grid interpolation

A sufficiently well parameterized model entails a high-dimensional parameter space. A higher dimensional model is often computationally difficult to simulate, deeming the parameter estimation problem intractable. On the other hand, the global diversity of model behaviors desired for accurate parameter estimation may be forfeited by compromising on the simulation effort. Computationally intensive models also present a similar obstacle, where it is desired to minimize the number of direct model evaluations as much as possible. Sparse grid interpolation presents a viable, parsimonious solution to these challenges. By sampling the parameter space strategically and selectively, sparse grid interpolants closely approximate the target model [2,5,13,15]. The interpolant is constructed by combining basis functions at a set of sparsely sampled points across the parameter space. By interrogating the interpolant rather than the target model, excessive and costly model evaluations can be avoided. Figure 1 demonstrates the application of sparse grid interpolation to a simple 3-dimensional exponential function. The concept of sparse grid interpolation can be traced back to the Russian mathematician Smolyak, who developed an efficient technique to extend tensor product formulas for numerical integration, or quadrature, to multiple dimensions [32]. Smolyak’s algorithm takes the partial tensor product of univariate quadrature rules instead of the full tensor product representation, minimizing the number of points used while maintaining an error up to a logarithmic factor [2,5,13,15]. Important features of sparse grid interpolation include hierarchical decomposition and dimensional adaptivity. The hierarchical property of sparse grids allows for points to be reused at higher levels of refinement, meaning only points unique to the higher level are evaluated [5,13,21]. Adaptive sparse grids place more points along dimensions of the parameter space that contribute most to the interpolation error to produce a smoother, more accurate interpolant [14,39].

Sparse grids have been used to aid efforts in parameter identification before [10,11]. In particular, Donahue et al. [11] demonstrated that the quality of an 18-dimensional sparse grid-based parameter estimation method improved when the number of model evaluations increased. Furthermore, the sparse grid approach outperformed a standard optimization method when the same number of model evaluations were considered for both. However, these early approaches tended to interpolate the cost function itself, which can be rather error-prone due to function irregularities. Later approaches [3,24] interpolated the actual model dynamics, resulting in a far more accurate interpolant with fewer model evaluations. Localized sparse grids were introduced in [3], and further explored in [24], to enable searches of local subspaces once the global search was exhausted. Interpolants constructed on local grids, when rendered sufficiently accurate, can improve upon the results obtained from their global counterparts. However, the original concept was intended
Figure 1. Exponential function evaluated on a grid $[-2, 2] \times [-2, 2]$. Both the original function (left) and the sparse grid interpolant (right) are shown. The interpolant was produced with a relative error of 0.021%, absolute error of 0.00039, and 129 support nodes.

towards model-based experiment design for reducing uncertainty in model dynamics. Furthermore, local grids were originally intended to identify a sufficient number of parameters to satisfy a given criterion, not necessarily to determine which parameters best minimized the difference between simulated outcomes and observed data. In this work, we define a parameter to be either a point within, or a dimension of, the search space, depending on the context. We also define a parameter value to be a particular numerical value for a parameter. To interpolate model dynamics using sparse grids, we use the Matlab-based Sparse Grid Interpolation Toolbox [20].

3. Parameter estimation algorithm

To demonstrate the advantages of sparse grid interpolation in approximating model trajectories, we propose a parameter estimation algorithm that operates at successive levels of refinement to match underlying data. The algorithm searches for potentially optimal parameters on both the global and local scales by constructing interpolants that effectively capture model dynamics within targeted subspaces of the global parameter space.

3.1. Global stage

The global stage scans the entire parameter space for potentially optimal regions by constructing a global interpolant. We stipulate that the interpolant must possess a relative error of less than 1%, regardless of how many points in the parameter space (i.e. model evaluations) are required. This level of accuracy ensures that parameters found at the global stage are suitable starting points for the local stage. The global interpolant is then sampled using Latin Hypercube Sampling (LHS) to obtain comprehensive coverage of the dynamics underlying the global space. LHS ensures this coverage by dividing the space into
equally sized intervals based on the number of samples desired. Samples are then acquired from these intervals according to some probability distribution. For our algorithm, in the absence of any prior information, we apply LHS using a uniform distribution. To measure the fitness of each sampled parameter, an unweighted sum of squared errors (SSE) cost function compares the interpolated trajectories with the available data. Then, we choose those parameters whose costs are below a model-dependent threshold to form the initial parameter set for the local stage.

### 3.2. Local stage

The local stage starts with the incoming parameter set collected during the global stage. Here, we apply cluster analysis to select the shape and number of local grids at each iteration. Each cluster will specify the space upon which a local grid will be created. Cluster analysis, seen as an extension of multistart methods in the context of global optimization, can avoid the redundancy of detecting the same local minima repeatedly by isolating neighbourhoods of local optima in order to conduct efficient, productive searches [25,30].

$k$-means clustering partitions the parameter set based on each parameter’s proximity to $k$ cluster means. These clusters are determined by using an objective associated with the silhouette method, where clusters are well-separated and cluster members are appropriately categorized:

$$k^* = \arg\max_k \bar{S}(k) \cdot S_{\min}(k),$$  

where $\bar{S}(k)$ is the average silhouette coefficient, and $S_{\min}(k)$ is the minimum silhouette coefficient, for all parameters in $k$ clusters. The objective function penalizes negative silhouette coefficients, which suggest mis-clustering and lack of cohesion within a cluster, and prioritizes higher average silhouette coefficients, which indicate good separation.

Gaussian mixture models (GMMs) assign each parameter soft membership to a cluster defined by a Gaussian distribution. In [3,24], GMMs were used to cluster parameters with a criterion that minimizes the volume of overlap between local grids, which we use here:

$$k^* = \arg\min_k V(k),$$

where $V(k)$ is the volume of overlap between $k$ clusters.

Once the local grid ranges have been specified, the local interpolants are then created, with more stringent accuracy requirements than the global grid. We impose a limit on the relative error of the local interpolants to $10^{-3}\%$, with a maximum number of model evaluations set to 1000. The overall accuracy of the interpolant must be maintained so that the algorithm does not return parameters whose interpolated trajectories differ greatly from the actual dynamics.

Parameters are then sampled from each local interpolant uniformly using LHS. The interpolant generates approximate trajectories for each parameter, which are then compared with observed data to yield a cost. Those parameters with the lowest costs across all local interpolants are retained. The process then returns to the clustering step of the local stage, with the newly found parameters. We devised two stopping criteria, upon which the algorithm will terminate:
(1) The number of overall iterations.
(2) When no change in the minimum cost is observed after 2 consecutive iterations.

Once either stopping criterion is satisfied, the algorithm will output the best (i.e. lowest SSE) parameters found in each iteration. The number of times the original model is evaluated is also reported for reference. The purpose of employing a sparse grid-enabled parameter estimation algorithm is to sparingly evaluate the original model so that the interpolant can serve as an accurate proxy, saving computational effort and time.

For comparison with the clustering methods at the local stage, we also perform parameter estimation using the Nelder-Mead simplex method, an unconstrained, derivative-free optimization method [22,27]. This is done by starting the Nelder-Mead optimization from the best parameter found in the global stage, thereby avoiding creation of local interpolants and instead evaluating the original model as needed.

4. Cholera models

4.1. Generalized logistic model

We first consider the generalized logistic model, a phenomenological model describing the number of cumulative cases as a function of four parameters:

\[ C(t) = \frac{K}{1 + g e^{-r(t-t_i)}}^{1/g}. \] (3)

The generalized logistic model was introduced for cholera in [29], one of the first analyses of the Yemen epidemic. The authors assimilated weekly incidence data, as they became available, to infer model parameters via a maximum likelihood estimation approach. With the updated parameters, they were able to produce real-time epidemic curves that excluded potential reporting biases. They estimated the epidemic would peak by the middle of 2017 and then decline in the subsequent weeks, ultimately reaching a cumulative incidence of 700–900,000 cases by the end of the current outbreak. However, their study used data from a shorter time interval, April 16 to July 1, 2017, which led to an underestimation of the final epidemic size and the peak timing of the epidemic, as cholera cases surpassed the million case mark by the end of 2017 [37].

Model variables and parameters are listed in Table 1. \( K \), the carrying capacity, defines the cumulative incidence at time infinity, or the final epidemic size. \( t_i \), the inflection point, defines the time of peak incidence. No prior parameter ranges are available for the revised dataset we use for parameter estimation, so we perform a preliminary search step to determine an initial parameter fit by performing Nelder-Mead optimization on a range of plausible values for \( t_i \) and \( r \). Once \( t_i \) and \( r \) have been determined, we solve for \( K \) using Equation 3 when \( g = 1 \). From there, we compute parameter ranges for all four parameters that when interpolated over, produce sufficiently accurate dynamics.

4.2. ODE model

The second model describes population dynamics using a compartmental approach to cholera. It was originally used to analyze a 2008–2009 epidemic of the water-borne disease
Table 1. Model variables and parameters of generalized logistic model along with feasible ranges. C and K denote individuals in thousands. State and parameters are as described in [29].

| State variables | Definition | Units | Range |
|-----------------|------------|-------|-------|
| C               | Cumulative cases | people | [500, 2000] |
| Parameters      | Definition   | Units           |        |
| K               | Carrying capacity | people | [0, 0.1] |
| r               | Growth rate | days$^{-1}$ | [50, 200] |
| ti              | Inflection point | days | [0, 0.2] |
| g               | Inflection point parameter | dimensionless | [5, 0] |

in Zimbabwe [26]. However, we modify the ordinary differential equation (ODE)-based model to concentrate solely on the environmental transmission pathway. This nonlinear transmission function reflects the dose-response of the pathogen to increased contact with humans in its natural reservoir [8, 26]. The ODEs are as follows:

\[
\frac{dS}{dt} = \mu N - \beta(t)S \frac{B}{\kappa + B} - \mu S \quad (4)
\]

\[
\frac{dI}{dt} = \beta(t)S \frac{B}{\kappa + B} - (\gamma + \mu)I \quad (5)
\]

\[
\frac{dR}{dt} = \gamma I - \mu R \quad (6)
\]

\[
\frac{dB}{dt} = \xi I - \delta B \quad (7)
\]

\[
\beta(t) = \beta e^{\alpha t}. \quad (8)
\]

The model notation and parameter values are summarized in Table 2. Population estimates were adapted from [6]. We also include a time-dependent transmission rate $\beta(t)$, with its own decay parameter $\alpha$, to account for time-dependent changes in the virulence of the cholera pathogen. We consider a 8-dimensional parameter space, whose ranges were suggested by [12, 16]. For comparison with our dataset, a variable representing the cumulative number of cases is described as follows:

\[
\frac{dC}{dt} = \beta(t)S \frac{B}{\kappa + B}. \quad (9)
\]

As the initial conditions for $B$ have not been clearly specified, we include them as variables for parameter estimation.

4.3. Data

Data on the number of cumulative cases reported from the current cholera outbreak in Yemen, spanning from May 22 to December 31, 2017, was used for parameter estimation [36].
Table 2. Model variables and parameters of ODE model along with feasible ranges and values. * denotes all variables being estimated. All state variables (except $B$) denote individuals in thousands.

| State variables | Definition                  | Units      | Value          |
|-----------------|-----------------------------|------------|----------------|
| $S$             | Susceptible pop.            | people     | 29932.971      |
| $I$             | Infectious pop.             | people     | 35.571         |
| $R$             | Recovered pop.              | people     | 0              |
| $B^*$           | Bacterial concentration     | cells/ml   | [0, $10^6$]    |
| $C$             | Cumulative cases            | people     | 35.571         |

Parameters

| Parameter  | Definition                        | Units     | Value          |
|------------|-----------------------------------|-----------|----------------|
| $\mu^*$   | Birth/death rate                  | days$^{-1}$ | $[10^{-5}, 5 \cdot 10^{-5}]$ |
| $k^*$     | Half-infectious bacterial concentration | cells/ml | $[10^5, 10^6]$ |
| $\gamma^*$| Recovery rate                     | days$^{-1}$ | $[\frac{1}{17}, \frac{1}{73}]$ |
| $\xi^*$   | Bacterial contamination rate      | cells/ml/day/person | [0, 10] |
| $\delta^*$| Bacterial death rate              | days$^{-1}$ | $[\frac{1}{17}, \frac{1}{3}]$ |
| $\beta^*$ | Transmission rate                 | days$^{-1}$ | [0, 1]         |
| $\alpha^*$| Transmission growth/decay rate     | days$^{-1}$ | $[-0.001, 0.001]$ |
| $N$       | Total population                  | people    | 29968.442      |

5. Results

5.1. Generalized logistic model

The global interpolant was created with 549 support nodes, a relative error of 0.94%, and an absolute error of 18. For each interpolant created at the global and local stage, we take 10,000 samples of the parameter space using LHS. We interpolate the corresponding trajectories using the global interpolant, and retain those parameters whose costs are below $10^5$ for use in the local stage. Five iterations of the local stage are completed. 10 independent runs of the algorithm (global + local) were completed, and their resulting best-fit trajectories are illustrated in Figure 2. All variants of the overall algorithm appear to qualitatively fit the trends exhibited by the data, although all the curves fail to fit the initial set of cases. On closer inspection of the quantitative differences (Table 3), there is a marked contrast in performance. Each method possesses a distinct level of fitness, as evidenced by the average SSEs and their non-overlapping confidence intervals. $k$-means edges out GMM when using local grids, but uses more model evaluations to do so. Nelder-Mead, somewhat surprisingly, outperforms all other methods to deliver a vastly superior fit with diminished computational effort. Even after five iterations at the local stage, the clustering methods are unable to provide the same level of fit. Overall, we find reductions in mean SSE from the global stage of 74, 68, and 85% for $k$-means, GMM, and Nelder-Mead, respectively.

The global stage trajectories appear to show some deviation in the first 100 days by initially over- and then underestimating the number of cases. The local stage methods mitigate this deviation. Analyses of the parameter values, shown in Table 4, suggests a later inflection point may be to blame. In fact, all three local stage methods identify earlier inflection points to accommodate later data points. Examining other parameter values obtained from the 10 runs, Table 4 shows that the global stage parameter estimates are all greater than their local counterparts. The global stage predicts a lower cumulative incidence and higher growth rate coupled with a later epidemic peak. On the other hand, the local methods, when sorted by increasing fit, predict progressively higher cumulative incidences, lower growth rates, and earlier epidemic peaks.
Figure 2. Respective fits of each variant of parameter estimation algorithm with data (black dots) for generalized logistic model. Curves represent average of 10 runs.

Table 3. Performance of sparse grid-enabled parameter estimation algorithm for generalized logistic model. Parentheses indicate 95% confidence intervals.

| Number of SSE Model Evaluations | Model | SSE       | Lower CI | Upper CI | SSE       | Lower CI | Upper CI |
|--------------------------------|-------|-----------|----------|----------|-----------|----------|----------|
| Global                         |       | 28990.21  | (26809.42, 31189.27) |       | 2086.4    | (1906.13, 2266.67) |
| k-means                        |       | 7535.15   | (6932.82, 8137.48) | 2086.4 | 1978.2    | (1846.69, 2109.71) |
| GMM                            |       | 9352.02   | (8489.74, 10214.29) | 1978.2 | 395.53    | (385.02, 406.13)   |
| Nelder-Mead                    |       | 4147.69   | (4142.1, 4153.33)    |       | 395.53    | (385.02, 406.13)   |

We include results from [29] in Table 4 for a more comparative picture as to how the epidemic evolves through our updated dataset. One of the clearest discrepancies is in final epidemic size. We predict upwards of 1-1.2 million cases of cholera by the end of the outbreak, greatly dwarfing the estimate of 767,029 proposed by [29]. The predicted final epidemic size is significantly larger when case counts from the end of 2017 are incorporated. The increased growth rate reported by Nishiura et al. [29] may have tried to compensate for the low carrying capacity in order to describe future trends with a limited dataset. Nishiura et al. [28] also conceded their early inflection point and low carrying capacity diverged from the actual trends reported after their observation period, attributing it to structural model mismatch and shifts in the prevailing epidemiological landscape.
Table 4. Parameter estimates for generalized logistic model. Parentheses indicate 95% confidence intervals. $K$ denotes individuals in thousands. $t_i$ denotes number of days elapsed since May 22, 2017. Results from [29] listed in final row.

| Method      | $K$        | $r$          | $t_i$       | $g$          |
|-------------|------------|--------------|-------------|--------------|
| Global      | 1036.718   | 0.026837     | 86.6        | 1.065        |
|             | (1019.717, 1053.687) | (0.025001, 0.028675) | (84.1, 89.1) | (0.911, 1.212) |
| $k$-means   | 1086.954   | 0.017768     | 72.9        | 0.209        |
|             | (1073.429, 1100.479) | (0.017029, 0.018506) | (71.7, 74.2) | (0.144, 0.274) |
| GMM         | 1071.146   | 0.019411     | 75.9        | 0.356        |
|             | (1062.24, 1080.052) | (0.018683, 0.020139) | (74.3, 77.6) | (0.290, 0.422) |
| Nelder-Mead | 1197.228   | 0.011586     | 55.1        | −0.383       |
|             | (1195.942, 1198.525) | (0.011553, 0.011619) | (55.0, 55.2) | (−0.386, −0.381) |
| [29]        | 767.029    | 0.18         | 37.3        | 6.9          |
|             | (690.877, 871.671) | (0.09, 0.65) | (33.1, 44.7) | (2.4, 33.2)   |

5.2. ODE model

In contrast to the generalized logistic model, it is difficult to produce an accurate global sparse grid interpolant encompassing all eight parameters of the ODE model since the corresponding dynamics induced by the various parameter ranges would be too diverse to capture with a few thousand model evaluations. In light of this obstacle, we decide to select a subset of the parameters (and their ranges) to serve as the parameter space for the global interpolant. We choose the following three parameters:

1. $\beta$: The number of cumulative cases is fairly sensitive to the transmission rate. Depending on whether transmission is fairly high ($\beta \approx 1$) or virtually non-existent ($\beta \approx 0$), the number of cases may rise exponentially or remain constant.
2. $\alpha$: This parameter is introduced into our model to modulate the influence of transmission over time. Since no prior information exists on this parameter, exploring possible values at the global stage would be beneficial.
3. $B(0)$: The initial pathogen concentration is practically unknown in all epidemic contexts, but like $\beta$, can influence the spread of cholera, especially if several orders of magnitude are explored.

After sampling 100,000 points in this 3-dimensional parameter space, we select viable parameters with costs less than $10^6$ to estimate the chosen parameters in the local stage. Five iterations of the local stage are completed. After the local stage is completed, we then perform Nelder-Mead optimization with the best local parameters to optimize the remaining five parameters. We compare these results with Nelder-Mead optimization across all eight parameters simultaneously when initialized with the best global parameters. The 3-dimensional global sparse grid interpolant required 3,297 ODE model evaluations with an estimated relative error of 0.73% and an absolute error of 220.

Table 5 presents the performance from 10 runs of the parameter estimation algorithm. Reductions in SSE of at least 95% are observed for all three local methods when compared against the global stage, an observation depicted visually in Figure 3. $k$-means and GMM surpass Nelder-Mead in computational effort but deliver superior fits, reducing the SSE by half. Both clustering methods prove to be reasonably similar in determining better fits after each iteration. Figure 3 illustrates the model simulations aligning with the given dataset.
Table 5. Performance of sparse grid-enabled parameter estimation algorithm for ODE model. Parentheses indicate 95% confidence intervals.

| Method       | SSE               | Number of Model Evaluations |
|--------------|-------------------|----------------------------|
| Global       | 114150.31 (13274.26, 215026.36) | —                          |
| k-means      | 3609.33 (2629.36, 4589.3)       | 4089.6 (3314.62, 4864.58)  |
| GMM          | 2969.36 (2936.44, 3002.28)      | 4660.22 (4196.76, 5123.69) |
| Nelder-Mead  | 6233.7 (2680.26, 9787.14)       | 1276.2 (1094.37, 1458.03)  |

Figure 3. Respective fits of each variant of parameter estimation algorithm with data (black dots) for ODE model. Curves represent average of 10 runs. Shaded regions indicate 95% confidence intervals.

The global stage trajectories are relatively uncertain, as the shaded regions indicate. There is a significant deviation starting at the 125 day mark as the average starts to diverge from the data, incurring higher costs in the process. In spite of the higher SSE attributed to the Nelder-Mead optimization, the other three curves are visually indistinguishable from one another in terms of faithfully representing the data.

Table 6 displays the parameter estimates from the 10 replications. The local methods output lower population birth/death rates, indicative of declining populations. Lower values of \( \mu \) may be more consistent with the time period observed in the dataset so as to render the population essentially constant. Also of interest are the values of \( \beta, \alpha \), and \( B(0) \). \( \beta < 1 \) seems to be preferred, with \( \alpha < 0 \) suggesting a slightly declining transmission rate over the nearly 250 day period. Considering that \( B(0) \) has a range that varies by several orders
Table 6. Parameter estimates for ODE model. Parentheses indicate 95% confidence intervals.

|       | μ         | κ         | \( γ^{-1} \) | \( ξ \) |
|-------|-----------|-----------|-------------|--------|
| Global| 4.3E-5    | 1000000   | 2.9         | 10     |
| K-means| -2.4E-4  | 1036693   | 3.03        | 11.2   |
|       | (1E-4, -7.4E-5) | (1003048, 1070338) | (2.79, 3.26) | (10.2, 12.1) |
| GMM   | 3.77E-5   | 952658    | 2.72        | 9.6    |
|       | (-2.8E-4, 3.6E-4) | (888281, 1017036) | (2.48, 2.95) | (8.7, 10.5) |
| Nelder-Mead| 1.95E-5 | 1028785   | 2.91        | 10     |
|       | (-2.1E-5, 6E-5) | (872603, 1184966) | (1.79, 4.04) | (7.0, 13.0) |

Table 7. Number of cases developed by the end of 2018 and basic reproductive numbers for each variant of sparse grid-enabled parameter estimation algorithm. Parentheses indicate 95% confidence intervals.

|       | \( C(December \, 31, \, 2018) \) | \( R_0 \) |
|-------|----------------------------------|---------|
| Global| 1090.585 (830.644, 1350.525)     | 1.86 (1.50, 2.22) |
| K-means| 1099.945 (1082.99, 1116.901)     | 1.43 (1.30, 1.57) |
| GMM   | 1095.333 (1092.661, 1098.006)    | 1.53 (1.45, 1.61) |
| Nelder-Mead| 1164.957 (1049.287, 1280.627)  | 1.62 (1.33, 1.91) |

of magnitude, clustering may be the key to effectively grouping and detecting successful parameter fits. The algorithm tended to favor initial bacterial concentrations between 10 and 10,000 cells/ml. Finally, we forecast the number of cases developed by the end of 2018, and the basic reproductive number \( R_0 \), with the particular parameters these methods have determined. Table 7 shows agreement with the \( K \) values of the generalized logistic model on the final epidemic size, as both anticipate 1-1.2 million cases by the end of 2018. As for \( R_0 \), we predict the disease will not die out since \( R_0 > 1 \). The number of infections are forecast to only increase at this time.

6. Discussion

This work has exploited the approximation properties of sparse grid interpolation to calibrate two epidemiological models of differing complexity. Sparse grid interpolants summarize diverse ensembles of smooth model trajectories using compact multidimensional polynomials, enabling quicker, efficient scans of the parameter space. Our parameter estimation algorithm selects and samples local subspaces, informed by cluster analysis, to repetitively narrow a large search space in the hopes of determining progressively better fits to observed data. Definitive gains in parameter estimation are expected when transitioning from global to local spaces. We show through two examples that definitive gains in performance can be made to produce simulated outcomes consistent with available epidemiological data. The global-local paradigm enables a significant improvement in fitness.
Once suitable candidates at the global stage are identified, the algorithm can isolate and refine several neighbourhoods, or niches, of promising parameters.

For our case study in parameter estimation, we present results pertaining to both a phenomenological (generalized logistic) and mechanistic (ODE) analyses of the cholera outbreak in Yemen. Results from both models confirm the escalation of the epidemic above the 1 million case mark by its conclusion. The generalized logistic model simplifies the epidemiology of cholera to an empirical relationship between disparate phenomena. On the other hand, the ODE model explicitly represents mechanisms of transmission, contamination, demography, and recovery. Yet, both models comparably elucidate the Yemen outbreak data to provide contending theories as to the growth of the epidemic, a testament to the diversity of cholera models. This is not surprising as [23] stressed the importance of broader data collection beyond case reports in their comparative modelling study. Existing studies of the Yemen outbreak examine reporting biases [29], environmental drivers [6], and medical resources [17]. To compare the overall quality of both models in this context, we make use of the Akaike Information Criterion (AIC). We calculate AIC values of 141.76 and 146.97 for the generalized logistic model and the ODE model, respectively. So while the simpler model may be favoured holistically, the larger model achieved the better fit.

One feature of this work not to be overlooked is the integration of an unconstrained, direct search method like Nelder-Mead with sparse grid interpolation. The initial scan of the entire parameter space using the global interpolant supplies candidate parameters for Nelder-Mead optimization. Both steps compliment each other well. The advantages in deploying an unconstrained optimization technique like Nelder-Mead is that the search for an optimal point does not adhere to bound constraints like parameter ranges, seeking minimizers wherever they may be. Venturing outside the prescribed parameter ranges reaps rewards for both models. Specifically, we find that both $g$ and $\mu$, parameters assumed to be positive, converge to negative values. Such information updates our knowledge of these parameters. However, Nelder-Mead optimization works differently for both models. For the first model, initiating the simplex method from the global stage seems to be sufficient. This is not the case for the second model. When starting from the global stage, it cannot locate better optima than when starting from the local stage, after the clustering methods have identified the first three parameters. The higher dimensionality and presence of local minima may be complicating factors for successfully calibrating the ODE model using Nelder-Mead optimization. In spite of this difficulty, we observe improvements when pairing the clustering methods, which estimates three of the eight parameters of the ODE model, with Nelder-Mead, which estimates the rest.

For the purposes of parameter inference in epidemiology, the parameter estimation algorithm presented here can easily be extended to more complicated models, such as metapopulation models, whose parameters are more numerous yet spatially refined. Fitting parameters for larger models will provide more detailed information on the local status of a disease and facilitate computation of regional epidemiological metrics. Understanding the epidemic on the local level will enable researchers to craft interventions better tailored to the afflicted population, saving lives and halting the spread of devastating diseases. Furthermore, our cluster selection criteria enable optimal coverage of the parameter space previously inhabited by parameters obtained in the global stage. However, the use of cluster analysis comes into question when a local grid range delineated by a cluster becomes too large. The resulting local interpolant can become poor in quality, a problem exacerbated
by the dimensionality of the parameter space, as is the case for the ODE model, forcing us to select three of the possible eight parameters. While we experience no such obstacle in a low-dimensional parameter space, it is a potential hurdle that will have to be overcome by designing more advanced cluster selection criteria, which may produce more clusters occupying smaller spaces. However, an increase in granularity will inevitably lead to an increase in computational effort.

While sparse grid interpolation permits global interpolation of the model dynamics, we cannot claim for certain that the results obtained from our parameter estimation strategy are truly optimal, except that based on our ensemble of simulations, we found no alternative optima. The limitations in resolving all dynamics over large parameter ranges or dimensions prevents a comprehensive analysis. However, we are confident that the large gains in matching simulation outcomes with existing data exhibited by our computational experiment presents a compelling parameter estimation procedure for similar models. Moreover, the model dynamics captured by each interpolant will enable further analysis to tweak and tune parameter values and ranges. The uses for sparse grid interpolation are visibly abundant. Any situation requiring the clarification of unknown or uncertain model dynamics would benefit from its use.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Aditya Sai  http://orcid.org/0000-0003-2887-479X

References

[1] O. Akman and E. Schaefer, An evolutionary computing approach for parameter estimation investigation of a model for cholera, J. Biol. Dyn. 9 (2015), pp. 147–158.
[2] V. Barthelmann, E. Novak, and K. Ritter, High dimensional polynomial interpolation on sparse grids, Adv. Comput. Math. 12 (2000), pp. 273–288.
[3] J.N. Bazil, G.T. Buzzard, and A.E. Rundell, A global parallel model based design of experiments method to minimize model output uncertainty, Bull. math. biol. 74 (2012), pp. 688–716.
[4] L.M.A. Bettencourt and R.M. Ribeiro, Real time bayesian estimation of the epidemic potential of emerging infectious diseases, PLoS ONE 3 (2008). pp. e2185.
[5] H.J. Bungartz and M. Griebel, Sparse grids, Acta Numer. 13 (2004), pp. 147–269.
[6] A. Camacho, M. Bouhenia, R. Alyushi, A. Alkohlanli, M.A.M. Naji, X. de Radiguës, A.M. Abubakar, A. Almoalimi, C. Seguin, M.J. Sagrado, M. Poncin, M. McRae, M. Musokey, A. Rakesh, K. Porten, C. Haskew, K.E. Atkins, R.M. Eggo, A.S. Azman, M. Broekhuijsen, M.A. Saatcioglu, L. Pezzoli, M.L. Quilici, A.R. Al-Mesbahy, N. Zagaria, and F.J. Luquero, Cholera epidemic in Yemen, 2016–18: An analysis of surveillance data, Lancet Glob. Health 6 (2018), pp. e680–e690.
[7] G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, Infect. Dis. Model. 2 (2017), pp. 379–398.
[8] C.T. Codeço, Endemic and epidemic dynamics of cholera: The role of the aquatic reservoir, BMC Infect. Dis. 1 (2001), pp. 209.
[9] F.C. Coelho, C.T. Codeço, and M.G.M. Gomes, A Bayesian framework for parameter estimation in dynamical models, PLoS ONE 6 (2011), pp. 1–6.
[10] M.M. Donahue and G.T. Buzzard, Parameter identification with adaptive sparse grid-based optimization for models of cellular processes, Methods Bioeng. (2009), pp. 1–32.
[11] M.M. Donahue, G.T. Buzzard, and A.E. Rundell, Robust parameter identification with adaptive sparse grid-based optimization for nonlinear systems biology models, in 2009 American Control Conference, St. Louis, MI, USA. IEEE, 2009, pp. 5055–5060.

[12] I.C.H. Fung, Cholera transmission dynamic models for public health practitioners, Emerg. Themes Epidemiol. 11 (2014), pp. 1–11.

[13] T. Gerstner and M. Griebel, Numerical integration using sparse grids, Numer. Algorithms 18 (1998), pp. 209–232.

[14] T. Gerstner and M. Griebel, Dimension-adaptive tensor-product quadrature, Computing 71 (2003), pp. 65–87.

[15] T. Gerstner and M. Griebel, Sparse grids, in Encyclopedia of Quantitative Finance, Vol. 13, John Wiley & Sons, Ltd, Chichester, UK, 2010, pp. 147–269.

[16] Y.H. Grad, J.C. Miller, and M. Lipsitch, Cholera Modeling, Epidemiology 23 (2012), pp. 523–530.

[17] D. He, X. Wang, D. Gao, and J. Wang, Modeling the 2016-2017 Yemen cholera outbreak with the impact of limited medical resources, J. Theor. Biol. 451 (2018), pp. 80–85.

[18] L. Held, Bayesian Methods in Epidemiology, in Handbook of Epidemiology, Springer New York, New York, NY, 2014, pp. 1161–1193.

[19] G. Katriel, R. Yaari, A. Huppert, U. Roll, and L. Stone, Modelling the initial phase of an epidemic using incidence and infection network data: 2009 H1N1 pandemic in Israel as a case study, J. R. Soc. Interface 8 (2011), pp. 856–867.

[20] A. Klimke, Sparse grid interpolation toolbox user’s guide, Tech. Rep., 2008.

[21] A. Klimke and B. Wohlmuth, Algorithm 847: spinterp: Piecewise multilinear hierarchical sparse grid interpolation in MATLAB, ACM Trans. Math. Softw. 31 (2005), pp. 561–579.

[22] J.C. Lagarias, J.A. Reeds, M.H. Wright, and P.E. Wright, Convergence properties of the nelder–mead simplex method in low dimensions, SIAM J. Optim. 9 (1998), pp. 112–147.

[23] E.C. Lee, M.R. Kelly, B.M. Ochocki, S.M. Akinwumi, K.E. Hamre, J.H. Tien, and M.C. Eisenberg, Model distinguishability and inference robustness in mechanisms of cholera transmission and loss of immunity, J. Theor. Biol. 420 (2017), pp. 68–81.

[24] T. Mdluli, G.T. Buzzard, and A.E. Rundell, Efficient optimization of stimuli for model-based design of experiments to resolve dynamical uncertainty, PLOS Comput. Biol. 11 (2015), pp. e1004488.

[25] C.G. Moles, P. Mendes, and J.R. Banga, Parameter estimation in biochemical pathways: A comparison of global optimization methods, Genome res. 13 (2003), pp. 2467–74.

[26] Z. Mukandavire, S. Liao, J. Wang, H. Gaff, D.L. Smith, and J.G. Morris, Estimating the reproductive numbers for the 2008–2009 cholera outbreaks in Zimbabwe, Proc. Natl. Acad. Sci. 108 (2011), pp. 8767–8772.

[27] J.A. Nelder and R. Mead, A simplex method for function minimization, Comput. J. 7 (1965), pp. 308–313.

[28] H. Nishiura, S. Tsuzuki, and Y. Asai, Forecasting the size and peak of cholera epidemic in Yemen, 2017, Future Microbiol. 13 (2018), pp. 399–402.

[29] H. Nishiura, S. Tsuzuki, B. Yuan, T. Yamaguchi, and Y. Asai, Transmission dynamics of cholera in Yemen, 2017: A real time forecasting, Theor. Biol. Med. Model. 14 (2017), pp. 1316.

[30] A. Rinnooy Kan and G. Timmer, Stochastic global optimization methods part i: Clustering methods, Math. Programming 39 (1987), pp. 27–56.

[31] A. Smirnova and G. Chowell, A primer on stable parameter estimation and forecasting in epidemiology by a problem-oriented regularized least squares algorithm, Infect. Dis. Model. 2 (2017), pp. 268–275.

[32] S. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Sov. Math., Dokl. 4 (1963), pp. 240–243.

[33] J. Sun, J.M. Garibaldi, and C. Hodgman, Parameter estimation using meta-heuristics in systems biology: A comprehensive review, IEEE/ACM Trans. Comput. biol. bioinformatics/IEEE, ACM 9 (2012), pp. 185–202.
[34] L.F. White and M. Pagano, *Transmissibility of the influenza virus in the 1918 pandemic*, PLoS ONE3 (2008), pp. e1498

[35] L.F. White, J. Wallinga, L. Finelli, C. Reed, S. Riley, M. Lipsitch, and M. Pagano, *Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA*, Influenza Respir. Viruses 3 (2009), pp. 267–276.

[36] WHO, *Situation report no. 40 - Yemen conflict*, Tech. Rep. 40, 2017.

[37] WHO, *Weekly Epidemiological Bulletin - W52 2017 (Dec 25 - Dec 31)*, Tech. Rep., 2017.

[38] R. Yaari, G. Katriel, L. Stone, E. Mendelson, M. Mandelboim, and A. Huppert, *Model-based reconstruction of an epidemic using multiple datasets: Understanding influenza A/H1N1 pandemic dynamics in Israel*, J. R. Soc. Interface 13 (2016), pp. 20160099.

[39] H. Yserentant, *Sparse grids, adaptivity, and symmetry*, Computing (Vienna/New York) 78 (2006), pp. 195–209.