An evolutionary computing approach for parameter estimation investigation of a model for cholera

Olcay Akman\textsuperscript{a} and Elsa Schaefer\textsuperscript{b}

\textsuperscript{a}Department of Mathematics, Illinois State University, Normal, IL 61790, USA; \textsuperscript{b}Department of Mathematics, Marymount University, Arlington, VA 22207, USA

(Received 20 June 2014; accepted 6 April 2015)

We consider the problem of using time-series data to inform a corresponding deterministic model and introduce the concept of genetic algorithms (GA) as a tool for parameter estimation, providing instructions for an implementation of the method that does not require access to special toolboxes or software. We give as an example a model for cholera, a disease for which there is much mechanistic uncertainty in the literature. We use GA to find parameter sets using available time-series data from the introduction of cholera in Haiti and we discuss the value of comparing multiple parameter sets with similar performances in describing the data.

Keywords: genetic algorithms; cholera; parameter fitting; parameter estimation

AMS Subject Classification: 92B01; 92B20; 62B10

1. Introduction

The purpose of this paper is to suggest an intuitive and easily implemented method to inform parameter values for a given deterministic model using available time-series data. The immediate investigation was spurred by the study of several models for cholera, a disease for which many mechanistic studies are in contradiction and for which the available data are flawed, thus making parameter selection challenging.

We consider a model with 14 unknown parameters and use comparison with time-series data to choose optimal values for the parameters. Clearly, a stochastic approach is needed, since if \( n \) values are allowed for each of \( m \) model parameters, a search grid for parameter sets would have \( n^m \) parameter combinations. In the case of cholera, we have seen references to an Markov Chain Monte Carlo approach for data fitting [1,37], the use of the least-squares fitting routine lsqcurvefit from the MATLAB (MATLAB and Parallel Computing, Optimization, and Statistics and Machine Learning Toolboxes Release 2014a, The MathWorks, Inc., Natick, Massachusetts, USA) optimization toolbox [29], and the application of semi-parametric regression [22]. The focus of the references is understandably on the characteristics of the system being studied and
conclusions for the control of cholera, and any parameter fitting is performed in the background with minimal explanation. The purpose of this paper is not to evaluate the current methods available for parameter estimation, both for cholera models or more generally in any deterministic modelling setting, but rather to suggest an approach that is easy to implement, that is independent of a particular software package, and that uses the principles of natural selection in a process called evolutionary computing.

In the following, we introduce the concept of genetic algorithms (GAs), and also, for the sake of illustration, we present a sample model for cholera. We show how, in the absence of distributional assumptions, the principles of natural selection make GAs an invaluable tool in modelling. The parameter sets that are output in the GA process additionally give insight into the relative contributions of certain parameters that are needed to fit the observed data, and can add to the important information gained through a sensitivity analysis.

2. GAs: a brief overview

A GA is an optimization technique which is inspired by biological evolution under the influence of natural selection. Evolutionary search algorithms – GAs – were first introduced by Goldberg [12] and were further developed by Holland [17]. This technique has been shown to be valid and capable of dealing with highly multimodal and discontinuous search landscapes where traditional optimization techniques fail [4]. In these settings, the traditional optimization methods may get trapped in the local optimal points while searching within multimodal landscapes. The technique of GAs was developed to avoid this problem.

GAs are usually applied by using computer simulations in which the optimization problem is specified. In such an optimization problem, members of candidate solutions, called individuals, are represented by the biological term chromosomes. The GA process is an iterative working process in which a set of individuals called the population evolves towards an objective, or fitness, function. The evolutionary process of a GA is a highly simplified simulation of the biological version. GA starts with a population of individuals randomly generated according to some probability distribution, generally chosen to be a uniform distribution unless there is a priori knowledge suggesting a differing distribution. The GA process updates this population in steps called generations. A population of possible solutions, that is, the chromosomes, are evaluated using the objective function and thus given a fitness value. The chromosomes with the best fitness values are allowed to mate with other chromosomes, mutate, and move on to the next generation. This process is repeated until a predetermined stopping rule is met: when there is only a small chance to obtain a better solution due to the diminishing improvement of previous generations, that is, minute chance to produce an individual with a fitness value that is significantly better than the members of the most recent generation. At the end of this process, the individual having the best fitness value is considered to be the solution of the optimization problem. Implementing GAs has a unique advantage due to its evolutionary nature.

Most traditional GA research has concentrated in numerical function optimization. GAs have been shown to be able to outperform conventional optimization techniques on aggressively oscillating, discontinuous, multimodal, or noisy functions [4]. Perhaps the most delightful application of GAs is the travelling salesperson problem, where the task is to find the shortest route for visiting a specified group of cities. Among others, [18–20] are just a few of the studies where GAs were implemented in travelling salesperson problem. Ref. [16] successfully used GAs in building phylogenetic trees showing the inferred evolutionary relationships among various biological species based on similarities or differences in their morphological and/or genetic characteristics. Applications of GAs extend to medicine [43], biotechnology [44], and transportation [10], among many others. In addition, the abundance of GA software in the literature brings a GA
parameter estimation approach within the reach of many researchers. Although we have used our own code, our results have been replicated using commercially available packages such as MATLAB, running under the Optimization toolbox.

3. Description of a model for cholera

The inspiration for the current investigation has its origins in a desire to improve and validate several mechanistically driven models for cholera. With significant outbreaks in Zimbabwe and later Haiti during the past decade, much thought has been given to modelling cholera in recent years. The ultimate goal, of course, of these efforts is to better understand the dynamics of a cholera outbreak so that we can greatly reduce the morbidity which results from the disease. There is a disconnect between the mechanistic understanding that we can infer from a microbiology/human studies perspective [2,5,11,21,26,27,30–34,36,40] and the simplified models chosen, in part, due to the limited data available to justify parameter choices for more complex models [7,28,29,37,41,42]. Compounding this disparity is the inconsistency regarding choices for key parameter values across many carefully researched papers by top scientists. Since the focus of this work is not cholera, but rather a technique for using available time-series data to inform parameter choices, we save a careful discussion of these discrepancies for a future work, and simply present a mechanistically complex model for cholera for which we desire appropriate parameter choices (Figure 1).

We consider here a model which was developed by Schaefer with Fister, Gaff, and Lenhart, and a similar model appears in PDE form [8] in an age-based consideration of cholera dynamics. The model structure allows testing of hypotheses regarding the role of a partially immune class of individuals in the spread of cholera.

Humans with no immunity:

\[
\frac{dS}{dt} = -\beta_B \frac{BS}{\kappa + B} - \beta_{I_s} \frac{I_A S}{N} - \beta_{I_s} \frac{I_s S}{N} + bN - dS + \omega_{s} \hat{S},
\]

Figure 1. The SSIIRRB model allows for a partially immune class of humans with differing parameters for infectivity, length of illness, and length of immunity.
\[
\frac{dI_S}{dt} = p \left[ \frac{\beta_B BS}{\kappa + B} + \beta_{I_S} I_S \frac{I_S}{N} + \beta_{I_A} I_A \frac{I_A}{N} \right] - dI_S - \gamma_{I_S} I_S,
\]
\[
\frac{dR_S}{dt} = -dR_S + \gamma_{I_S} I_S - \omega_{R_S} R_S. 
\] (1)

Humans with partial immunity:
\[
\frac{d\hat{S}}{dt} = -\beta_B \frac{B\hat{S}}{\kappa + B} - \beta_{I_S} \frac{I_S\hat{S}}{N} - \beta_{I_A} \frac{I_A\hat{S}}{N} - d\hat{S} + \omega_{R_A} R_A + \omega_{R_S} R_S - \omega_{S\hat{S}},
\]
\[
\frac{dI_A}{dt} = \beta_B \frac{B\hat{S}}{\kappa + B} + \beta_{I_S} I_S \frac{\hat{S}}{N} + \beta_{I_A} I_A \frac{\hat{S}}{N} + (1-p) \left[ \frac{\beta_B BS}{\kappa + B} + \beta_{I_S} I_S \frac{I_S}{N} + \beta_{I_A} I_A \frac{I_A}{N} \right] - dI_A - \gamma_{I_A} I_A,
\]
\[
\frac{dR_A}{dt} = -dR_A + \gamma_{I_A} I_A - \omega_{R_A} R_A. 
\] (2)

Bacteria:
\[
\frac{dB}{dt} = \eta_{I_A} I_A + \eta_{I_S} I_S - \delta B. \] (3)

The model notation and parameter values are summarized in Table 1. Here, we wish simply to highlight the discordance within the literature concerning some important parameter values. First, we note that in [7], the authors consider the question of identifiability of parameters in a simple susceptible-infected-water-removed/recovered model for cholera, and they find that in the presence of noise in the data, parameter identifiability is simply not possible. In light of the tremendous noise in the available data and the likely interdependencies of the parameters, we do not expect to find unique parameter sets with best performance in describing the cholera data from Haiti.

There is disagreement regarding the relative importance of asymptomatic and symptomatic infections [11,22,31,37], which would influence the values of the person-to-person infectivity parameters \( \beta_{IS} \) and \( \beta_{IA} \), as well as the shedding rates \( \eta_{I_A} \) and \( \eta_{I_S} \). We further note that none of these parameters could be determined empirically with the available data. Also, the relative contact of individuals with the water supply versus the immediate contact with human-contaminated household items, \( \beta_B \) versus \( \beta_{I_A} \) and \( \beta_{I_S} \) is unknown, and while earlier articles have attempted to quantify \( \beta_B \) based on drinking rates [3,15], we are not certain this practice adequately captures the infectivity due to ingestion of water. The Michaelis constant \( \kappa \) may also be poorly estimated – there is tremendous variation in the amount of bacteria needed to cause infection [21,31] – but we feel that variation in the corresponding rate \( \beta_B \) alone is adequate for an investigation – there is no need to vary both \( \beta_B \) and \( \kappa \) given the redundancy in the way those changes would impact model output.

The parameter \( p \) describes the proportion of infections among those with no immunity that would be symptomatic (versus asymptomatic), and not only is there the uncertainty as described above about the rate of symptomatic infections in a given outbreak, but also it is not clear what proportion of total symptomatic infections are from disease-inferred partial immunity versus a natural resistance to disease.

A recent critique about models for cholera pointed out that many research groups assume that bacteria in the water are essentially nonviable in 30 days and suggests that in fact bacteria may survive anywhere from 3 to 40 days, causing an enormous difference in model predictions [13]. Thus, we include large variation for the parameter \( \delta \), the ‘death’ rate for bacteria, in our parameter search.
Table 1. Model variables and parameters along with feasible ranges.

| Variables          | Description                                      | Feasible Range       |
|--------------------|--------------------------------------------------|----------------------|
| $S(t)$             | Susceptible humans                               |                      |
| $\hat{S}, \hat{I}$ | Susceptible with partial immunity                |                      |
| $I_S(t)$           | Symptomatic infected humans                      |                      |
| $I_A(t)$           | Asymptomatic infected humans                     |                      |
| $R_S(t)$           | Humans recovered from symptomatic infection      |                      |
| $R_A(t)$           | Humans recovered from asymptomatic infection     |                      |
| $N(t)$             | Total population in all human compartments       |                      |
| $B(t)$             | Concentration of vibrio in environment           |                      |

Parameters with fixed values

| Parameter | Description                                      | Value         |
|-----------|--------------------------------------------------|---------------|
| $b$       | Natural birth rate of humans                     | 0.000078      |
| $d$       | Natural death rate of humans                     | 0.000025      |
| $\kappa$  | Half-saturation constant of vibrios              | $10^6$        |

Parameters chosen using GA approach

| Parameter | Description                                      | Feasible Range |
|-----------|--------------------------------------------------|----------------|
| $h$       | Proportion of (symptomatic) infected hospitalized | $0 - 1$       |
| $p$       | Prob. of new infected from $S$ to be asymptomatic | $0.001 - 1$   |
| $\beta_B$ | Ingestion rate of vibrio from environment       | $0.001 - 0.5$ |
| $\beta_{IS}$ | Infectivity for susceptible– symptomatic infected | $.001 - .5$  |
| $\beta_{IA}$ | Infectivity for susceptible– asymptomatic infected | $.001 - .5$  |
| $\gamma_S$ | Cholera recovery rate for symptomatic infected | $\frac{1}{14} - \frac{1}{2}$ |
| $\gamma_A$ | Cholera recovery rate for asymptomatic infected  | $\frac{1}{7} - 1$ |
| $\omega_{RS}$ | Rate of waning immunity from $R_A$ to $\hat{S}$ | $0.0001 - \frac{1}{1000}$ |
| $\omega_{RS}$ | Rate of waning immunity from $R_S$ to $\hat{S}$ | $0.00001 - \frac{1}{3600}$ |
| $\omega_S$ | Rate of waning immunity from $S$ to $\hat{S}$    | $0.0001 - \frac{1}{1000}$ |
| $\eta_A$ | Rate of contribution by asymptomatic infected    | $0.001 - 10$  |
| $\eta_S$ | Rate of contribution by symptomatic infected     | $0.001 - 10$  |
| $\delta$ | Death rate of vibrios from HI to non-HI state    | $\frac{1}{360} - \frac{1}{3}$ |
| $r$       | Initial proportion of $\chi$ in the environment  | $0.001 - 1$   |

4. The GA approach

To illustrate the use of GA for parameter selection, we consider the model for cholera given in Section 3, and we seek choices for the model parameters so that the resulting model best describes the data provided by the Haitian Ministry of Health for numbers of hospitalizations due to cholera for the first 438 days of the initial cholera outbreak in Nord Department [35]. In addition to the work presented in this manuscript, we also considered four additional models for cholera and data from all departments. We will write about the wider results in a later publication that explores techniques for model comparison and selection.

As discussed in the previous section, while specific values for most parameters needed for the model are not known, we are able to suggest realistic ranges for each parameter value as described in Section 3 and listed in Table 1. We then employ an evolutionary computing approach to identify parameter sets within these feasible ranges that provide strong goodness of fit with the given Haitian data. In our particular case, given a parameter set, we solved the system given in Equations (1)–(3). From this, we could calculate the number of individuals who would be newly infected on each day of the simulation – note that we are tracking the incidence and not the prevalence. This incidence function, scaled using parameter $h$ of hospitalized cases, was compared to the time-series data provided by the Ministry of Health [35]. We calculated the average error between the data and the simulated incidence on each day, and we thus defined the goodness of fit of the model paired with the given parameter set. The goodness of fit is a measure of the error between the model and the data, and we note that minimizing this error is equivalent to maximizing the fitness assigned to a parameter set.
We seek the ‘most fit’ parameter set through the following steps:

4.1. **Step 1: Latin hypercube sampling**

We first use a Latin hypercube sampling (LHS) scheme to randomly select \( N \) parameter sets within the allowed ranges. The LHS code at the Kirshner Lab website \([23]\) provides an excellent starting point for this process. We use a uniform distribution over our ranges, and we find that the use of at least \( N = 5000 \) parameter sets of length \( \ell = 14 \) as the starting population provides stable performance.

4.2. **Step 2: Create a single ‘elite’ individual**

Beginning with the \( N \) parent genes, we seek to create a single elite individual by creating new generations until a stopping criterion is met. There are many ways in which this can reasonably accomplished with similar results. We chose the following:

1. First, we retain the top-performing 10% of the parents for the next generation; these are the generational elite individuals. We found that the method is insensitive to the percentage of parents retained in its ability to find similarly performing elites provided that the population pool is not small.

2. Every 10th generation, we allow migration. We repeat the LHS process to create .1N ‘fresh’ parent parameter sets for the next generation. Migration is introduced to prevent the scheme from obtaining local rather than global optimization results.

3. The remaining slots in our new generation are filled by creating children. To do this, we randomly choose one parent from the top-performing genes identified in number (1). We then randomly choose any parent from the prior generation without regard to the modelling error. Once two parents are identified, we complete a mating process. We identify four mating approaches below. We found that approach (a) was most effective in consistently selecting top-performing elite individuals. In our step of creating single ‘elite’ individuals, we limit our attention to approach (a), but in Step 3 below we apply all methods sequentially so that we might benefit from the advantages of each.

   (a) Uniform crossover: To create the child, we consider each of the \( \ell \) parameters (genes) individually, randomly selecting the parameter (gene) value from one parent or the other parent. An advantage of this approach is its ability to preserve gene values that are on the endpoints of the allowed range. The disadvantage is that this method does not allow for fine-tuning between existing gene values.

   (b) Single-point crossover: Rather than consider one single parameter at a time, we splice the parent sets so that the first \( s \) parameters are from parent 1, while the second \( \ell - s \) parameters are from parent 2. This method does not allow as much generational change as the previous method.

   (c) Gene-averaging crossover: To create the child, each of the \( \ell \) parameter (gene) values is chosen by averaging the respective values for each parent. The advantage of this approach is that it allows fine-tuning of parameter sets that may be near optimal, and a disadvantage is that the averaging will encourage movement away from the endpoints of the allowed parameter ranges.

   (d) Hybrid crossover: This final method combines the first and third approaches. For each of the offspring’s genes, a random choice is made to either average the two parents’ gene values, accept parent 1’s gene value, or accept parent 2’s gene value.
Once the offspring have been generated, we allow mutation to modify randomly selected genes of no more than 5% of the individuals. To accomplish this, we randomly select up to 5% of the offspring to be subject to mutation. This proportion, as well as the other values used in various selection steps are conventionally accepted values for optimal computing time or convergence, not much different than the traditional 95% level for a typical confidence interval. Most commercially available GA software allow users change these values. For each selected offspring, each of the \( \ell \) parameters has a \( \frac{1}{\ell} \) chance of being mutated. If selected for mutation, a gene is replaced by a new parameter value that is randomly chosen from the feasible range.

We continue to create new generations until we satisfy a stopping criterion. Given \( N \) members in a generation, for each member we compute the model’s goodness of fit for the given data set (as defined at the beginning of Section 4). We then calculate the average goodness of fit over all \( N \) members. We sum the absolute value of the difference between each population member’s goodness of fit and the average goodness of fit, and we name this quantity \( \text{stoperr} \). When the value of \( \text{stoperr} \) is sufficiently small (we used 0.001), we stop.

Once the stopping criterion is met, or we deplete our predetermined resources, we choose the population member with the lowest error, and this is our ‘elite’ individual.

**Step 2 Modification: shrinking population size**

Following Hallam et al. [14], we implement an adaptive scheme which allows a decrease in populations in relation to the performance of each generation. This allows us to cast a wide net in the LHS sampling scheme for our initial population, but to quickly lower our computational cost by decreasing the population size once we have benefited from the initial sweep. The shrinking method was not sensitive to precise rules but would not consistently converge if small populations were allowed to shrink ‘too quickly’. We used the following algorithm:

1. If the initial population size is greater than 1000, we immediately reduce the population to the most fit 1000 before applying any GA techniques.
2. If the population size is larger than 300, we reduce the size by 80% in each generation.
3. If the population size is larger than 100 and the \( \text{stoperr} < 100 \), then we reduce the population size by 50%.
4. Otherwise, in any generation satisfying \( \text{stoperr} < 1 \), we reduce the population size by 10% in that generation.

This algorithm allowed us to find elite individuals with equivalent fitnesses to those found from the non-adaptive process, but at a tremendous computational time benefit.

**4.3. Step 3: Apply the evolutionary process to \( M \) elite individuals to choose one or more parameter sets**

We have found that some parameter optimization techniques, such as MATLAB’s lsqcurvefit, do not employ stochasticity and thus return identical parameter choices with each run. Given that the data itself are flawed, our viewpoint was that a single ‘best’ parameter set that describes that data are not valuable; rather, we would like to apply multiple similarly performing parameter sets to our model to investigate a range of possible simulation outcomes depending on the range of reasonable parameter choices available to describe the data.

Thus, as an additional step, due to stochasticity, as well as in some cases the lack of uniqueness of the most fit parameter set, we implement the previous algorithm multiple times to obtain, say, \( M \) elite individuals. In our particular case, while each elite individual found for a given model/data set has a similar error when compared to the data, the individual parameter choices often show variation. In other words, we repeat Steps 1 and 2 above \( M \) times resulting in \( M \) elite
Table 2. Step 3 results from GA fitting of the model to data from Nord Department in Haiti.

| $h$     | $\beta_B$ | $\beta_{I_A}$ | $\beta_{I_S}$ | $\gamma_A$ | $\gamma_S$ | $\omega_{R_A}$ | $\omega_{R_S}$ | $\omega_S$ | $\eta_A$ | $\eta_S$ | $p$ | init prop $\kappa$ | $\delta$ | Fit   |
|---------|-----------|----------------|----------------|-----------|----------|----------------|----------------|-----------|----------|---------|-----|-------------------|-------|-------|
| 0.080845 | 0.139670  | 0.255439       | 0.182161       | 0.739927  | 0.138097 | 0.010873      | 0.002322       | 0.000933  | 8.572746  | 0.001589 | 0.136353  | 0.007276 | 0.306280 | 29.257 |
| 0.080845 | 0.144565  | 0.255439       | 0.182161       | 0.739927  | 0.138337 | 0.010873      | 0.002322       | 0.000933  | 8.572746  | 0.000154 | 0.136353  | 0.007276 | 0.306280 | 29.373 |
| 0.080845 | 0.144565  | 0.255439       | 0.182161       | 0.739927  | 0.138337 | 0.010873      | 0.002322       | 0.000933  | 8.572746  | 0.000160 | 0.136353  | 0.007276 | 0.306280 | 29.374 |
| 0.080845 | 0.144565  | 0.255439       | 0.182161       | 0.739927  | 0.138337 | 0.010873      | 0.002322       | 0.000933  | 8.572746  | 0.000157 | 0.136353  | 0.007276 | 0.306280 | 29.271 |
| 0.080845 | 0.144565  | 0.255439       | 0.182161       | 0.739927  | 0.138337 | 0.010873      | 0.002322       | 0.000933  | 8.572746  | 0.000154 | 0.136353  | 0.007276 | 0.306280 | 29.165 |
| 0.080845 | 0.144565  | 0.255439       | 0.182161       | 0.739927  | 0.138337 | 0.010873      | 0.002322       | 0.000933  | 8.572746  | 0.000154 | 0.136353  | 0.007276 | 0.306280 | 29.165 |
| 0.080845 | 0.144565  | 0.255439       | 0.182161       | 0.739927  | 0.138337 | 0.010873      | 0.002322       | 0.000933  | 8.572746  | 0.000154 | 0.136353  | 0.007276 | 0.306280 | 29.165 |
| 0.080845 | 0.144565  | 0.255439       | 0.182161       | 0.739927  | 0.138337 | 0.010873      | 0.002322       | 0.000933  | 8.572746  | 0.000154 | 0.136353  | 0.007276 | 0.306280 | 29.165 |

Notes: We observe that multiple parameter sets share similar fitness results. The bottom row gives the parameters output by MATLAB’s lsqcurvefit.
individuals. We then apply the GA process described in Step 2 $K$ times to this same initial set of $M$ parents in the GA process.

5. Results

In Table 2, we share results using $M = 500$ and $K = 20$ for parameter choices that describe data from Nord Department in Haiti. We observe that despite stochasticity in every step, many of the top-performing resulting parameter sets are quite similar, and also quite different from the parameter set chosen by MATLAB’s lsqcurvefit which is included at the bottom of the table. The resulting parameter set with the least error can be chosen as the ‘best’ parameter set to allow the given model to describe the given parameter set, but as we described at the beginning of this section, when the data itself are noisy, we see value in considering the variation within similarly performing parameter sets. As a comparison, we provide graphs of the model’s ability to describe the reported incidence in Nord Department using lsqcurvefit and the top – performing GA output (Figure 2). The difference in values chosen from within the parameter landscape between the optimization methods is intriguing, and highlights the limitations a modeller might find in choosing a single best parameter set from a single optimization method in describing the data.

We further note that neither of the simulations resulting from the chosen parameter sets in Figure 2, and indeed none of the simulations resulting from our other ‘best-fit’ parameter sets, seems to fit the end data well. This is not uncommon in cholera modelling; in fact, we were pleased to see the second ‘hump’ in our simulations. The models that are able to simulate the data’s repeated epidemic waves include some sort of environmental forcing term that mimics what we believe to be the cause of the repeating epidemic waves. While efforts at tying cholera outbreaks to rain or flooding have been attempted in both specific and general circumstances [9,22,24,25,38], and recently have been attempted to explain the data in Haiti [6], unless the models are simulating historic data they can only give a stochastic or even periodic ‘guess’ at what future epidemics might experience. It may be an interesting question for a future study to consider how one might choose parameters if stochastic forcing is allowed, with attention to the progress made in [24] but from the point of view of models similar to these.

In [39] and in future work, we use the variation in parameters to fully assess possible intervention outcomes. We also believe that the variation in parameter sets can give us an additional tool.

Figure 2. Fitness performance with two parameter estimation methods.
to consider model sensitivities. It is important to note that our approach has an advantage over most traditional methods, in that it does not assume the existence of a unique solution. Especially as models become more complicated and contain a large number of parameters, seeking a unique solution may not be appropriate. In such cases, we believe there is danger in considering a single ‘best’ parameter set as provided by lsqcurvefit or any technique, when this single set is used to draw a unique set of informative conclusions, when instead the consideration of the model output using multiple parameter solutions may be much more informative. Hence, we believe our approach, and others, for which multiple parameter solutions may be discovered, provide an important change in the common mindset of seeking a single solution. The randomness built in the GA approach helps in producing new information with comparable fits; and if a unique solution exists, then the GA will return very similar, if not identical solutions.

6. Future work

While the immediate goal of this project is to consider GAs as a tool for parameter selection, the results, such as those given in Table 2 have fed a number of new projects. As presented in [39], we are currently working to consider the impact of parameter uncertainty on optimal control advice for the best mix of sanitation and vaccination strategies to effectively contain an epidemic. Separately, we are considering a comparison between GA output and sensitivity analysis results, and these tools can be used to paint a more complete picture about model sensitivities. In addition, this GA approach can be easily extended to consider the question of model selection. For cholera, model selection was considered by Rinaldo et al. [37], and we will revisit the issue from a different point of view.

7. Conclusion

As the foundation of evolutionary computing, the basic principles of GAs differ radically from those of most of the traditional optimization methods. GAs work with a population of points instead of a single point, which many traditional optimization methods use, and they use previously obtained information in a biologically evolving manner. Furthermore, the GA approach alleviates a main flaw of using some traditional methods, such as maximum likelihood (ML) or least squares (LS), in which parameters are estimated directly from the data that has built-in randomness. The models of our type are distinctly different in that we do not estimate the parameters directly from the data. In our case, the chosen parameters define a model, whose output is compared to the data in order to evaluate the fitness of the chosen parameters. In a way, the randomness in the data does not directly affect our parameter selection, as is the case in ML or LS approaches. The deterministic model we implement is, in fact, a layer between the parameter choice and the data. In contrast, with the ML approach the likelihood of a model to produce the selected sample is maximized by optimizing the model parameters. Similarly in the LS approach, a function of the error is minimized by optimizing the model parameters. In both cases, the resulting parameter estimates vary based on the randomness built in the sampling process. However, in our case, we are in fact optimizing the fitness of a deterministic model whose validity is connected to the accuracy of the selected parameters and the resulting (ODE) system solution. In the application of parameter selection for a model, we particularly appreciate the method’s output of multiple parameter sets that provide a similar fit to the data, and we hope that researchers would consider more than one ‘best’-fitting parameter set when using model simulations to inform decisions.

In this and other applications, GAs use the principles of natural selection so that previously obtained good solutions contribute to the global solution using a reproduction operator and
are propagated adaptively through crossover and mutation operators. Another advantage with a population-based search algorithm is that multiple optimal solutions can be captured in the population, which increases the efficiency in converging to the global solution. Hence, since there is more than one string of parameter values being processed simultaneously, it is very likely that the expectation of the GA solution may be a global solution when such a solutions exists. In a way, by their nature, GA operators exploit the similarities in string-structures to make an effective search. Because GAs use a population of string-coded variables, GAs eliminate complications arising from working with discrete or even discontinuous functions. If desired, the GAs can require only function values at various discrete points, which make GAs useful for a large variety of models.

Acknowledgments

The model for cholera was developed in prior work with Holly Gaff, Renee Fister, and Suzanne Lenhart. In addition, Schaefer is grateful for the support of her undergraduate student Marina Romadan who helped to evaluate GA outputs.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Science Foundation grant DMS-0813563.

References

[1] J.R. Andrews and S. Basu, Transmission dynamics and control of cholera in haiti: An epidemic model, The Lancet 377 (2011), pp. 1248–1255.
[2] D.D. Anh, A.L. Lopez, V.D. Thiem, S.L. Grahek, T.N. Duong, J.K. Park, H.J. Kwon, M. Favorov, N.T. Hien, and J.D. Clemens, Use of oral cholera vaccines in an outbreak in vietnam: A case control study, PLoS Neglected Trop. Dis. 5 (2011), e1006 (pp. 1–7).
[3] C. Codeco, Endemic and epidemic dynamics of cholera: The role of the aquatic reservoir, BMC Infectious Dis. 1 (2001), pp. 1–14.
[4] K. De Jong, Genetic algorithms: A 30 year perspective, Persp. Adaptation Nat. Artif. Syst. 11 (2005), pp. 125–134.
[5] L.K. Durham, I.M. Longini, M.E. Halloran, J.D. Clemens, N. Azhar, and M. Rao, Estimation of vaccine efficacy in the presence of waning: Application to cholera vaccines, Amer. J. Epidemiol. 147 (1998), pp. 948–959.
[6] M.C. Eisenberg, G. Kujbida, A.R. Tuite, D.N. Fisman, and J.H. Tien, Examining rainfall and cholera dynamics in haiti using statistical and dynamic modeling approaches, Epidemics 5 (2013), pp. 197–207.
[7] M.C. Eisenberg, S.L. Robertson, and J.H. Tien, Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease, J. Theoret. Biol. 324 (2013), pp. 84–102.
[8] K.R. Fister, H. Gaff, E. Numfor, E. Schaefer, and S. Lenhart, Optimal control of an age-based model for cholera, (2014), To be submitted.
[9] K. Fister, H. Gaff, E. Schaefer, G. Buford, and B.C. Norris, Investigating cholera using an sir model with age-class structure and optimal control, Involve (2015), to appear.
[10] A. George, B. Rajakumar, and D. Binu, Genetic algorithm based airlines booking terminal open/close decision system, Proceedings of the International Conference on Advances in Computing, Communications and Informatics, Chennai, India, 2012, pp. 174–179.
[11] R.I. Glass, S. Becker, M.I. Huq, B.J. Stoll, M.U. Khan, M.H. Merson, J.V. Lee, and R.E. Black, Endemic cholera in rural Bangladesh, 1966–1980, Amer. J. Epidemiol. 116 (1982), pp. 959–970.
[12] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Vol. 412, Addison-wesley, Reading Menlo Park, 1989.
[13] Y.H. Grad, J.C. Miller, and M. Lipsitch, Cholera modeling: Challenges to quantitative analysis and predicting the impact of interventions, Epidemiology (Cambridge, Mass.) 23 (2012), pp. 523–530.
[14] J.W. Hallam, O. Akman, and F. Akman, Genetic algorithms with shrinking population size, Comput. Stat. 25 (2010), pp. 691–705.
[15] D.M. Hartley, J.G. Morris Jr, and D.L. Smith, Hyperinfectivity: A critical element in the ability of v. cholerae to cause epidemics?, PLoS Med 3 (2005), e7 (pp. 0063–0069).
