MODELING CHOLERA AS A STOCHASTIC PROCESS

MURAT O. AHMED

Abstract. We present a discrete-state stochastic model for Cholera to explain the behavior of an environmental reservoir. The resulting birth-death-immigration model does a reasonable job explaining the expected number of infecteds but does poorly in explaining the variance of the process. A continuous state limit is derived to provide an SDE approximation to the process. Using data from Dhaka we fit the parameters of our models and discuss the implications.

1. Introduction

The dynamics of disease rates depend on number of people who can contract and spread the disease. This dependence leads to periodic behavior in the number of infected individuals. The periodicity implies that certain times are characterized by very high and low rates. Additionally, certain diseases can be contracted through an environmental reservoir, requiring no infected individuals initially to spread the disease. When disease rates are very low the environmental reservoir is determining how the disease is spread, since the spreading of disease from person to person is minimal. To explain the effect of an environmental reservoir we identify those times when disease rates are very low and study how the disease behaves shortly after.

Cholera can be contracted person to person as well as from an environmental reservoir. From figure 1 we see highly periodic behavior in disease rates. When disease rates are very high the susceptible population is low and this slows the spread of disease so that rates decrease. Likewise, when rates are low the susceptible population is high which allows the disease to spread. From the ACF, figure 2 left, we see a strong correlation at periods of one and two years, this plot suggests there is a strong annual cycle with cholera. Additionally, the periodogram (figure 2 right) has a strong peak at one year cycles, as well as smaller peaks for six-month and quarterly periods. From this we can conclude that the behavior of Cholera is periodic annually. We should then analyze the behavior of this disease within a year time-frame.

We see from figure 3 that the yearly behavior of this disease is consistent across time. The disease seems to peak around April then steadily decline to a minimum around August, from which it begins to rise again. This pattern clearly repeats itself every year. Around August we would expect the effect from the environmental reservoir to be the strongest. We want to model the rise in disease rates after the disease is at its minimum. Cholera rates seem to rise up until November then start to decrease around December. However, the rate of increase slows down after November suggesting that between August and November is when the environmental

Date: July 8, 2005.

The author thanks Ed Ionides and Nikola Petrov for all their advice and help.

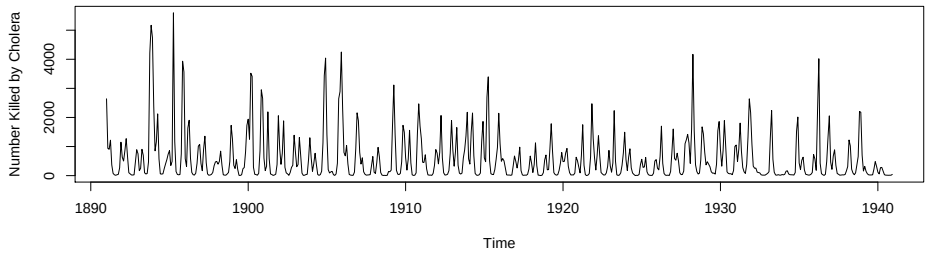

FIGURE 1. Time-Series of Cholera Mortality

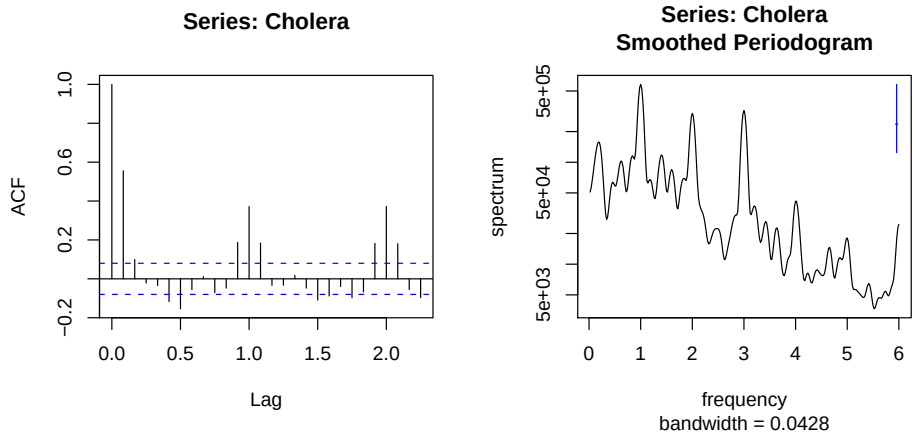

FIGURE 2. ACF and Power Spectrum of Cholera Infections

reservoir effect is greatest and the crowding-out effect is negligible. Our goal is to offer a stochastic model for the behavior of Cholera between August and November.

Modeling using a counting process allows us to estimate the effects of person-to-person infection and infection from the reservoir. The data are monthly Cholera mortality rates in Dhaka, Bangladesh. Our estimates of the parameters get close to the moments of the data, but are not able to fit them exactly. The variance of the data becomes too large as the number of infecteds grow for our model to fit the data. Consequently, our model can explain the expected value of the number of infecteds fairly well, but fails to accurately describe the variance of year to year infection rates.

2. Modeling with Count Processes

2.1. **Discrete State Continuous Time Modeling.** To model disease rates when there is a reservoir present we shall use a birth-death-immigration process. Let $X(t)$ be the number of infected individuals. Since the disease is contagious, anyone with

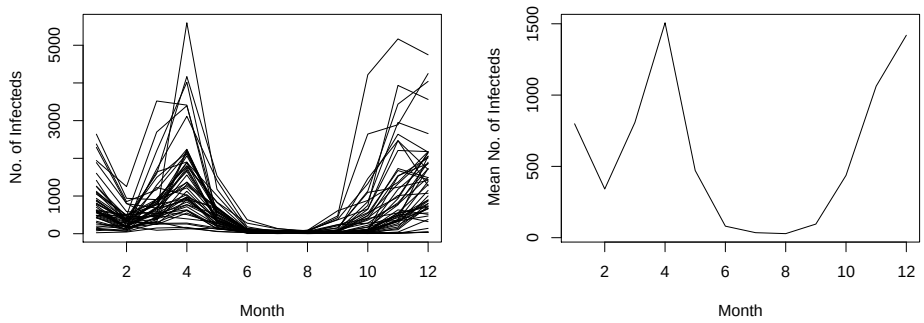

FIGURE 3. Superposition of Yearly Infections and Mean Infections

the disease can spread it and therefore creates a type of branching process. Diseases contracted from other individuals will be part of the birth process. Obviously removal from the population of infected individuals is the death process. Additionally, since there is a reservoir from which healthy people can contract cholera we add an immigration term for those who enter into the population of infected individuals, but have not contracted the disease from someone else.

For the immigration rate we will assume a Poisson process with constant rate ν. First we must find the generating function for the birth-death process. We begin by modeling the short-time transition probabilities. We allow the birth and death rates to be nonlinear, with respect to the population size. Let λ_n be the birth rate when there are n individuals with the disease and μ_n be the corresponding death rate. We define the rates as, $\lambda_n = n^\alpha \lambda + \nu$ and $\mu_n = n^\beta \mu$ where $\alpha, \beta \in \mathbb{R}$. The short-time transition probabilities are then,

(2.1.1)

$$\mathbb{P}\left(X(t+h) = j \mid X(t) = i\right) = \begin{cases} (i^\alpha \lambda + \nu)h + o(h), & j = i+1 \\ 1 - (\lambda i^\alpha + \nu + \mu i^\beta)h + o(h), & j = i \\ i^\beta \mu h + o(h), & j = i-1 \\ o(h), & |j| > i+1. \end{cases}$$

From this we are able to derive a system of forward equations. Let $p_j(t) := \mathbb{P}\left(X(t) = j\right)$ and $p_{ij}(h) := \mathbb{P}\left(X(t+h) = j \mid X(t) = i\right)$. It follows from the Chapman-Kolmogorov equations that,

$$\begin{aligned} \mathbb{P}\left(X(t+h) = j\right) &= \sum_{i \in \mathcal{S}} \mathbb{P}\left(X(t) = i\right) \mathbb{P}\left(X(t+h) = j \mid X(t) = i\right) \\ &= \mathbb{P}\left(X(t+h) = j \mid X(t) = j-1\right) p_{j-1}(t) \\ &+ \mathbb{P}\left(X(t+h) = j \mid X(t) = j\right) p_j(t) \\ &+ \mathbb{P}\left(X(t+h) = j \mid X(t) = j+1\right) p_{j+1}(t) + o(h) \\ &= (((j-1)^\alpha \lambda + \nu)h) p_{j-1}(t) + (1 - (\lambda j^\alpha + \nu + \mu j^\beta)h) p_j(t) \\ &+ \mu (j+1)^\beta h p_{j+1}(t) + o(h). \end{aligned}$$

We subtract $\mathbb{P}(X(t)=j)$ and divide by h, then take the limit as $h \downarrow 0$ while noting that $\lim_{h\downarrow 0}\frac{o(h)}{h}=0$. We are then left with,

$$\lim_{h\downarrow 0}\frac{p_j(t+h)-p_j(t)}{h} = ((j-1)^{\alpha}\lambda+\nu)p_{j-1}(t)-(\lambda j^{\alpha}+\nu+\mu j^{\beta})p_j(t) + \mu(j+1)^{\beta}p_{j+1}(t).$$

Therefore we have that,

$$p_j'(t) = (\lambda(j-1)^{\alpha}+\nu)p_{j-1}(t)-(\lambda j^{\alpha}+\nu+\mu j^{\beta})p_j(t)+\mu(j+1)^{\beta}p_{j+1}(t). \tag{2.1.2}$$

Take the special case where $\alpha=\beta=1$, if we multiply by s^j then sum we have that,

$$\sum_{j=0}^{\infty}s^j p_j'(t) = \lambda s^2\sum_{j=1}^{\infty}(j-1)s^{j-2}p_{j-1}(t)+\nu s\sum_{j=1}^{\infty}s^{j-1}p_{j-1}(t)-(\lambda+\mu)s \cdot\sum_{j=0}^{\infty}js^{j-1}p_j(t)-\nu\sum_{j=0}^{\infty}s^j p_j(t)+\mu\sum_{j=0}^{\infty}(j+1)s^j p_{j+1}(t)$$

Since the probability generating function of the process is $G(s,t)=\sum_{j=0}^{\infty}s^j p_j(t)$ we are left with a first order partial differential equation.

$$\frac{\partial G}{\partial t}=(s-1)\left[\nu G+(\lambda s-\mu)\frac{\partial G}{\partial s}\right] \tag{2.1.3}$$

Subject to the boundary condition $G(s,0)=s^I$, given $X(0)=I, \because p_j(0)=\delta_{ij}$. We solve this equation by the method of characteristics. First we rewrite it in the form $\tau(s,t)\frac{\partial G}{\partial t}+\sigma(s,t)\frac{\partial G}{\partial s}=\rho(s,t,G)$. Then we solve the equations $\frac{dt}{\tau(s,t)}=\frac{ds}{\sigma(s,t)}=\frac{dG}{\rho(s,t,G)}$, noting that the constants that occur after solving can be expressed as functions. That is $C_1=\Phi_1(s,t,G)$ and $C_2=\Phi_2(s,t,G)$, then the general solution of the PDE has the form $\Phi_1(s,t,G)=f(\Phi_2(s,t,G))$, where f is an arbitrary function of one variable. We have that $\frac{\partial G}{\partial t}+(s-1)(\mu-\lambda s)\frac{\partial G}{\partial s}=\nu(s-1)G$ so that we need to solve $dt=\frac{ds}{(s-1)(\mu-\lambda s)}$ and $\frac{ds}{\mu-\lambda s}=\frac{dG}{\nu G}$. This yields that

$$C_1=\frac{\ln G}{\nu}-\frac{\ln(\lambda s-\mu)}{\lambda}\text{ and }C_2=t+\frac{1}{\lambda-\mu}\ln\left(\frac{1-s}{\mu-\lambda s}\right).$$

Taking exponents of both expressions we have that

$$G(\lambda s-\mu)^{-\frac{\nu}{\lambda}}=f\left(e^t(1-s)^{\frac{1}{\lambda-\mu}}(\mu-\lambda s)^{\frac{1}{\mu-\lambda}}\right).$$

Let $z=z(s)$ be the argument inside f, then by using the boundary condition we can solve for $f(z)$.

$$G(s,0)=f\underbrace{\left((1-s)^{\frac{1}{\lambda-\mu}}(\mu-\lambda s)^{\frac{1}{\mu-\lambda}}\right)}_{z(s)}=s^I(\lambda s-\mu)^{\frac{\nu}{\lambda}}$$

$$\Rightarrow f(z)=\left(\frac{\mu z^{\lambda-\mu}-1}{\lambda z^{\lambda-\mu}-1}\right)^I\left(\frac{\mu z^{\lambda-\mu}-1}{\lambda z^{\lambda-\mu}-1}\cdot\lambda-\mu\right)^{\frac{\nu}{\lambda}}$$

So that, $G(s,t) = \left(\frac{\mu-\lambda}{\lambda z^{\lambda-\mu}-1}\right)^{\frac{\nu}{\lambda}} \left(\frac{\mu z^{\lambda-\mu}-1}{\lambda z^{\lambda-\mu}-1}\right)^{I}$, where $z = e^t(1-s)^{\frac{1}{\lambda-\mu}}(\mu-\lambda s)^{\frac{1}{\mu-\lambda}}$. Plugging in for z we finally obtain,

$$G(s,t) = \left(\frac{\mu-\lambda}{e^{t(\lambda-\mu)}(s-1)\lambda+\mu-s\lambda}\right)^{\frac{\nu}{\lambda}} \left(\frac{e^{t(\lambda-\mu)}(s-1)\mu+\mu-s\lambda}{e^{t(\lambda-\mu)}(s-1)\lambda+\mu-s\lambda}\right)^{I} \tag{2.1.4}$$

We see that $G(1,t) = 1, \forall t$, so $X(t)$ is honest $\forall \lambda, \mu$, and ν. From the moment generating function we can find the mean and variance of this process. We can find the first moment since $\mathbb{E}(X(t)) = \frac{\partial G}{\partial s}\big|_{s=1}$. Also, the variance is $\mathbb{V}(X(t)) = G_{ss}(1,t) + G_s(1,t) - (G_s(1,t))^2$. So that,

$$\mathbb{E}(X(t)) = Ie^{t(\lambda-\mu)} + \nu\frac{(e^{t(\lambda-\mu)}-1)}{\lambda-\mu} \tag{2.1.5}$$

Or equivalently,

$$\mathbb{E}(X(t)) = Ie^{(\lambda-\mu)t} + \nu\int_0^t e^{(\lambda-\mu)s}ds. \tag{2.1.6}$$

We see that the expected value is linearly proportional to the immigration term times the integral of the expected value of a birth-death process. This is because each time an immigration occurs it starts a new birth death process from the time it occurs to the final time t. Additionally, the variance of the process at time t is

$$\mathbb{V}(X(t)) = I\frac{\lambda+\mu}{\lambda-\mu}e^{t(\lambda-\mu)}(e^{t(\lambda-\mu)}-1) + \nu\frac{(\lambda e^{t(\lambda-\mu)}-\mu)(e^{t(\lambda-\mu)}-1)}{(\lambda-\mu)^2}. \tag{2.1.7}$$

As we would expect, if we take the limit as $\nu \downarrow 0$ we are left with the expected value and variance of a simple birth-death process found in Grimmett and Stirzaker [3].

$$\lim_{\nu\downarrow 0}\mathbb{E}(X(t)) = Ie^{t(\lambda-\mu)}, \qquad \lim_{\nu\downarrow 0}\mathbb{V}(X(t)) = I\frac{\lambda+\mu}{\lambda-\mu}e^{t(\lambda-\mu)}(e^{t(\lambda-\mu)}-1)$$

Intuitively we expect that as time goes to infinity if the birth rate is larger than the death rate this process would explode and if the death rate is larger, the immigration rate should determine the expected value of the process.

$$\lim_{t\to\infty}\mathbb{E}(X(t)) = \begin{cases}\infty, & \text{if } \lambda > \mu \\ \frac{\nu}{\mu-\lambda}, & \text{if } \lambda < \mu\end{cases} \tag{2.1.8}$$

The expected value behaves as expected. When $\lambda < \mu$ we have that the expected value is a ratio between the immigration rate and the difference between the death and birth rate, say the net death rate. It has been shown that the asymptotic distribution of death-immigration process is $Poisson\left(\frac{\nu}{\mu}\right)$, [3]. So that the expected value would be $\frac{\nu}{\mu}$, which is essentially what we have.

Also, as the birth rate approaches the death rate we expect births to cancel deaths and only our immigration term should affect the expected value. Indeed, $\lim_{\lambda\to\mu}\mathbb{E}(X(t)) = I + \nu t$, which is the initial population size plus the expected value of a *Poisson* process with rate ν.

Obviously, the asymptotic probability of extinction is zero, since there is constant immigration. However, we can still say something about the process eventually

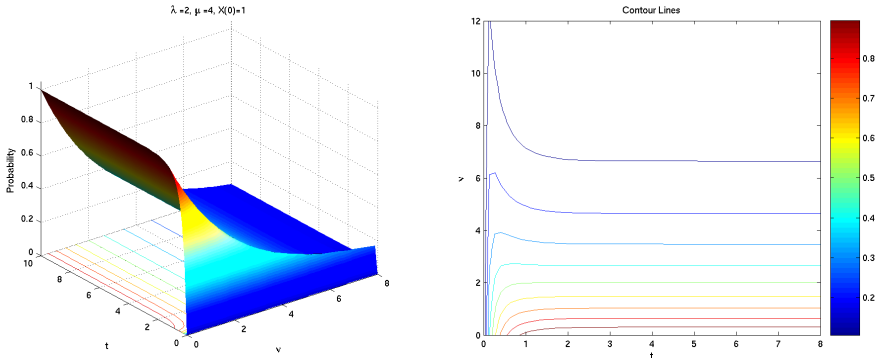

FIGURE 4. Probability $X(t) = 0$ when $\lambda < \mu$

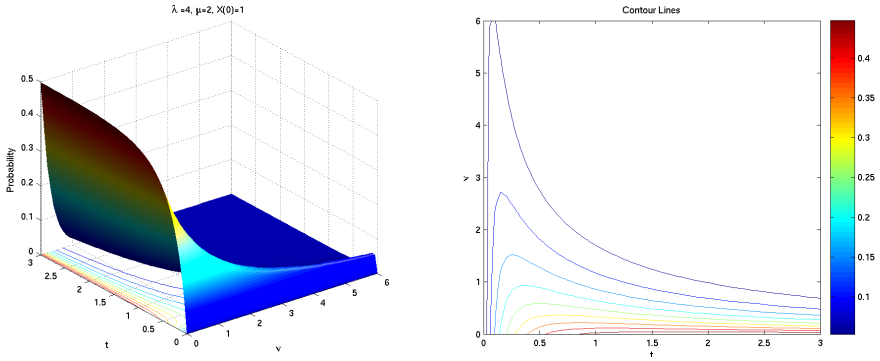

FIGURE 5. Probability $X(t) = 0$ when $\lambda > \mu$

going to 0. We have that $\mathbb{P}(X(t) = 0) = G(0,t) = \left(\frac{\mu-\lambda}{\mu-\lambda e^{t(\lambda-\mu)}}\right)^{\frac{\nu}{\lambda}} \left(\frac{\mu e^{t(\lambda-\mu)}-\mu}{\lambda e^{t(\lambda-\mu)}-\mu}\right)^{I}$, so

$$\lim_{t\to\infty} \mathbb{P}(X(t) = 0) = \begin{cases} \left(\frac{\mu-\lambda}{\mu}\right)^{\frac{\nu}{\lambda}}, & \text{if } \lambda < \mu \\ 0, & \text{if } \lambda > \mu. \end{cases}$$

Surprisingly, $\lim_{t\to\infty} \mathbb{P}(X(t) = 0) < 1, \forall \nu > 0$ in either case. For the simple birth death process considered in [3] the asymptotic probability of a zero population is 1, if $\mu > \lambda$, and less than one but positive, if $\mu < \lambda$. This is particularly important since diseases with environmental reservoirs do not go extinct even if all infecteds have been removed and in our model sometimes will never even reach a zero population.

2.2. **Numerical Methods for Finding Transition Probabilities.** One method for solving the forward equations is by Laplace transforms. We know that $\mathcal{L}[f'](\theta) =$

$\theta\mathcal{L}[f](\theta) - f(0)$. We assume that $p_j(0) = \delta_{ij}$, given that $X(0) = i$. So the transformed forward equations are,

$$\widehat{p}_j(\theta) = \frac{\delta_{ij} + (\lambda(j-1)^\alpha + \nu)\widehat{p}_{j-1}(\theta) + \mu(j+1)^\beta \widehat{p}_{j+1}(\theta)}{(\theta + \lambda j^\alpha + \nu + \mu j^\beta)} \tag{2.2.1}$$

One can then solve these equations by numerically by inverting the Laplace transform. This would then give the transition probabilities when the birth and death rates are non-linear with respect to the population size.

2.3. **Simulating From the Process.** Given a stochastic semigroup $\{\mathbf{P}_t : t \geq 0\}$, with entries $p_{ij}(s,t)$ for some $s \leq t$, we can construct the generator of the Markov chain. Since $p_{ij}(h)$ is approximately linear when h is small there exist constants $\{g_{ij} : i, j \in \mathcal{S}\}$, where $\mathcal{S} = \mathbb{N} \cup \{0\}$ is the state space, such that $p_{ij}(h) \backsimeq g_{ij}h$, if $i \neq j$ and $p_{ii}(h) \backsimeq 1 + g_{ii}h$. The generator matrix, $\mathbf{G} = (g_{ij})$, gives us the probability of jumping or staying put for a small time interval h. We have that either nothing happens during $(t, t+h)$ with probability $1+g_{ii}h+o(h)$, or the chain jumps to state $j \neq i$ with probability $g_{ij}h + o(h)$.

From our construction we have that $g_{ii} = -(i(\lambda+\mu)+\nu), g_{i,i+1} = i\lambda+\nu, g_{i,i-1} = i\mu$ for $i \neq 0$, and $g_{00} = -\nu, g_{01} = \nu$. For any other jump $g_{ij} = 0$, thus

$$\mathbf{G} = \begin{pmatrix} -\nu & \nu & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu+\nu) & \lambda+\nu & 0 & 0 & \cdots \\ 0 & 2\mu & -(2(\lambda+\mu)+\nu) & 2\lambda+\nu & 0 & \cdots \\ 0 & 0 & 3\mu & -(3(\lambda+\mu)+\nu) & 3\lambda+\nu & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Let $X(s) = i$, then define $U = \inf\{t \geq 0 : X(t+s) \neq i\}$ to be the time it takes for the chain to jump. It has been shown that U is exponentially distributed with parameter $-g_{ii}$, [3]. There are only two possible places the chain can jump if it is at state i, either to $i+1$ or $i-1$. This implies that the probability of jumping to state $j \neq i$, given that a jump occurs, is $-\frac{g_{ij}}{g_{ii}}$. Therefore, this distribution is *Bernoulli* with parameter $-\frac{g_{i,i+1}}{g_{ii}}$. If there is a success we jump up one, otherwise we jump down one. From this we can easily simulate the process as long as we set the initial conditions and parameter values.

We see that in figure 6 the growth is exponential. We have that as the immigration parameter increases the growth becomes more linear than exponential. This is as expected since the immigration parameter is always constant so if ν dominates the process the growth should be linear. Conversely, if the birth rate is dominant then the growth should be exponential since $\lambda_n = n\lambda$ increases as the population size increases.

When the death rate is larger than the birth rate we know from (2.1.8) that the immigration term will determine the key features of the process. We see that the process quickly goes to the steady-state value, then behaves like a stationary process. Also, we have that the variance increases as ν increases. Further evidence of the importance of the immigration term is given by the time it takes for the jumps to occur. When $\lambda > \mu$ we see that regardless of ν it takes around 2 time units for there to be 1000 jumps. However, when $\lambda < \mu$ we find that it takes about 180, 47, 23 and 14 time units for 1000 jumps as we increase the immigration parameter. Based on the characteristics of the simulations and from figure 3, during

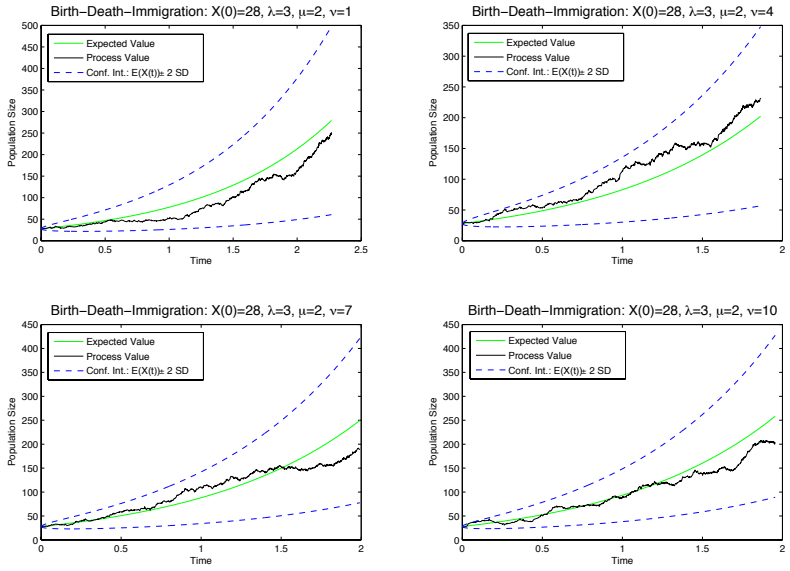

FIGURE 6. Simulation Realizations with 1000 Jumps, $\lambda > \mu$

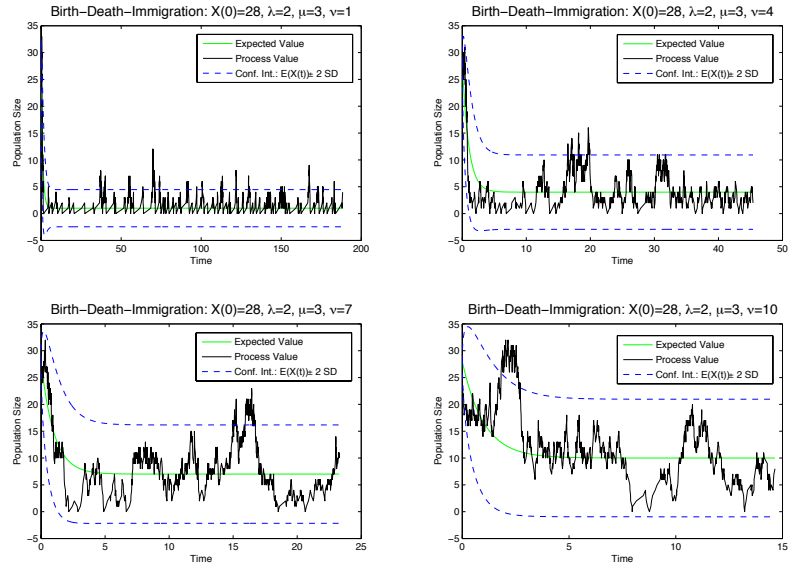

FIGURE 7. Simulation Realizations with 1000 Jumps, $\lambda < \mu$

the months of August to October, we would expect cholera to behave as a process whose birth rate is larger than the death rate.

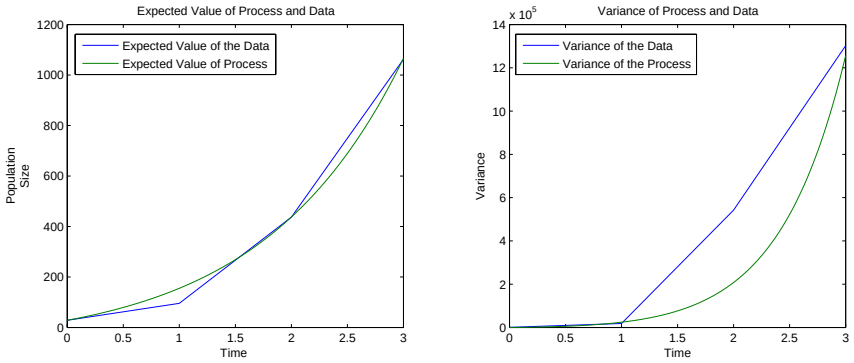

FIGURE 8. Expected Value and Variance of Process and Data, $I = 28.48, \lambda = 75.8, \mu = 75, \nu = 60$

2.4. **Estimating Parameters: Method of Moments.** One method for estimating the parameters of our model is by using the expected value and variance of our data and of the process to set up a system of equations. These equations can then be solved to determine the parameter values. Since there are three parameters to estimate we need three moments from the data. Our process is intended to model the data from August ($t = 0$) to around November ($t = 3$), so that using the mean and variance of the data we can get up to 6 moments. By inspecting Figure 3, we see that the data is very volatile after August so we use only the first moments of the months September to November.

Results from the MOM estimation indicate that our model has problems explaining the variance of the process. The numerical root finding is very sensitive to initial conditions and encounters many singular points during the search. The main problem is that the variance of this process gets large too quickly for it to be explained by our model, which has exponential growth in the variance. Even when we use only first moments to estimate the parameters we encounter similar problems. In fact the only time we find a stable solution is when we make a change of variable of $a = \lambda - \mu$ and use the means of October and November. This gives us the estimate for the difference between the birth and death rate as well as the immigration rate.

We estimate that $a = \lambda - \mu \backsimeq 0.8$ and $\nu \backsimeq 60$. Of course we still do not know that exact values of λ and μ separately, although we do have a linear equation that they satisfy. We choose λ and μ so that the processes variance is close to the variance of the data. In figure 2.4 we see that the choice $\lambda \backsimeq 75.8$ and $\mu \backsimeq 75$ allows us to accurately estimate the variance during September and November, but is far from accurate for the month of October. The problem is that the large increase in the variance from September to October cannot be explained by the exponential growth given by our model. This increase is just too drastic.

Once we have estimates for $a = \lambda - \mu$ and ν we would like to get a confidence interval. We use the bootstrap to get an estimate of the variability of our parameters. Since our model does not do a good job explaining the variance of the data, essentially being too small, we use a non-parametric bootstrap. Since we have 50 realizations with our data (August to November), we first randomly draw

with replacement from our data 50 times. Then we use the resulting expected values to estimate the parameters with MOM. The variance across our estimates then becomes the basis for the construction of our confidence interval. One way to construct a confidence interval is to use the expected value found in the original MOM estimate plus or minus two standard deviations from the bootstrap. This would give us an approximate 95% confidence interval for our parameter estimates. However, we can also use the 97.5% and 2.5% quantiles of the fitted estimates. Our confidence interval would not be symmetric but it would give a better idea of the true range of possible values, since if the variance were very high for an estimate a symmetric confidence interval may include negative values.

	$\lambda - \mu$	ν
Estimate	0.7977	60.3917
97.5% Quantile	1.1141	148.8714
2.5% Quantile	0.5025	7.6462

TABLE 1. Estimates and Confidence Intervals for Parameters

We see that the confidence interval for the difference between the birth and death rate contains all positive values. The standard deviation is about 0.1598 which does not seem all that large. However, for the immigration term we have a standard deviation of about 38.2580, which would actually go below zero for a 95% confidence interval lower bound. This suggests a high degree of variability in our estimation for the rate of infection from the reservoir. We cannot have too much confidence in our estimate nor even state whether the reservoir plays a very strong role in spreading the disease or just a small one. If ν were around 7 then only around 7 individuals would be expected to contract the disease from the reservoir. This is insignificant as the number of infected individuals from the data is over 400 in October and over 1000 in November. If however the rate were closer to 140 or so then the reservoir is extremely important in determining the initial spread of Cholera.

3. Diffusion Approximation to the Counting Process

3.1. **SDE from Transition Probabilities.** Let $X = \{X(t) : t \geq 0\}$ be a diffusion process. Suppose there exist functions $a(t, x)$ and $b(t, x)$, such that

$$\mathbb{E}\left(X(t+h) - X(t) \mid X(t) = x\right) = a(t, x)h + o(h)$$

and

$$\mathbb{E}\left([X(t+h) - X(t)]^2 \mid X(t) = x\right) = b^2(t, x)h + o(h).$$

Then we identify $a(x, t)$ and $b^2(x, t)$ as the infinitesimal mean and variance of the diffusion process. It follows that the differential form of this is $dX_t = a(t, X_t)dt + b(t, X_t)dW_t$, where W_t is a standard Wiener process, or equivalently in its integrated form

$$X_t = X_0 + \int_0^t a(s, X_s)ds + \int_0^t b(s, X_s)dW_s$$

where the second integral is an Itô integral.

Using the short-time transition probabilities given by (2.1.1) we have that $a(x, t) = (\lambda - \mu)x + \nu$ and $b^2(x, t) = (\lambda + \mu)x + \nu$. This gives us the SDE,

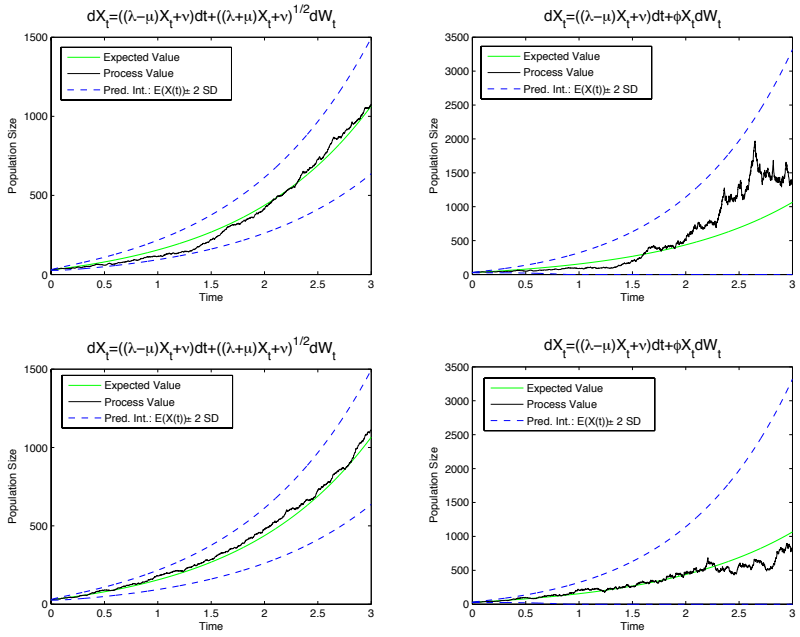

FIGURE 9. Two Sample Paths of $dX_t = ((\lambda - \mu)X_t + \nu)dt + \sqrt{((\lambda + \mu)X_t + \nu)}dW_t$ (left) and $dX_t = ((\lambda-\mu)X_t+\nu)dt+\varphi X_t dW_t$ (right); $I = 28.48, \lambda = 2.8, \mu = 2, \nu = 60, \varphi = 1/2$

$$dX_t = ((\lambda - \mu)X_t + \nu)\, dt + \sqrt{(\lambda + \mu)X_t + \nu}dW_t \tag{3.1.1}$$

Due to the nonlinear stochastic term and no easy way of transforming the variable to get a linear SDE we attempt to solve (3.1.1) numerically. We use an *Euler approximation*, with equidistant discretization times, given in [6].

From figure 9 we see that the SDE derived from the birth-death-immigration process is very similar to our BDI simulation. The expected value for the BDI process is very close to the SDE solution. However from figure 10, it is apparent that the expected value of the data is proportional to the standard deviation. This would imply that the stochastic term in the SDE should be linear in X_t rather than the square root. This would also insure a much larger variance in the process, which is one shortcoming or the BDI model. We do not just square the stochastic term in (3.1.1). This is because the $\sqrt{(\lambda + \mu)X_t + \nu}$ term has a meaning as the infinitesimal variance of the BDI process. But, $(\lambda+\mu)X_t+\nu$ has no direct meaning, we just want X_t to appear linearly with the stochastic part, dW_t. Therefore we choose the SDE,

$$dX_t = ((\lambda - \mu)X_t + \nu)dt + \varphi X_t dW_t \tag{3.1.2}$$

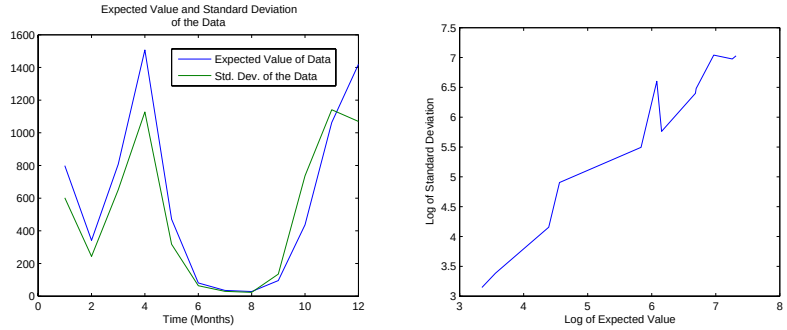

FIGURE 10. Proportionality Between the Expected Value and Variance of the Data

where φ is a constant. This SDE is linear so we have a solution (taken from [6])

$$X_t = \Phi_t \left(X_0 + \nu \int_0^t \Phi_s^{-1} ds \right) \tag{3.1.3}$$

where $\Phi_t = e^{((\lambda-\mu)-\frac{1}{2}\varphi^2)t+\varphi W_t}$.

One reason for choosing this SDE to model the data is that the variability in the BDI model does not become large enough to explain the variance in the data. This can be attributed to the "environmental stochasticity" for which the BDI model does not consider. Rather our model considers only the "population stochasticity". This suggests we model the process with additional sources of variability. From (3.1.3) we can find the expected value and variance of this process and try to re-estimate the parameters using MOM.

Let $a = \lambda - \mu$, then we have that the expected value of this process is

$$\begin{aligned}
\mathbb{E}(X_t) &= \mathbb{E}\left(\Phi_t X_0 + \nu \int_0^t \Phi_t \Phi_s^{-1} ds \right) \\
&= X_0 e^{(a-\frac{\varphi^2}{2})t} \mathbb{E}(e^{\varphi W_t}) + \nu \int_0^t e^{(a-\frac{\varphi^2}{2})(t-s)} \mathbb{E}\left(e^{\varphi(W_t - W_s)} \right) ds \\
&= X_0 e^{(a-\frac{\varphi^2}{2})t + \frac{\varphi^2}{2}t} + \nu \int_0^t e^{(a-\frac{\varphi^2}{2})(t-s) + \frac{\varphi^2}{2}(t-s)} ds \\
&= X_0 e^{at} + \nu \frac{e^{at} - 1}{a} \\
&= X_0 e^{(\lambda-\mu)t} + \nu \int_0^t e^{(\lambda-\mu)s} ds.
\end{aligned}$$

We are using the fact that φW_t and $\varphi(W_t - W_s)$ are distributed $N(0, \varphi^2 t)$ and $N(0, \varphi^2(t-s))$ respectively, so that the expectation is over lognormal r.v.'s. Notice that this is the same value as in (2.1.6). Since we did not change the deterministic part of the equation it is no surprise that we have the same expected value as in the BDI. In figure 9 we use the same parameters as in the BDI simulations. This allows us to easily compare the two processes, noting the strong similarities between the BDI process and its corresponding SDE approximation. We also see how much greater the variability is in the SDE approximation with "environmental stochasticity".

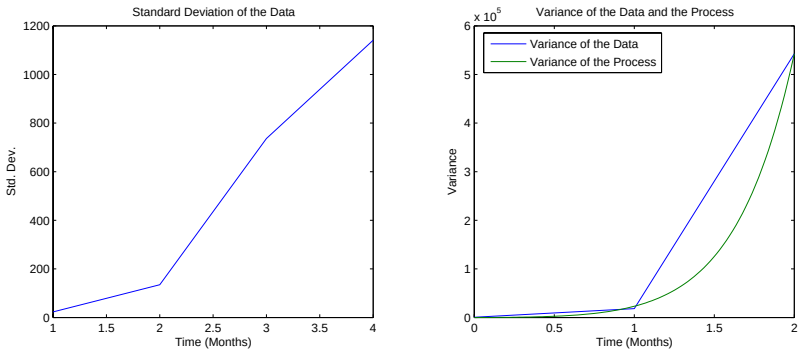

FIGURE 11. Standard Deviation Aug.-Nov. (left) and Fitted Diffusion Variance (right)

We can also find the variance of this process. First we find the second moment,

$$\begin{aligned}
\mathbb{E}(X_t^2) &= \mathbb{E}\left(X_0^2\Phi_t^2 + 2X_0\nu\int_0^t \Phi_t^2\Phi_s^{-1}ds + \nu^2\int_0^t\int_0^t \Phi_t^2\Phi_u^{-1}\Phi_v^{-1}dudv\right)\\
&= X_0^2 e^{(a-\frac{1}{2}\varphi^2)2t}\mathbb{E}(e^{2\varphi W_t}) + 2X_0\nu\int_0^t e^{(a-\frac{1}{2}\varphi^2)(2t-s)}\mathbb{E}(e^{\varphi(2W_t-W_s)})ds\\
&\quad + \nu^2\int_0^t\int_0^t e^{(a-\frac{1}{2}\varphi^2)(2t-u-v)}\mathbb{E}(e^{\varphi(2W_t-W_u-W_v)})dudv\\
&= X_0^2 e^{(a-\frac{1}{2}\varphi^2)2t+2\varphi^2} + 2X_0\nu\int_0^t e^{(a-\frac{1}{2}\varphi^2)(2t-s)+(2t-\frac{s}{2})\varphi^2}ds\\
&\quad + \nu^2\int_0^t\int_0^t e^{(a-\frac{1}{2}\varphi^2)(2t-u-v)+(2t-\frac{u+v}{2})\varphi^2}dudv\\
&= X_0^2 e^{(2a+\varphi^2)t} + 2X_0\nu\frac{e^{(a+\varphi^2)t}(e^{at}-1)}{a} + \nu^2\frac{e^{\varphi^2 t}(e^{at}-1)^2}{a^2}\\
&= e^{\varphi^2 t}\cdot\frac{(e^{at}(\nu+aX_0)-\nu)^2}{a^2}
\end{aligned}$$

which implies that,

$$\begin{aligned}
\mathbb{V}(X_t) &= \mathbb{E}(X_t^2) - (\mathbb{E}(X_t))^2\\
&= e^{\varphi^2 t}\cdot\frac{(e^{at}(\nu+aX_0)-\nu)^2}{a^2} - \left(\frac{(e^{at}(\nu+aX_0)-\nu)}{a}\right)^2\\
&= (e^{\varphi^2 t}-1)\frac{(e^{at}(\nu+aX_0)-\nu)^2}{a^2} = (e^{\varphi^2 t}-1)(\mathbb{E}(X_t))^2.
\end{aligned}$$

From this it is apparent that the standard deviation of the process is $\sigma_t = \sqrt{(e^{\varphi^2 t}-1)}\cdot\mathbb{E}(X_t)) \propto \mathbb{E}(X_t)$, which agrees with the data.

Now we can estimate φ using the variance of the data. Since the expected value of this diffusion is the same as the BDI process we can use our estimates of the a and ν parameters we obtained with the MOM. To estimate φ we use a nonlinear least-squares approach. We minimize the sum of squares difference

between the data and the variance given by the diffusion. We see in figure 11 that the standard deviation of the data does not display exponential growth from October to November, however our diffusion does. We therefore cannot include the month of November in our model since it clearly does not have properties similar to our diffusion. We only use the data for the variance in September and October and obtain an estimate of $\varphi = 0.8197$. We have very closely fitted the variance of the data for the months of September and October, something we were unable to do with the BDI variance. This is most likely due to the fact that this models variance is proportional to its expected value, whereas in the diffusion approximation the variance was proportional to the expected value.

3.2. **Constructing a Limiting Process.** Our goal is to construct an infinitesimal mean and variance from our birth-death-immigration process, so that we can find the forward equation of the diffusion process and the resulting SDE approximation. We first attempt to find the characteristic function of the process when we let the jumps sizes get very small but the amount of jumps grow very large. We look at $\mathbb{P}(X(t+h) = \xi j \mid X(t) = \xi i)$, for some $\xi > 0$ and initial population $X(0) = I$. If we let $\xi \downarrow 0$ the jumps get very small, and since $\lambda_n = n\lambda$ and $\mu_n = n\mu$, letting $X(0) \to \infty$ will create a large number of jumps. However the immigration term is just a constant, unless we make ν large the immigration aspect of this process will vanish, so we also let $\nu \to \infty$. We select our limits so that $I\xi \to \tau$ and $\nu\xi \to \sigma$.

The characteristic function of a stochastic process is defined as $\varphi(s,t) = \mathbb{E}(e^{isX(t)}) = \sum_{k=0}^{\infty} e^{is\xi k} p_k(t)$. We could of course reformulate our forward equation to find a PDE that this function satisfies. However, if we notice that $G(s,t) = \mathbb{E}(s^{X(t)})$ then we have that $G(e^{is\xi}, t) = \varphi(s,t)$. So that,

(3.2.1)

$$\varphi(s,t) = \left(\frac{\mu - \lambda}{e^{t(\lambda-\mu)}(e^{is\xi}-1)\lambda + \mu - e^{is\xi}\lambda}\right)^{\frac{\nu}{\lambda}} \left(\frac{\mu(1-e^{is\xi}) - (\mu - \lambda e^{is\xi})e^{-t(\lambda-\mu)}}{\lambda(1-e^{is\xi}) - (\mu - \lambda e^{is\xi})e^{-t(\lambda-\mu)}}\right)^{I}$$

We now take the appropriate limits and are left with,

$$\lim_{I \to \frac{\tau}{\xi}} \lim_{\nu \to \frac{\sigma}{\xi}} \lim_{\xi \downarrow 0} \varphi(s,t) = e^{is\left(\sigma \frac{e^{t(\lambda-\mu)}-1}{\lambda-\mu} + \tau e^{t(\lambda-\mu)}\right)}. \tag{3.2.2}$$

We see that by letting the initial population and immigration rate grow large we eliminate the population dynamics in the system which removes all the variance in the process and leaves us with a deterministic solution. The "population stocahsticity" has been completely removed so we have no variability left in our process. In fact we have the characteristic function of a normal random variable with expected value basically identical to the BDI process and zero variance. We are unable to get a limiting solution that is interesting, since all the variance is removed from the process. This prevents us from deriving a limiting diffusion.

4. Conclusion

Our attempt to model the spread of Cholera during the months of August through November has failed to explain the variance of the data. Although the variance of a simple birth-death-immigration process is much larger than the expected value it grows too slow to explain the variance of the Cholera data. The behavior of the expected value and variance of our process, namely exponential

growth, seems to be appropriate however. Additionally, we find that the SDE approximation to the BDI process seems reasonable. The numerical solution of the SDE looks very similar to the simulated discrete state process.

We should note that neither the mean nor variance of the BDI process can fit all three data points (September-November). The growth in both the expected value and variance from September to October is too large and cannot be fit using our exponential model. It seems that the process from August to September and from October to November need to be modeling separately. During August to September the immigration term, reservoir effect, dominates since the infected population is so small. The growth seems better modeled linearly. However, from October to November the population is large enough for exponential growth to appropriately model the data. We could try letting the constants be functions of time. The immigration term should be very large comparatively in the beginning of the process but as the population grows the difference between the birth and death rate should dominate the process.

References

[1] Engen, S., Lande, R., Sæther, B. E., *Stochastic Population Dynamics in Ecology and Conservation*, Oxford University Press, New York, 2003.

[2] Gardiner, C. W., *Handbook of Stochastic Methods*, 3^{rd} edition, Springer, Berlin, 2004.

[3] Grimmett, G. R., Stirzaker, D. R., *Probability and Random Processes*, 3^{rd} edition, Oxford University Press, New York, 2004.

[4] Karlin, S., Taylor, H. M., *A Second Course in Stochastic Processes*, Academic Press, New York, 1981.

[5] Kiffe, T. R., Matis, J. H., *Stochastic Population Models*, Springer, New York, 2000.

[6] Kloeden, P. E., Platen, E., *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, 1999.

[7] Koelle, K., Pascual, M., *Disentangling Extrinsic from Intrinsic Factors in Disease Dynamics: A Nonlinear Time Series Approach with an Application to Cholera*, The American Naturalist, 163:901-913, 2004.

[8] Rice, J. A., *Mathematical Statistics and Data Analysis*, 2^{nd} edition, Duxbury Press, Belmont, Ca., 1995.

[9] Øksendal, B. K., *Stochastic Differential Equations*, 6^{th} edition, Springer, Berlin, 2003.

5. Source Code for MATLAB

Birth-Death-Immigration Simulation

```
figure
M=input('Enter parameters as a matrix, each row is [X(0) lambda mu nu]: ');
for k=1:4
    v=M(k,:);
    X(1)=v(1);U(1)=0;
    if X(1)~=0
        U(2)=exprnd(1/(X(1)*(v(2)+v(3))+v(4)));
        T(1)=binornd(1,(X(1)*v(2)+v(4))/(X(1)*(v(2)+v(3))+v(4)));
        if T==1
            X(2)=X(1)+1;
        else
            X(2)=X(1)-1;
        end
    else
        U(2)=exprnd(1/v(4));
        X(2)=X(1)+1;
    end
    for i=1:5000
        if X(i+1)~=0
            U(i+2)=U(i+1)+exprnd(1/(X(i)*(v(2)+v(3))+v(4)));
            T(i+1)=binornd(1,(X(i)*v(2)+v(4))/...
                (X(i)*(v(2)+v(3))+v(4)));
            if T(i+1)==1
                X(i+2)=X(i+1)+1;
            else
                X(i+2)=X(i+1)-1;
            end
        else
            U(i+2)=U(i+1)+exprnd(1/v(4));
            X(i+2)=X(i+1)+1;
        end
    end
    t=linspace(0,U(5002),1001);
    exval=v(1)*exp(t*(v(2)-v(3)))+(v(4)*(exp(t*(v(2)-v(3)))-1))./...
        (v(2)-v(3));
    vari=max(v(1)*(v(2)+v(3))/(v(2)-v(3))*exp(t*(v(2)-v(3))).*...
        (exp(t*(v(2)-v(3)))-1)+(v(4)*(v(2)*exp(t*(v(2)-v(3)))...
        -v(3)).*(exp(t*(v(2)-v(3)))-1))./(v(2)-v(3))^2,0);
    sd=sqrt(vari); subplot(2,2,k)
    plot(t,exval,'g',U,X,'k',t,exval+2*sd,'--b',t,exval-2*sd,'--b')
    xlabel('Time');ylabel('Population Size');
    title(['Birth-Death-Immigration: X(0)=',num2str(v(1)),...
        ', \lambda=',num2str(v(2)),', \mu=',num2str(v(3)),...
        ', \nu=',num2str(v(4))],'FontSize',14);
    legend('Expected Value','Process Value','Conf. Int.: E(X(t))\pm 2 SD',0);
end
```

Non-Parametric Bootstrap Estimation of Conf. Int.

```
uiopen('/Users/muratoa/Desktop/Mathematics/REU/Dhaka.xls',1)
for i = 1:49
    M(i,:) = Dhaka(12*i+1:12*(i+1));
end
N = [Dhaka(1:12)';M]; C = N(:,8:11);
%Generates random sampling w/replacement from data%
%and calculates the paramter estimates%
for k = 1:200
    R = floor(50*rand(50,1)+1);
    for j = 1:50
        dat(j,:) = C(R(j),:);
    end
    xb(k,:) = mean(dat);
    x(:,k) = fsolve(@(x)[xb(k,1)*exp(2*x(1))+...
    (x(2)*(exp(2*x(1))-1))./x(1)-xb(k,3);...
    xb(k,1)*exp(3*x(1))+...
    (x(2)*(exp(3*x(1))-1))./x(1)-xb(k,4)],[.8 ; 60]);
end
%Makes sure all estimates are positive%
for l = 1:200
    if x(1,l) < 0 || x(2,l) < 0
        L(l) = l;
    else
        L(l) = 0;
    end
end
s = sort(L);
n = s(201-mean(L>=1)*200:200);
for i=1:length(n)
    x(:,n(i)-(i-1))=[];
end
%Calculates the std. dev. of parameter est.%
%so we can create a conf. int.%
sig = sqrt(var(x'))
%or alternatively we can use the quantiles of the dist%
Y = [quantile(x(1,:),.975) quantile(x(1,:),.025);...
    quantile(x(2,:),.975) quantile(x(2,:),.025)]
```

Euler-Maruyama Approximation of SDE's

```
figure
M=input('Enter parameters as a matrix, each row is [T N X(0) lambda mu nu phi]: ');
for k = 1:2
    v = M(k,:);
    n = v(1)/v(2); Y1(1) = v(3); Y2(1) = v(3);
    W = normrnd(0,sqrt(n),v(2)-1,1);
    for i = 1:v(2)-1
        Y1(i+1) = max(0,Y1(i)+((v(4)-v(5))*Y1(i)+v(6))*n...
            +sqrt((v(4)+v(5))*Y1(i)+v(6))*W(i));
        Y2(i+1) = Y2(i)+((v(4)-v(5))*Y2(i)+v(6))*n...
            +v(7)*Y2(i)*W(i);
    end
    t = linspace(0,v(1),length(Y1));
    exval = v(3)*exp(t*(v(4)-v(5)))+(v(6)*(exp(t*(v(4)-v(5)))...
        -1))./(v(4)-v(5));
    vari = v(3)*(v(4)+v(5))/(v(4)-v(5))*exp(t*(v(4)-v(5))).*...
        (exp(t*(v(4)-v(5)))-1)+(v(6)*(v(4)*exp(t*(v(4)-v(5)))...
        -v(5)).*(exp(t*(v(4)-v(5)))-1))./(v(4)-v(5))^2;
    vari2 = (exp(t*v(7)^2)-1).*(exp(t*(v(4)-v(5)))*(v(6)...
        +(v(4)-v(5))*v(3))-v(6)).^2/(v(4)-v(5))^2;
    sd = sqrt(vari);
    sd2 = sqrt(vari2);
    subplot(2,2,2*k-1)
    plot(t,exval,'g',t,Y1,'k',t,exval+2*sd,'--b',t,max(0,exval-2*sd),'--b')
    xlabel('Time');ylabel('Population Size');
    title(['dX_t=((\lambda-\mu)X_t+\nu)dt+((\lambda+\mu)X_t+\nu)^{1/2}dW_t'],...
        'FontSize',14);
    legend('Expected Value','Process Value',...
        'Pred. Int.: E(X(t))\pm 2 SD',0);
    subplot(2,2,2*k)
    plot(t,exval,'g',t,Y2,'k',t,exval+2*sd2,'--b',t,max(0,exval-2*sd2),'--b')
    xlabel('Time');ylabel('Population Size');
    title(['dX_t=((\lambda-\mu)X_t+\nu)dt+\phiX_tdW_t'],...
        'FontSize',14);
    legend('Expected Value','Process Value',...
        'Pred. Int.: E(X(t))\pm 2 SD',0);
end
```

Summer REU 2005 for the Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48109

E-mail address: muratoa@umich.edu