Final Report: Crime Modeling

Yunbai Cao, Kun Dong, Beatrice Siercke & Matt Wilber

August 9, 2013

1 Introduction

The study of crime time series is an area that contains a wealth of information, but very little knowledge and much disagreement. For example, McDowall, Loftin and Pate (2011) note the large extent to which seasonality has been identified in crime data, but also that many studies that came to the opposite conclusion [8]. While some studies claim crime spikes occur in the summer, others observe them in the winter, and others in the months in between. Beyond this, much variation has been found in seasonality with respect to crime type. It has been believed that property crimes peak in the winter, and violent crimes in the summer.

However, we do know that crime has a level or organization and thus predicatability. It is well known that criminals are "moderately regular," meaning they are more successful when targeting a similar area repeatedly [4]. Such observations have led to self-exciting burglary process models that simulate individual criminals within a neighborhood [9, 12]. These have been based off of models that represent an "attractiveness field" of homes that govern movement of criminals within a neighborhood [10, 11].

Much investigation into seasonality of crime time series has gone into empirical statistical findings or behavioral theories explaining their occurrence. Temperature-aggression theories have been posited, arguing that crime increases with temperature. It has also been argued that seasons associated with more outdoor activity causes a greater susceptibility of households to theft and other crimes, since the owner is more often removed from the potential crime site.

On the other hand, less research has gone into the mathematical modeling of crime seasonality of time series, particularly in the field of differential equations. Current models are largely statistical, and are unable to provide significant insight or understanding of the seasonality of crime. A simple differential equations model would allow criminologists to develop a better understanding of the forces that cause crime to oscillate over time, and perhaps even provide predictability power.

With these advantages in mind, the group intends to study property crime, specifically, burglary, and develop a differential equations model that can well describe seasonal crime data. The group considered crime data from Los Angeles, California and Houston, Texas over the periods of 2005-2013 and 2009-2013, respectively. These data contained burglary rates in their respective city on a daily basis. These data are plotted below, with LA in blue and Houston in red.

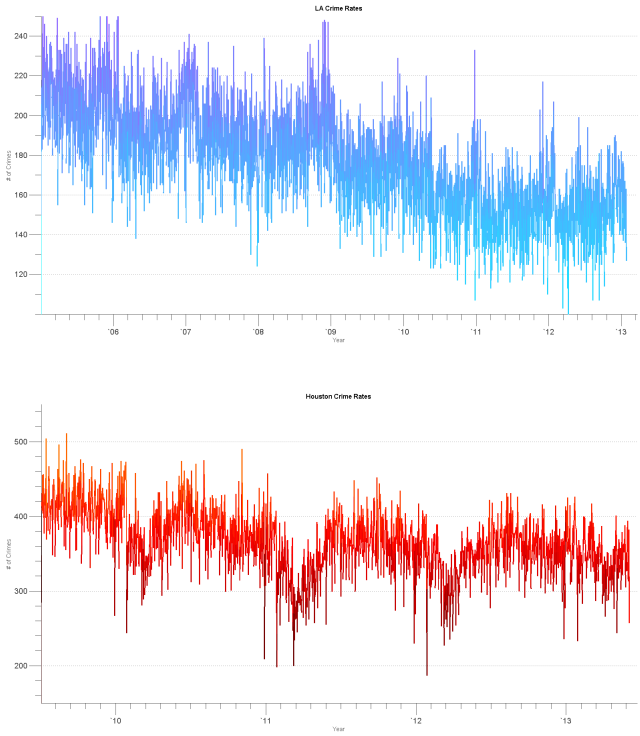

Figure 1: Los Angeles (blue) and Houston (red) burglary rates from '05-'13 and '09-'13, respectively.

The noisy data lends itself to a stochastic differential equation (SDE) model, a differential equation where one of the terms includes white noise. A well-known example of an SDE is Geometric Brownian Motion, which leads to the Black-Scholes partial differential equation for options pricing,

$$\frac{dP}{P} = \mu\, dt + \sigma\, dW \tag{1}$$

where P is represents the price of the option as a function of time, and μ represents the drift of the price, or the rate at which it is expected to increase or decrease. The σ parameter is a scaling factor that determines the magnitude of the white noise dW that is included in the model.

Before producing our own SDE model, however, we first need to extract the seasonal part of our crime data in order to simplify our analysis.

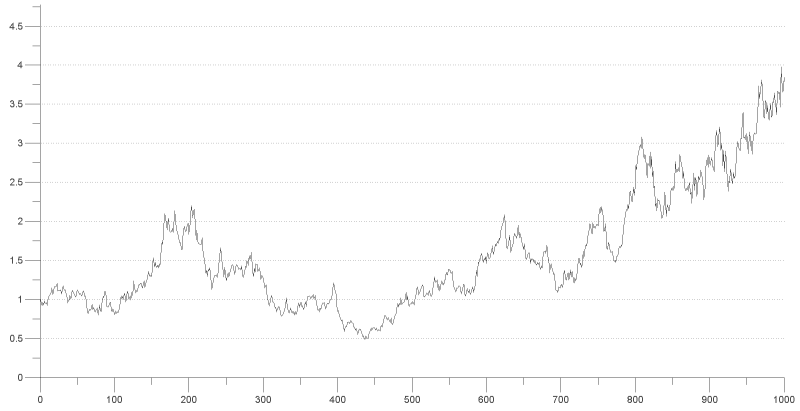

Figure 2: An example of Geometric Brownian Motion in Equation (1), used in the derivation of the Black-Scholes PDE.

2 Separating long term trends, seasonal trends, and noise

In order to better view the seasonality of the data, we extract the long term trend from the data by using Singular Spectrum Analysis (SSA) on each time series, a nonparametric method to obtain a low-rank approximation of our data. Given a time series vector X of length N and components $x_1, x_2, \ldots, x_N$, we choose a window length $L = 365$ corresponding the to the number of days in a year and suggesting yearly seasonality. Letting $K = N - L + 1$, we now produce a *trajectory matrix* of *lagged vectors* of length L,

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ x_3 & x_4 & x_5 & \cdots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_N \end{pmatrix}. \tag{2}$$

We then perform a standard singular value decomposition (SVD) on the trajectory matrix, decomposing it as

$$\mathbf{X} = \mathbf{X_1} + \ldots + \mathbf{X_L}. \tag{3}$$

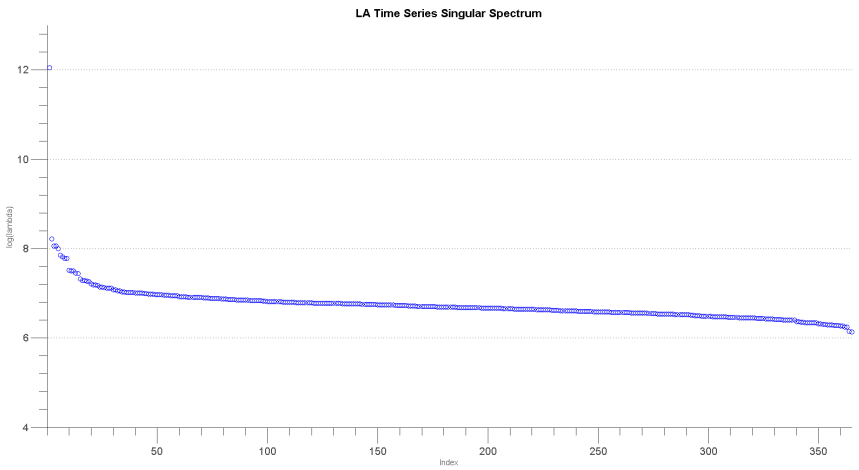

Figure 3: Logarithm of singular spectrum for the Los Angeles data set.

Next, we reconstruct each elementary matrix $\mathbf{X_I}$ into a time series of length N, where the kth term is the average of the antidiagonal $i + j = k + 1$, where $\mathbf{X_I}_{ij}$ is the element of $\mathbf{X_I}$ in the ith row and jth column. We result in L elementary reconstructed series, $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(L)}$, corresponding to L *singular values* $\lambda_1, \ldots, \lambda_L$. The Los Angeles data results in the singular values plotted in Figure 3.

The series corresponding to the largest λ_i is considered to be the long-term trend of our time series, as it is several orders of magnitude larger than the remaining singular values, as is the time series it corresponds to. To produce the seasonal component of our data, we sum element-wise all elementary reconstructed series with a period of approximately 365 days. The modes corresponding to the six largest singular values for the Los Angeles data are also plotted in Figure 4. As a result, we can single out the seasonality and noise components of the data, by subtracting the trend from the original data and adding back the mean of the trend, resulting in the figure below.

It is now easier to view and model the seasonality of the data. We observe distinct yearly upward spikes in the LA data, and downward spikes in the Houston data. This strongly suggests a seasonality in burglary in the two areas of interest. However, one peculiarity in the LA data is the lack of a distinct spike near the beginning of 2008, while a spike is observed in all other years in the data set. While this may be attributed to environmental factors such as the 2008 recession, the group

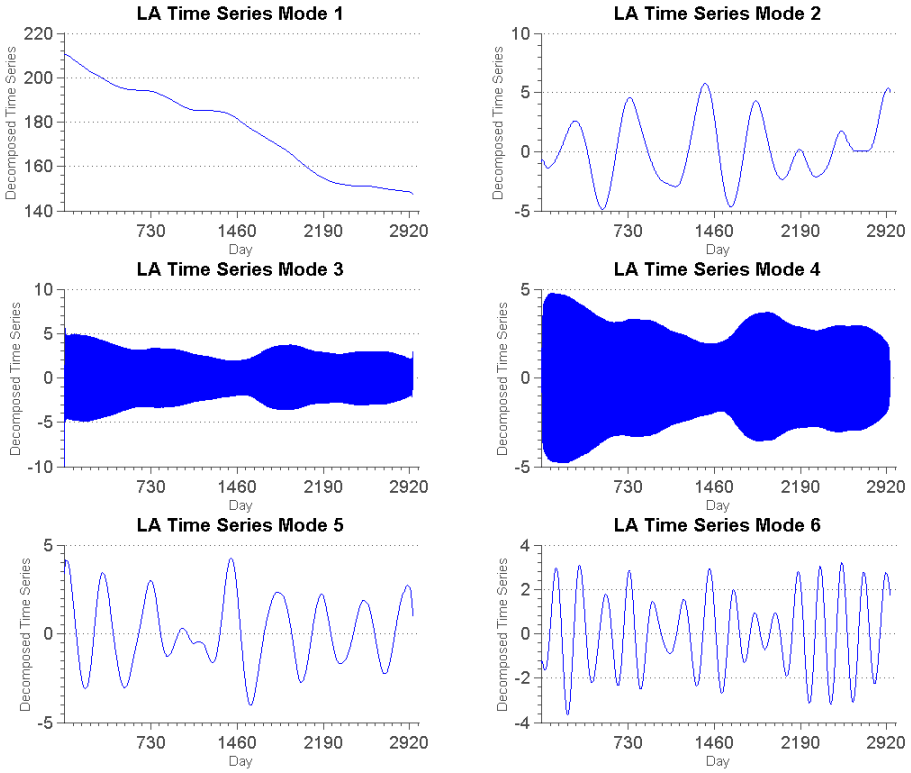

Figure 4: Modes corresponding to the six largest singular values of the LA Angeles data set. Mode 1 represents the long-term trend of the data, whereas the sum of Modes 2 and 5 provide the seasonal component of our data, with each having periods of approximately 365 days.

seeks to discover whether this can be explained by a simple differential equations model.

We may also compare seasonality in different types of crime data. We are able to use the same type of spectrum analysis to observe yearly periodicity in Houston Aggravated Assault data, as seen in Figure 6. On the other hand, Figure 7 reveals a semi-annual periodicity in Houston robbery data, and we are unable to find periodicity in Houston murder data, possibly due to daily murder counts being relatively very low.

In terms of the long term trends of each crime type, we see a general decrease

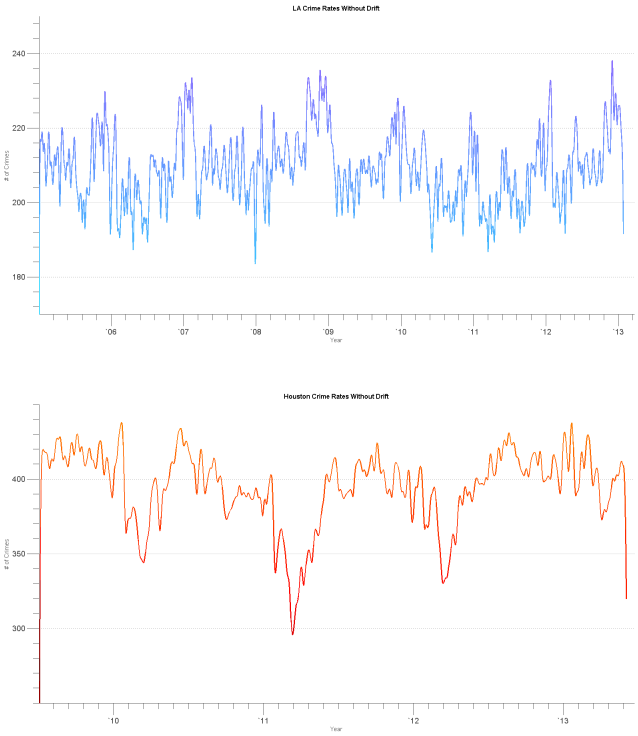

Figure 5: Smoothed LA and Houston data with trend extracted

from 2009-2013 in the Houston data. However, Auto Theft, Roberry, and Theft each show a distinct increase in the second half of the data set for the mode corresponding to their largest singular value. Each turning point occurs around the 730 day mark, during the summer of 2011. At the same time, we note that the Burglary and Rape data feature a short increase around 730 days, before they continue to fall. The table below provides the decrease of the long-term trend for each type of crime, as a percentage of the initial rate.

Crime Type	**Agg**	**Auto**	**Burg**	**Murd**	**Rape**	**Robb**	**Theft**
Percentage	19.75	5.73	15.69	29.18	21.66	18.15	10.74

Figure 7: Percentage decreases of Houston crime types relative to their rate on June 1, 2009.

We find that Murder is decreasing at the greatest rate in Houston, relative to its prevalence, and Auto Theft is decreasing at the slowest rate. However, all crime types are decreasing over time.

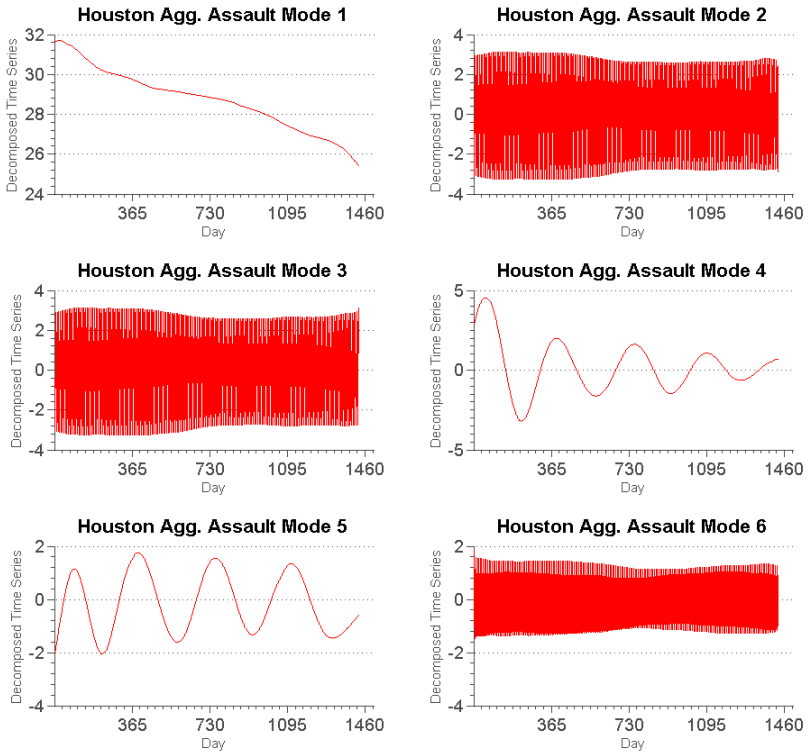

Figure 6: Modes corresponding to the six largest singular values of the Houston Aggravated Assault data set.

3 Modeling seasonal crime

In general, a two-dimensional Stochastic Differential Equation (SDE) can be written in the form

$$dX_t = f(X_t, Y_t)\,dt + g(X_t, Y_t)\xi(t), \qquad X_0 = X(0) \tag{4}$$

$$dY_t = \tilde{f}(X_t, Y_t)\,dt + \tilde{g}(X_t, Y_t)\tilde{\xi}(t), \qquad Y_0 = Y(0) \tag{5}$$

where $\xi(t)$ and $\tilde{\xi}(t)$ represent "white-noise" terms, which in our case will be equivalent to the Brownian motion term, dW. An SDE can be thought of as a type of differential equation with "randomness" built in. For example, one of the most important applications of SDEs is in economics, where geometric Brownian motion, an

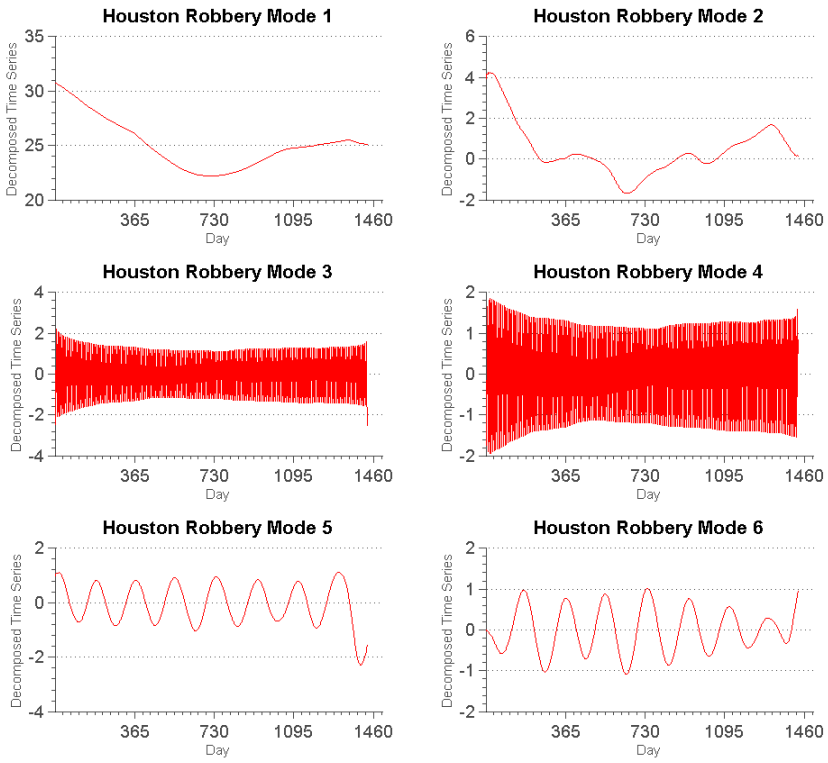

Figure 8: Modes corresponding to the six largest singular values of the Houston Robbery data set.

SDE, motivates the derivation of the Block-Scholes option pricing partial differential equation.

An SDE follows Itô's Chain Rule, rather than the standard chain rule. Given a function $u(X(t), t)$, its differential becomes

$$du = \frac{\partial u}{\partial t}\, dt + \langle \nabla f, d\mathbb{X}_t \rangle + \frac{1}{2} \langle \mathrm{Hess}(f) d\mathbb{X}_t, d\mathbb{X}_t \rangle \tag{6}$$

where

$$d\mathbb{X}_t = \begin{bmatrix} dX_t \\ dY_t \end{bmatrix}.$$

To qualitatively justify the use of SDEs to model the crime data, an SDE was simulated that has stable oscillations about the unit circle,

$$\begin{aligned} dX_t &= [r(1-r)X_t - Y_t]\,dt + \sigma_1\,dW_t \\ dY_t &= [r(1-r)Y_t - X_t]\,dt + \sigma_2\,dV_t \end{aligned} \tag{7}$$

where $r = \sqrt{{X_t}^2 + {Y_t}^2}$ and σ_1 and σ_2 are parameters that scale the independent, identically distributed white noise terms, dW_t and dV_t. The plots below compare the qualitative nature of this SDE to a set of aggravated assault data from our Houston dataset.

Figure 9: Houston aggravated assault data, smoothed (top) and SDE (2) (bottom).

3.1 Lotka–Volterra

The Lotka-Volterra Model is a linear system of differential equations, widely used in ecology to represent predator-prey interactions, and applicable in many fields. The system of equations is provided below.

$$\begin{aligned} \dot{x} &= x(\alpha - \beta y), \quad X_0 = X(0) \\ \dot{y} &= -y(\gamma - \delta x), \quad Y_0 = Y(0) \end{aligned} \tag{8}$$

Here, we consider x the prey population as a function of time, and y the predator population. The positive constants α and γ, respectively, relate the growth rate of the prey population and the decay rate of the predator population. The positive parameters β and δ relate the magnitude of the effects of interaction between the populations on each population.

The Lotka-Volterra system of equation have previously been used to model gang interactions (Brantingham et al. 2012) and can behave well in a criminology setting [1]. In the case of property crime, the predator population can be interpreted as crime rates in a given area, and the prey either a measure of property susceptibility or of property attractiveness to criminals.

The deterministic model has equilibria at $(x, y) = (0, 0)$ and $(x, y) = \left(\frac{\gamma}{\delta}, \frac{\alpha}{\beta}\right)$. The model permits periodic solutions about the nontrivial equilibrium, similar to the behavior of the Los Angeles and Houston crime data. The system also has a conserved energy,

$$E = \alpha \log y + \gamma \log x - \beta y - \delta x \tag{9}$$

where $x, y \neq 0$. This can be shown to be constant in time. Taking the time derivative of the equation, we find

$$\begin{aligned} \dot{E} &= \frac{\alpha}{y}\dot{y} + \frac{\gamma}{x}\dot{x} - \beta\dot{y} - \delta\dot{x} \\ &= -\left(\frac{\alpha}{y} - \beta\right) y(\gamma - \delta x) + \left(\frac{\gamma}{x} - \delta\right) x(\alpha - \beta y) \\ &= -\alpha\gamma + \alpha\delta x + \beta\gamma y - \beta\delta xy + \alpha\gamma - \beta\gamma y - \alpha\delta x + \beta\delta xy \\ &= 0. \end{aligned}$$

This energy is also a maximum at the non-trivial equilibrium, and decreases as orbits get further from equilibrium.

3.2 A stochastic Lotka–Volterra equation

The model we propose is a Stochastic approach to a Lotka-Volterra model, as shown below.

$$\begin{aligned} dX_t &= X_t(\alpha - \beta Y_t)\,dt + \sigma_1 X_t\,dW_t, \qquad X(0) = X_0 \\ dY_t &= -Y_t(\gamma - \delta X_t)\,dt + \sigma_2 Y_t\,dV_t, \qquad Y(0) = Y_0. \end{aligned} \tag{10}$$

This system of equations is able to capture both, the periodic element of our original data by means of the Lotka-Volterra aspect, as well as representing noise found in the data by means of the Stochastic aspect of the model. The predator-prey equations become our functions $f(X_t, Y_t)$ found in equation (4) and $\tilde{f}(X_t, Y_t)$ found in equation (5). The parameters α, β, γ, and δ remain to have the same significance as described in the previous section. In the second term $\sigma_1 X_t$ and $\sigma_2 Y_t$ are the $g(X_t, Y_t)$ and $\tilde{g}(X_t, Y_t)$ functions found in equations (4) and (5) respectively. Here σ_1 and σ_2 are constants that scale the identically independent distributed noise generated by the terms dW_t and dV_t which conform to a Wiener process. Driven to have a realistic representation of the data, the white noise is multiplied by the population, in this manner the noise will be proportional to the population; so that greater oscillations will occur when there are high rates of crimes.

Because of the presence of Lotka-Volterra equations we naturally considered the energy of the model. Using Itô's Calculus (6) we found the following:

$$E_t = E_0 + \left[\int_0^t \sigma_1(\delta - \gamma X_t)\,dW_t + \sigma_2(\alpha - \beta Y_t)\,dV_t\right] - \frac{1}{2}({\sigma_1}^2\gamma + {\sigma_2}^2\alpha)t \tag{11}$$

The expectation of the energy, shown below, should obey the Stochastic equation using Itô's lemma.

$$\mathbb{E}(E_t) = \mathbb{E}(E_0) - \frac{1}{2}({\sigma_1}^2\gamma + {\sigma_2}^2\alpha)t. \tag{12}$$

As time increases the second term will also increase, therefore the energy is expected to decrease over time. Another expectation is that the orbits tend to leave equilibrium, which would result in our data expanding outward. The data would be expanding to orbits with smaller periods, and this heeds that there is built-in period variation between orbits. The last expectation is the greatest one, the expected value can be used as a means to validate a weak order of the scheme.

Simulating our model many times with the same number of timesteps as days in the Los Angeles data, we are able to compute the simulation with the minimum

sum of the least-squares difference between the instance values and the data itself. Figure 3 provides the instance with the minimum score generated, plotted against seasonal data extracted from the Los Angeles data set using SSA.

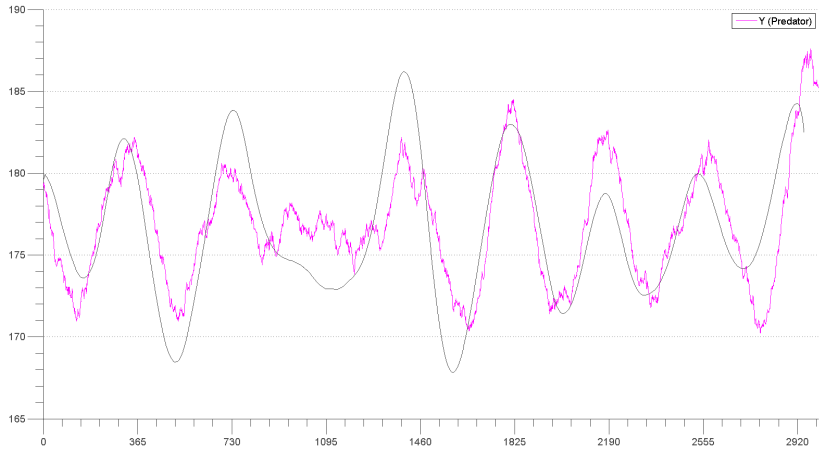

Figure 10: Simulation of Lotka–Volterra numerical scheme (magenta) in Equation (5), using parameters $\alpha = 6.805$, $\gamma = 0.42888$, $\beta = \delta = 0.0385$, $X_0 = 10.436$, $Y_0 = 179.545$. Only the values for Y_t are plotted. Plotted against SSA-extracted seasonal data from Los Angeles (black).

Note that we were able to reproduce a "missed period" near the end of the third year, as we observed in the Los Angeles data. Later we will discuss the ability of our SDE model to consistently produce such a pattern, which would indicate the possibility of a non-environmental explanation for the missed period.

Similarly, we were able to use a least-squares scoring method to compare our model to the raw data, with the long term trend (found by SSA) subtracted. Figure 4 on the next page gives plots of the minimum-scoring runs for both the Los Angeles and Houston data sets. Again, we are able to reproduce the desired period skip in the Los Angeles plot.

Figure 11: Minimum least squares scoring trials of SDE model for both raw LA and raw Houston data, each with their long term trend removed. Both data sets were smoothed using a moving window average for viewing purposes. Raw data is plotted in blue, the SDE in red.

4 Numerics for the model

4.1 Numerical scheme

In order to numerically simulate Equation (10), we use a semi-implicit method, which can be derived by setting

$$\begin{aligned} X_{t+1} &\approx X_t + (\alpha X_t\,\Delta t - \beta X_{t+1} Y_t\,\Delta t + \sigma_1 X_t\,\Delta W_t) \\ Y_{t+1} &\approx Y_t + (\gamma X_t Y_t\,\Delta t - \delta Y_{t+1}\,\Delta t + \sigma_2 Y_t\,\Delta V_t) \end{aligned} \tag{13}$$

where Δt is the timestep and $\Delta W_t, \Delta V_t \sim \sqrt{\Delta t} \cdot N(0,1)$. Solving for X_{t+1}, Y_{t+1} results in the scheme

$$\begin{aligned} X_{t+1} &= \frac{1 + \alpha\Delta t + \sigma_1 \Delta W_t}{1 + \beta Y_t \Delta t} X_t \\ Y_{t+1} &= \frac{1 + \gamma X_t \Delta t + \sigma_2 \Delta V_t}{1 + \delta \Delta t} Y_t \end{aligned} \tag{14}$$

Experimentally, this numerical scheme has first-order weak convergence, but has the advantage of being able to be run very quickly in mathematical software.

Now let's compare the semi-implicit method with the traditional Euler-Maruyama method

$$\begin{aligned} X_{t+1} &\approx X_t + (\alpha X_t\,\Delta t - \beta X_t Y_t\,\Delta t + \sigma_1 X_t\,\Delta W_t) \\ Y_{t+1} &\approx Y_t + (\gamma X_t Y_t\,\Delta t - \delta Y_t\,\Delta t + \sigma_2 Y_t\,\Delta V_t) \end{aligned} \tag{15}$$

which calculates X_{t+1}, Y_{t+1} directly from X_t and Y_t. If we drop off the random noise part in the system, then Equation (14) becomes

$$\begin{aligned} X_{t+1} &= \frac{1 + \alpha\Delta t}{1 + \beta Y_t \Delta t} X_t \\ Y_{t+1} &= \frac{1 + \gamma X_t \Delta t}{1 + \delta \Delta t} Y_t \end{aligned} \tag{16}$$

and Equation (15) becomes

$$\begin{aligned} X_{t+1} &\approx X_t + (\alpha X_t\,\Delta t - \beta X_t Y_t\,\Delta t) \\ Y_{t+1} &\approx Y_t + (\gamma X_t Y_t\,\Delta t - \delta Y_t\,\Delta t) \end{aligned} \tag{17}$$

It can be easily observed from the above equations that the semi-implicit method guarantees the positivity of X_{t+1}, Y_{t+1} when X_t, Y_t and other parameters are positive. In contrast, the Euler-Maruyama method doesn't have this advantage.

If we consider the random noise part and calculate X_{t+1}, Y_{t+1} as in Equation (14) and Equation (15), we can derive from Equation (14) that the probability that $X_{t+1} \geq 0$ given $X_t, Y_t \geq 0$

$$P\left(X_{t+1} \geq 0 \mid X_t, Y_t \geq 0\right) \tag{18}$$

is equal to

$$P\left(1 + \alpha\Delta t + \sigma_1\sqrt{\Delta t}N(0,1) \geq 0\right) \tag{19}$$

which is

$$P\left(N(0,1) \geq \frac{-1-\alpha\Delta t}{\sigma_1\sqrt{\Delta t}}\right) \tag{20}$$

If we use Euler-Maruyama method as our numerical method we will get

$$P\left(N(0,1) \geq \frac{\beta Y_t\Delta t - 1 - \alpha\Delta t}{\sigma_1\sqrt{\Delta t}}\right) \tag{21}$$

for the probability that $X_{t+1} \geq 0$ given $X_t, Y_t \geq 0$ which can be derived from Equation (15). Observing the two equations above one can see that the semi-implicit method provides a higher possibility for each step to be positive than the traditional Euler-Maruyama method when given the previous step is positive. In another words, the semi-implicit method is more stable and thus allows one to take larger step sizes.

4.2 Extracting system parameters

It is always a crucial step for any model to get a close estimation of the parameters in order to reproduce the data we attempt to simulate. In our case, the stochastic Lotka-Volterra model requires six parameters α, β, γ, δ, σ_1, σ_2 as well as initial data x_0, x_0. (10)

The common approach of parameter fitting for stochastic model is using the maximum-likelihood function. However this method does not yield a satisfatory result for our model, due to both the number of parameters and the noisiness in our data. In order to efficiently maximize the likelihood function, we need either a better method to locate global maximum or an extremly accurate initial guess. To avoid this issue, we took a different approach.

We perform the least square fitting on our data against the second variable y in the solution of the ordinary Lotka-Volterra equations and extract the most suitable values for parameters α, β, γ, δ and intial data x_0, y_0. For simplicity we make the assumption that impact of interaction on the predator and the prey is quantitatively equivalent, hence $\beta = \delta$ in Equation (8). After renaming the parameters such that $\frac{\alpha}{\delta} = \hat{\alpha}$, $\frac{\gamma}{\delta} = \hat{\gamma}$ we adopt the following form of Lotka-Volterra equations.

$$\begin{aligned} \dot{x} &= \delta x(\hat{\alpha} - y), \qquad x_0 = x(0) \\ \dot{y} &= -\delta y(\hat{\gamma} - x), \quad y_0 = y(0) \end{aligned} \tag{22}$$

To conduct the fitting we use a Matlab function called `lsqcurvefit` which is written to find coefficients x that solve the problem

$$\min_x \|F(x, xdata) - ydata\|_2^2 \tag{23}$$

In our model F is the solution to Equation (22) produced by `ode45`, the fourth order Runge-Kutta method. Here, x represents the set of parameters $[\delta, \hat{\alpha}, \hat{\gamma}, x_0, y_0]$, $xdata$ is the vector $[0.01, 0.02, \cdots, 29.45]$ representing our time window of 2945 days and $ydata$ is our time series.

Obtaining an accurrate initial guess is essential for a reasonable parameter fitting and very beneficial to improve computational efficiency. The major goal when we are getting this initial guess is to ensure that the solution has period of around 365 days. Assuming the solution is within certain proximity of the equilibrium we can use the formula below to estimate the period.

$$T = \frac{2\pi}{\delta\sqrt{\hat{\alpha}\hat{\gamma}}} \approx 365 \tag{24}$$

Based on the properties of Equation (22), $(\hat{\gamma}, \hat{\alpha})$ is the non-trivial equilibrium of the system. While $\hat{\gamma}$ can be arbitrarily determined, we take $\hat{\alpha}$ to be the mean of our data. With values of $\hat{\alpha}$ and $\hat{\gamma}$, as well as the above formula, we can determine a desired value of δ. Furthermore, the values of x_0 and y_0 are taken to be some point nearby the equilibrium. To get the best fitting possible, we start with the initial guess and keep iterating the method `lsqcurvefit` until the return values stay approximately constant. As an example of our results, we get the parameters of LA Burglary to be $\delta = 0.0395$, $\hat{\alpha} = 176.5514$, $\hat{\gamma} = 11.1315$, $x_0 = 10.4410$ and $y_0 = 179.5454$.

We can use the parameter approximation from the ordinary Lotka-Volterra model as our initial guess to get estimation of σ_1 and σ_2 in the SDE model through maximum-likelihood function.

5 Missed periods

Since the span of the data was 8 years it is reasonable to expect 8 peaks for a crime with annual seasonality. However, one may note in the Los Angeles burglary data that there are only 7 peaks, namely a missed peak near the end of 2007 and the beginning of 2008. This is intriguing because it disrupts the seasonal trend of the data. Rather than rationalizing this anomaly as an effect of the financial crisis of 2008 in the United States or some other global environmental cause, we are curious to know if the model could capture a missed period without any additional input.

To determine the number of peaks in a simulation, we first need to define the notion of a peak, since noisy data has many local maxima and we seek only those that occur at the largest value of an approximately yearly oscillation. Our definition of a peak was a point that surpassed a threshold, often 180 units in this case, and must also be 10 units above nearby data points.

We also set a threshold, *minpeakdistance*, for the distance between peaks. If two peaks are so far away from each other that they exceed this threshold, a period is considered missing. This threshold takes value between 450 and 500 in most cases, because 450 is the length of a year with an extra season and any peaks happen in a different season is potentially missing a period. Meanwhile, we would like to remove those cases that move towards orbits with much lower energy levels. Such simulations will have a much larger amplitude of oscillation as well as a longer period, thereby no longer resemble our data. We can filter out those cases with a threshold on the standard deviation of the peaks' heights, which is set as 5 through our observation.

Tables 1 and 2 illustrate our results with this method, under various sets of parameters. It is evident that the percentage of simulations with missing period varies little with respect to different σ_1 values, while it falls from around 55% at $\sigma_2 = 0.01$ to 5% at $\sigma_2 = 0.03$. Moreover, this percentage varies between 17.0% and 22.5% when minpeakdistance takes value between 450 and 550.

Clearly the model is less sensitive to changes in σ_2— as columns remain within

σ_1 / σ_2	**0.010**	**0.015**	**0.020**	**0.025**	**0.030**
0.000	56.9	36.3	21.3	11.0	5.5
0.010	56.8	37.6	20.2	11.2	5.0
0.020	59.8	35.7	19.4	10.6	5.0
0.030	55.0	34.6	18.7	8.2	4.0

Table 1: Percent Missed Periods vs. Noise Patameter Variations out of 1000 simulations. Uses Los Angeles parameters $\alpha = 6.805$, $\beta = 0.0385$, $\gamma = 0.042888$, $\delta = 0.0385$, $X_0 = 10.436$, and $Y_0 = 179.545$, with $minpeakdistance = 470$.

Min. Peak Distance	**450**	**475**	**500**	**525**	**550**
Percent Missed Peaks	22.5	21.7	19.2	16.0	17.0

Table 2: Percent Missed Periods vs. Min. Peak Distance Variations out of 1000 simulations. Uses Los Angeles parameters $\alpha = 6.805$, $\beta = 0.0385$, $\gamma = 0.042888$, $\delta = 0.0385$, $X_0 = 10.436$, $Y_0 = 179.545$, $\sigma_1 = 0$, and $\sigma_2 = 0.02$.

$\pm$ 2 units— than to changes in σ_1—as rows vary by more than 50 units. Quite obviously it is reasonably sensitive to the minimum distance between peaks considered. Consider the case in which the minimum is 450 days, the percentage of missed peaks is high since it automatically includes simulations that have missed peaks for a greater distance, such as 550 days apart.

A possible issue with this method is an undercounting of smaller peaks. For example, in a situation such as the Figure:12, the sixth and seventh peaks may discarded by this method due to insufficient height difference from nearby data, even though they show peak behavior. This will create a missing period between five and eight peaks, which is a debatable result since the missed period is a fairly qualitative event.

In order to ensure a lower bound on the number of missed peaks were found, a second method was developed to reinforce the first one. For this method peaks only need to be 7 units above the nearby data and 179.6 in height, thereby more peaks are taken into consideration. Any two peaks within 44 days from each other are attributed to noise and only the higher one is admitted. Additionally, peaks within 103 days from each other are group into peak clusters, which is considered as the seasonal peak as a whole. Minpeakdistance still applies the same, except now if the first appearance of a peak is beyond this threshold a period is also considered

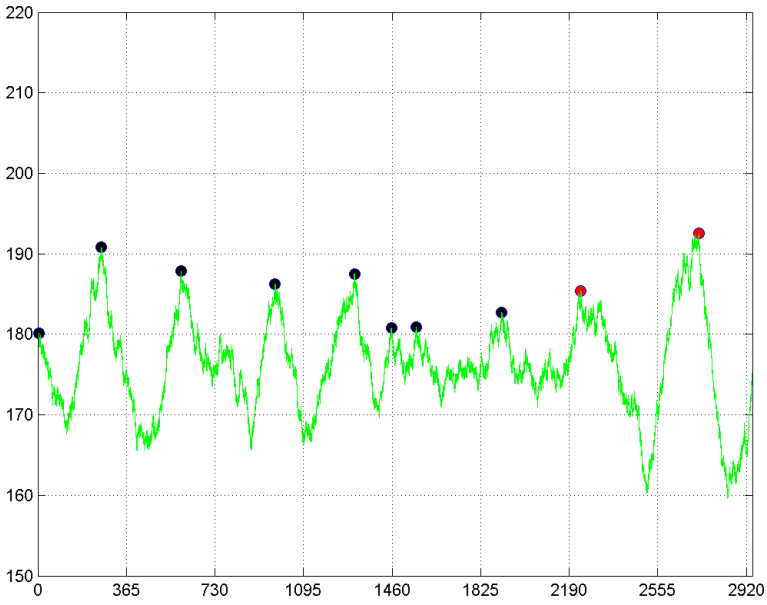

Figure 12: A simulation with missing period based on second method. Uses Los Angeles parameters $\alpha = 6.805$, $\beta = 0.0385$, $\gamma = 0.042888$, $\delta = 0.0385$, $X_0 = 10.436$, $Y_0 = 179.545$, $\sigma_1 = 0$, and $\sigma_2 = 0.02$.

missing. Also, if there are less than 8 peak clusters a period is missing. Finally, we remove simulations such that the standard deviation of peak heights is greater than 8. This method gives the results in Tables 3 and 4.

The same pattern demonstrated in first method persists, and we are consistently getting 10 to 20 percent of our simulations with a missing period. This supports our hypothesis that missing period is an intrinsic behavior of our stochastic model. Nevertheless, it is noticeable that the percentage is sensitive with respect to σ_2 in both methods, meaning we need either an accurate approximation on this parameter or some improvement on our methods for better stability.

We also viewed our solutions in the phase plane. We observe that during the missed period the solution is pushed towards the equilibrium and remains in close proximity, as seen in figure 10. This can explain mathematically why there is a missed period, namely the solution is lifted into a high-energy orbit by randomness, in which it possesses low oscillation, until it escapes to lower orbits under the

σ_1 / σ_2	**0.010**	**0.015**	**0.020**	**0.025**	**0.030**
0.000	32.4	27.9	19.5	12.1	6.9
0.010	38.9	30.6	18.0	13.9	6.7
0.020	39.5	27.2	17.4	13.4	8.1
0.030	40.5	28.0	19.3	12.6	8.1

Table 3: Percent Missed Periods vs. Noise Patameter Variations out of 1000 simulations. Uses Los Angeles parameters $\alpha = 6.805$, $\beta = 0.0385$, $\gamma = 0.042888$, $\delta = 0.0385$, $X_0 = 10.436$, and $Y_0 = 179.545$, with $minpeakdistance = 470$.

Min. Peak Distance	**450**	**460**	**470**	**480**	**490**	**500**
Percent Missed Peaks	27.5	22.6	20.5	18.5	14.1	12.8

Table 4: Percent Missed Periods vs. Min. Peak Distance Variations out of 1000 simulations. Uses Los Angeles parameters $\alpha = 6.805$, $\beta = 0.0385$, $\gamma = 0.042888$, $\delta = 0.0385$, $X_0 = 10.436$, $Y_0 = 179.545$, $\sigma_1 = 0$, and $\sigma_2 = 0.02$.

combined effort of noise and decreasing energy of the model. Thus, we are able to describe missing periods through populations dynamics, rather than through external causes.

6 SDE for other seasonal crime types

Beyond viewing the viability of the model over different geographic locations, it is also of interest to view its applicability across crime types. Using the SSA method described previously, trends may be extracted for Houston data limited to aggravated assault, auto theft, burglary, murder, rape, robbery, and theft.

As a whole, we see that the long term trend for each crime type is to decrease over time. However, around the end of the second year of the data, corresponding to the summer of 2011, we notice that auto theft, burglary, rape, and robbery display a distinct increase in crime rates, with only burglary and rape continuing to decrease afterwards. Similarly, we are able to extract a yearly seasonal component for aggravated assault, auto theft, burglary, rape, and theft in Houston.

It must be noted, however, that due to the relatively small number of rape crimes per day, we see very small oscillations on the order of 0.1 crimes per day. As a result, we may view the rape seasonality to be less reliable than the others. On

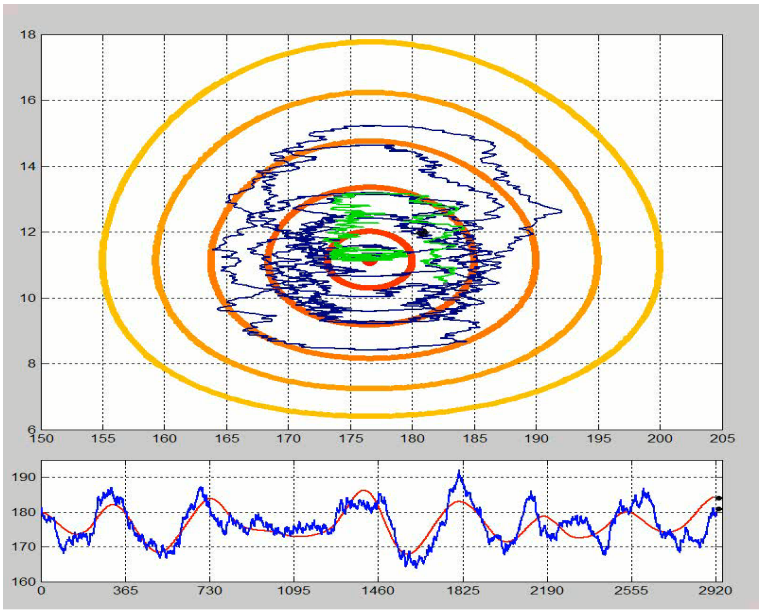

Figure 13: LA Burglary Data plotted on phase plane (above) and on x-y plane (below). Missed Period (green) remains near the equilibrium. Uses Los Angeles parameters $\alpha = 6.805$, $\beta = 0.0385$, $\gamma = 0.042888$, $\delta = 0.0385$, $X_0 = 10.436$, $Y_0 = 179.545$, $\sigma_1 = 0$, and $\sigma_2 = 0.02$.

the other hand, we are unable to extract a distinct yearly seasonal component from murder data or robbery data in Houston, using the SSA method described above. To illustrate this, we plot the six modes of each time series corresponding to the six largest singular values of their corresponding trajectory matrix. We find that, other than the trend, the modes are either noise or oscillate at frequencies much greater than once per year.

Using least squares parameter fitting on the seasonal component of both Los Angeles and Houston aggravated assault data, we were able to apply our stochastic Lotka-Volterra model to aggravated assault data, both from Los Angeles and Houston. Taking several runs and returning the minimum least-squares difference from the data, we result in the following plots.

Neither aggravated assault data set has a missed period, as in the Los Angeles burglary data, nor widely varying peaks, as in the Houston burglary data. As a result, we are able to produce an instance that behaves very similarly to the aggravated assault data. Both our ability to extract trend and seasonal components

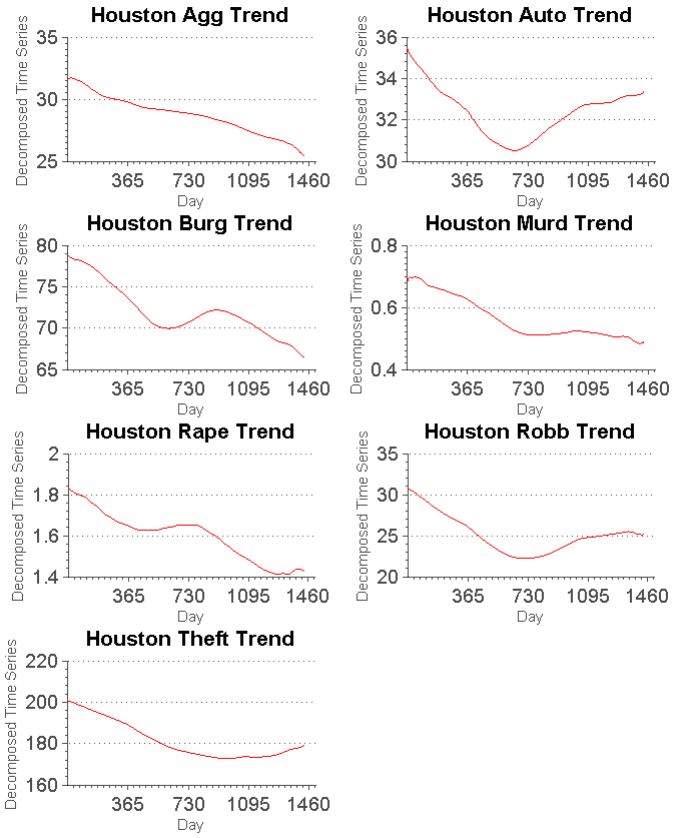

Figure 14: Long term trends for aggravated assault, auto theft, burglary, murder, rape, robbery, and theft in Houston.

of many crime types, as well as applying our SDE model to aggravated assault data, suggest that our model has applicability over a wide variety of crime types.

7 Conclusions

We have shown that the SSA method may be used to partition a crime time series into long term trend, seasonal, and noise components. In terms of applications of the technique, such partitioning may be used in the future to forecast crime trends, by providing law enforcement agencies a better understanding of the seasonality of a crime data set, and thus the ability to expect surges and troughs in crime frequency.

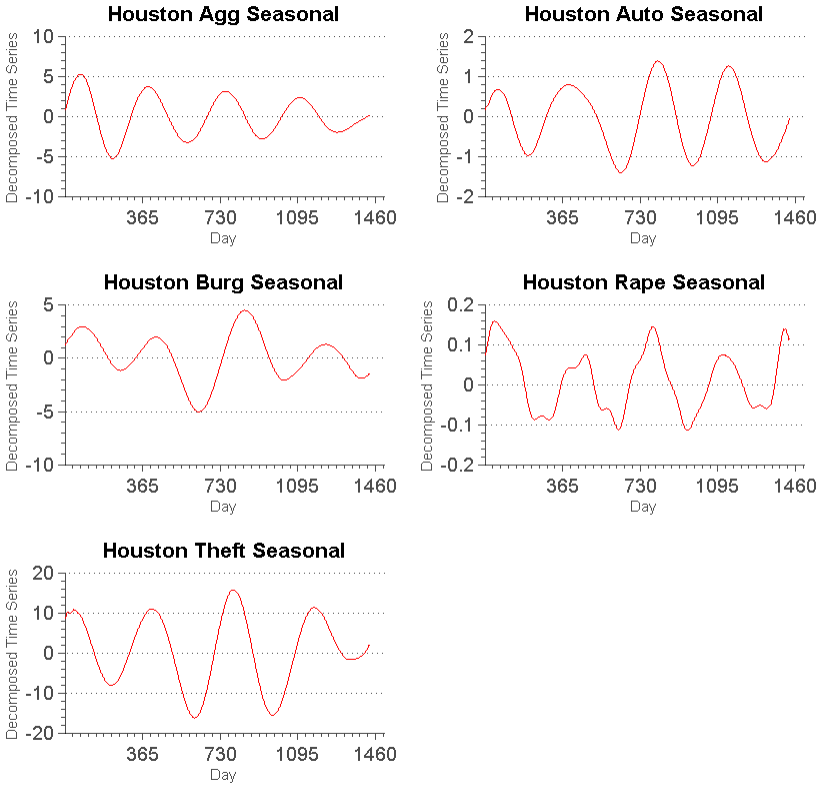

Figure 15: Yearly seasonality for aggravated assault, auto theft, burglary, rape, and theft in Houston.

Furthermore, rather than the manual division of modes into each of their three respective categories, clustering techniques may be used in the future to automate this process, in an unsupervised process. As a result, any missed periods may be quickly detected, allowing law enforcement to adjust their actions with warnings ahead of time. This may lead to more efficient use of resources, such as the time of police officers.

The apparent ability of the stochastic Lotka Volterra model described in this paper to simulate results very similar to those seen in our Los Angeles and Houston data sets can be seen as suggestive towards a relatively novel view of crime seasonality. Rather than attributing large seasonal fluctuations of crime rates to external forces, such as temperature or economic variation, we now consider the possibility of

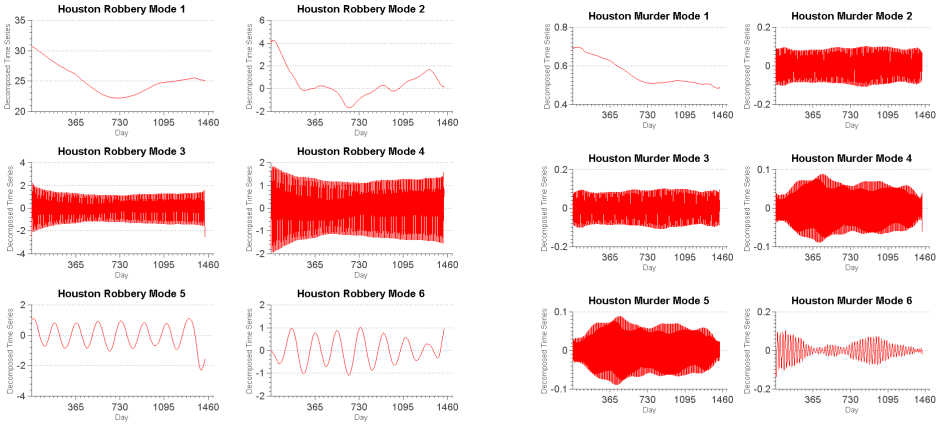

Figure 16: Modes of Houston robbery and murder time series corresponding to the six largest singular values of the corresponding trajectory matrix with window length $L = 365$ days. Note that Houston Robbery Mode 2 displays troughs on a yearly basis, but they are not nearly as distinct as in other data sets.

viewing crime rates as a population dynamics system. Future studies may consider whether crime can be described as a predator-prey system, with crime rates representing predation and susceptible targets per criminal representing prey. If this can be shown, a better understanding of the parameters that lead to certain orbits of crime may allow us to more efficiently and quickly decrease crime rates followed parameter sensitivity analysis. For example, maintaining a system very near equilibrium could lead to a stagnation of crime rates and much easier predictability of crime trends.

References

[1] P Jeffrey Brantingham, George E Tita, Martin B Short, and Shannon E Reid. The ecology of gang territorial boundaries*. *Criminology*, 50(3):851–885, 2012.

[2] Kevin Burrage, PM Burrage, and Tianhai Tian. Numerical methods for strong solutions of stochastic differential equations: an overview. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 460(2041):373–402, 2004.

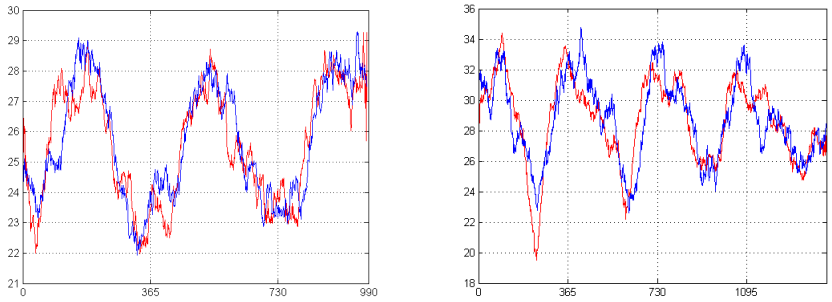

Figure 17: Stochastic Lotka-Volterra model applied to LA and Houston aggravated assault time series. Los Angeles parameters: $\alpha = 6.805$, $\beta = 0.0385$, $\gamma = 0.042888$, $\delta = 0.0385$, $X_0 = 10.436$, $Y_0 = 179.545$, $\sigma_1 = 0$, and $\sigma_2 = 0.02$. Houston parameters: $\alpha = 5.168$, $\beta = 0.0720$, $\gamma = 0.0720$, $\delta = 0.42888$, $X_0 = 10.436$, $Y_0 = 179.545$, $\sigma_1 = 0$, and $\sigma_2 = 0.03$.

[3] Lawrence C Evans. An introduction to stochastic differential equations version 1.2. *Department of Mathematics UC Berkeley, in internet*, 2001.

[4] MARCUS FELSON. What every mathematician should know about modelling crime. *European Journal of Applied Mathematics*, 21(4-5):275–281, 2010.

[5] Desmond J Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, 43(3):525–546, 2001.

[6] Joseph Ian Jeisman. Estimation of the parameters of stochastic differential equations. 2006.

[7] Changho Kim, Eok Kyun Lee, Peter Hänggi, and Peter Talkner. Numerical method for solving stochastic differential equations with poissonian white shot noise. *Physical review E*, 76(1):011109, 2007.

[8] David McDowall, Colin Loftin, and Matthew Pate. Seasonal cycles in crime, and their variability. *Journal of Quantitative Criminology*, 28(3):389–410, 2012.

[9] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493), 2011.

[10] Ashley B Pitcher and Shane D Johnson. Exploring theories of victimization using a mathematical model of burglary. *Journal of Research in Crime and Delinquency*, 48(1):83–109, 2011.

[11] Martin B Short, Maria R D'ORSOGNA, Virginia B Pasour, George E Tita, Paul J Brantingham, Andrea L Bertozzi, and Lincoln B Chayes. A statistical model of criminal behavior. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1249–1267, 2008.

[12] MB Short, MR DOrsogna, PJ Brantingham, and GE Tita. Measuring and modeling repeat and near-repeat burglary effects. *Journal of Quantitative Criminology*, 25(3):325–339, 2009.

[13] Angel Tocino and Jesus Vigo-Aguiar. Weak second order conditions for stochastic runge–kutta methods. *SIAM Journal on Scientific Computing*, 24(2):507–523, 2002.

[14] Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380, 2002.