A Stochastic Model for Seasonal Cycles in Crime

Yunbai Cao  Kun Dong  Beatrice Siercke  Matt Wilber

August 9, 2013
Introduction
Objective

We would like to create a mathematical model for the seasonal variation of burglary in Los Angeles, CA and Houston, TX, in order to forecast future crime trends. While seasonality of crime data is widely agreed upon, it is poorly understood, and we seek to determine if a model can reproduce seasonality for a wide variety of data sets without relying on environmental factors.
Criminology Background

Inconsistent Seasonal Folklore

- Disagreement over seasonality of crimes
- Variation with crime type
- Burglary peak in summer or winter
- Geographical variation of behavior
- Inconsistent hypotheses of what drives seasonality
Data Sets

- Data sets contain property crime rates
- LA data 2005-2013, Houston 2009-2013
- Would like to divide into long-term trend, seasonal, and noise components $X_t = T_t + S_t + N_t$
Methods
Singular Spectrum Analysis

- Decomposition of time series \( X = (x_1, x_2, \ldots, x_N) \) into elementary reconstructed series
- Window length \( L = N - K + 1 \)
- Low-rank approximation of original time series
- Allows us to extract long-term trend, seasonality, and noise

\[
M = \begin{pmatrix}
    x_1 & x_2 & x_3 & \cdots & x_K \\
    x_2 & x_3 & x_4 & \cdots & x_{K+1} \\
    x_3 & x_4 & x_5 & \cdots & x_{K+2} \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    x_L & x_{L+1} & x_{L+2} & \cdots & x_N
\end{pmatrix}
\]
Interested in the first six or so singular values
Singular Spectrum Analysis

- First mode corresponds to long-term trend
- Modes 2 and 5 display yearly periods
Trend Extraction

Los Angeles

Houston
Seasonality Extraction
Methods

Seasonality Extraction

Los Angeles

Houston

(UCLA, Harvey Mudd)
Crime Type Comparisons

- Aggravated Assault, Murder, and Theft have a clear downward trend
- Auto Theft and Robbery dip then rise
- Burglary and Rape briefly increase after two years, but otherwise decrease
Crime Type Comparisons

- Aggravated Assault, Auto Theft, Burglary, Rape, and Theft all have clear yearly seasonal components
Murder has no seasonal component, only noise

Robbery has a semi-annual seasonal component
Crime Type Comparisons

In summary,

- Extracted a clear trend for each crime type
- Most types have a yearly seasonal component
- Murder and Robbery do not have yearly components
We will consider two-dimensional SDEs of the form

\[ dX_t = f(X_t, Y_t) \, dt + g(X_t, Y_t) \, dW_t \]

\[ dY_t = \tilde{f}(X_t, Y_t) \, dt + \tilde{g}(X_t, Y_t) \, dV_t \]

- White-noise terms \( dW_t \) and \( dV_t \) represent Brownian motion
Ito Calculus uses a different chain rule. If we consider $u(t, x, y)$, then

$$du = \frac{\partial u}{\partial t} \, dt + \langle \nabla u, dX_t \rangle + \frac{1}{2} \langle \text{Hess}(u) dG, dG \rangle$$

where

$$dX_t = \begin{bmatrix} dX_t \\ dY_t \end{bmatrix} \quad \text{and} \quad dG = \begin{bmatrix} g(X_t, Y_t) \, dW_t \\ \tilde{g}(X_t, Y_t) \, dV_t \end{bmatrix}.$$
Lotka-Volterra

- Predator-Prey Model
- Growth/decay rate parameters $\alpha$, $\gamma$
- Interaction effect parameters $\beta$, $\delta$
- Used in criminology literature
- Natural periodic solutions

Follows the equations

$$\dot{x} = x(\alpha - \beta y), \quad x(0) = x_0$$
$$\dot{y} = -y(\gamma - \delta x), \quad y(0) = y_0$$
Lotka-Volterra

- Used a stochastic version of Lotka-Volterra,

\[ dX_t = X_t(\alpha - \beta Y_t) \, dt + \sigma_1 X_t \, dW_t, \quad X(0) = X_0 \]
\[ dY_t = -Y_t(\gamma - \delta X_t) \, dt + \sigma_2 Y_t \, dV_t, \quad Y(0) = Y_0. \]

- White noise, IID terms \( dW_t \) and \( dV_t \)
- Noise is proportional to size of population at time \( t \)
- \( X_t = \) Available targets per criminal
- \( Y_t = \) Crime rate
Energy of the Model

- Well-known energy of the Lotka-Volterra Model:

\[ E = \alpha \log y + \gamma \log x - \beta y - \delta x \]

- Using Itô's Lemma, we find

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

Expected Energy

- Taking expectation,

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

- Energy decreases over time
- Orbit tends to leave equilibrium
- Built-in period variation
- Can be used to validate weak order of scheme
Used semi-implicit scheme, derived by setting

\begin{align*}
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{align*}

where \( \Delta t \) 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

$$X_{t+1} = \frac{1 + \alpha \Delta t + \sigma_1 \Delta W_t}{1 + \beta Y_t \Delta t} X_t$$

$$Y_{t+1} = \frac{1 + \gamma X_t \Delta t + \sigma_2 \Delta V_t}{1 + \delta \Delta t} Y_t.$$

- Weak first order, but quick to compute
- Maintains positivity with very high probability
Results
Simulations

Plot

- $\sigma_1 = 0$, $\sigma_2 = 0.02$
- Least square parameter estimation
- Minimum least squares difference from data
- Missing period
Simulations

Los Angeles

Houston

- Reproduced missing period
Simulations

Los Angeles Aggravated Assault

- Data in red, simulation in blue
- Better fit without missed period or widely varying peak heights
- Shows that our model works across a variety of data
Simulations

Houston Aggravated Assault

- Data in red, simulation in blue
- Criminological implications
Results

LA Burglary Video
Houston Burglary Video
Want to identify peaks in time series

Use peakfinder.m, by Nathanael Yoder

Peakfinder.m

- Used for finding “local peaks or valleys ... in a noisy vector.”
- Input: A real vector
- Output: \([pks, \text{locs}],\) two vectors with peak heights and peak locations, respectively.
- Options
  - \(\text{sel}:\) The “amount above surrounding data for a peak to be identified”
  - \(\text{thresh}:\) A threshold above which maxima must be to be considered peaks
Peaks must be 10 units above nearby data

Peaks must be at least 180 units high
Method A

- If any pair of peaks are at least \( \text{minpeakdistance} \) days apart, a period has been missed.
Method $\Omega$

- Peaks must be 7 units above nearby data
- Peaks must be at least 179.6 units high
- Peaks must be at least 44 days apart
- Peaks within 103 days apart each other are grouped into the same peak cluster
Method Ω

- If any pair of peaks are at least minpeakdistance days apart, a period has been missed.
- If the first peak appears after minpeakdistance days, a period has been missed.
- If the number of peak clusters is less than 8, a period has been missed.
Noise Parameter Analysis: Method A

<table>
<thead>
<tr>
<th>( \sigma_1 / \sigma_2 )</th>
<th>0.010</th>
<th>0.015</th>
<th>0.020</th>
<th>0.025</th>
<th>0.030</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>56.9</td>
<td>36.3</td>
<td>21.3</td>
<td>11.0</td>
<td>5.5</td>
</tr>
<tr>
<td>0.01</td>
<td>56.8</td>
<td>37.6</td>
<td>20.2</td>
<td>11.2</td>
<td>5.0</td>
</tr>
<tr>
<td>0.02</td>
<td>59.8</td>
<td>35.7</td>
<td>19.4</td>
<td>10.6</td>
<td>5.0</td>
</tr>
<tr>
<td>0.03</td>
<td>55.0</td>
<td>34.6</td>
<td>18.7</td>
<td>8.2</td>
<td>4.0</td>
</tr>
</tbody>
</table>

- Percentage relatively sensitive to noise
- Varies little with \( \sigma_1 \)
- Decreases with \( \sigma_2 \)
Results

Noise Parameter Analysis: Method $\Omega$

Percent Missed Periods vs. Noise Parameters

<table>
<thead>
<tr>
<th>$\sigma_1/\sigma_2$</th>
<th>0.010</th>
<th>0.015</th>
<th>0.020</th>
<th>0.025</th>
<th>0.030</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>32.4</td>
<td>27.9</td>
<td>19.5</td>
<td>12.1</td>
<td>6.9</td>
</tr>
<tr>
<td>0.01</td>
<td>38.9</td>
<td>30.6</td>
<td>18.0</td>
<td>13.9</td>
<td>6.7</td>
</tr>
<tr>
<td>0.02</td>
<td>39.5</td>
<td>27.2</td>
<td>17.4</td>
<td>13.4</td>
<td>8.1</td>
</tr>
<tr>
<td>0.03</td>
<td>40.5</td>
<td>28.0</td>
<td>19.3</td>
<td>12.6</td>
<td>8.1</td>
</tr>
</tbody>
</table>

- Varies little with $\sigma_1$
- Decreases with $\sigma_2$
Minpeakdistance Analysis: Method A

- Both percentages decrease as minpeakdistance increases.
- Percentages relatively insensitive to minpeakdistance parameter.
- \( \sigma_1 = 0, \sigma_2 = 0.02 \)

<table>
<thead>
<tr>
<th>minpeakdistance</th>
<th>450</th>
<th>475</th>
<th>500</th>
<th>525</th>
<th>550</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Missed Peaks</td>
<td>22.5</td>
<td>21.7</td>
<td>19.2</td>
<td>16.0</td>
<td>17.0</td>
</tr>
</tbody>
</table>
Minpeakdistance Analysis: Method $\Omega$

- Both percentages decrease as minpeakdistance increases
- Percentages relatively sensitive to minpeakdistance parameter
- $\sigma_1 = 0$, $\sigma_2 = 0.02$

<table>
<thead>
<tr>
<th>minpeakdistance</th>
<th>450</th>
<th>460</th>
<th>470</th>
<th>480</th>
<th>490</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Missed Peaks</td>
<td>27.5</td>
<td>22.6</td>
<td>20.5</td>
<td>18.5</td>
<td>14.1</td>
<td>12.8</td>
</tr>
</tbody>
</table>
Conclusions

- Decomposed crime data into long-term trend, seasonal, and noise components
- Produced an SDE model to simulate burglary rates over several year spans
- Affirmed model’s ability to reproduce missed periods as in LA data
Acknowledgements

- Jeffrey Brantingham, for the Los Angeles crime data
- Houston Police Department, for the Houston crime data
- Scott McCalla and James von Brecht for lots of advice and mentoring