Spatio-Temporal Low Count Processes with Application to Violent Crime Events

Sivan Aldor-Noiman[1], Lawrence D. Brown[2], Emily B. Fox[3] and Robert A. Stine[2]

[1]The Climate Corporation
[2]University of Pennsylvania
[3]University of Washington

May 24, 2016

Supplementary Material

This supplementary material provides further details on our MCMC sampler in Section S1 and the baseline models to which we compare in Section S2. In Section S3, we elaborate upon the simulation studies outlined in the main paper. Finally, in Sections S4-S6, we provide additional details on our Washington, D.C. crime data and results.

S1 MCMC sampler derivation

In this section, we detail the derivation of the sampling steps presented in Section 5 of the main paper. We first introduce the following notation:

- $S_i = \sum_{t=1}^{T} \epsilon_{i,t}$
- $\boldsymbol{S} = [S_1, \ldots, S_N]$
- $\boldsymbol{S}_{-i} = [S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_N]$
- $z_{-i} = [z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_N]$
- $\Theta = \sum_{t=1}^{T} \theta_{s(t)}$

Step 1 – Sampling the innovations vectors In Step 1 of the MCMC sampler (Section 5) we motivate the sampling of the innovation vector. We have shown in (16) that conditional on the data and other model parameters, the innovations are independent of each other and we only need to sample $\epsilon_{i,t}$ when both the current observed value, $y_{i,t}$, and the previous observed value, $y_{i,t-1}$, have positive values. The values of $\epsilon_{i,t}$ in this case will range between $max\{0, y_{i,t} - y_{i,t-1}\} \leq \epsilon_{i,t} \leq y_{i,t}$. Otherwise, $\epsilon_{i,t}$ is set deterministically. The distribution of a single innovation, $\epsilon_{i,t}$ is a :

$$\begin{aligned} P(\epsilon_{i,t}|y_{i,t-1}, y_{i,t}, \alpha_i, \lambda_i^{'}, \boldsymbol{\theta}) &\propto P(y_{i,t}|\epsilon_{i,t}, y_{i,t-1}, \alpha_i) P(\epsilon_{i,t}|\lambda_i^{'}, \boldsymbol{\theta}) \\ &= \binom{y_{i,t-1}}{y_{i,t} - \epsilon_{i,t}} \alpha_i^{(y_{i,t} - \epsilon_{i,t})} (1-\alpha_i)^{(y_{i,t-1} - (y_{i,t} - \epsilon_{i,t}))} \frac{e^{-\lambda_{z_i}\theta_{s(t)}} \cdot (\lambda_{z_i} \cdot \theta_{s(t)})^{\epsilon_{i,t}}}{\epsilon_{i,t}!} \\ &\propto \frac{1}{\epsilon_{i,t}!(y_{i,t} - \epsilon_{i,t})!(y_{i,t-1} - (y_{i,t} - \epsilon_{i,t}))!} \left(\frac{\lambda_{z_i}\theta_{s(t)}(1-\alpha_i)}{\alpha_i}\right)^{\epsilon_{i,t}} \\ &= \frac{1}{C_i} \frac{1}{\epsilon_{i,t}!(y_{i,t} - \epsilon_{i,t})!(y_{i,t-1} - (y_{i,t} - \epsilon_{i,t}))!} \left(\frac{\lambda_{z_i}\theta_{s(t)}(1-\alpha_i)}{\alpha_i}\right)^{\epsilon_{i,t}} \end{aligned}$$

We can calculate the normalization constant C_i by summing over all the possible values of $\epsilon_{i,t}$ for a given set of values of $y_{i,t}$ and $y_{i,t-1}$.

Step 2 – Sampling the membership indicator In the following equations we

construct the posterior distribution for the membership indicator variable:

$$P(z_i = j|\boldsymbol{z}_{-i}, \boldsymbol{S}, \boldsymbol{\theta}) \quad \propto \quad P(S_i|S_l \quad l \in \{v : z_v = j, v \neq i\}, \boldsymbol{\theta}) \cdot P(z_i = j|z_l \quad l \in \{v : z_v = j, v \neq i\}).$$

It is straightforward to show that the above distribution has the following form:

$$P(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{S}, \boldsymbol{\theta}) \propto \begin{cases} \alpha \cdot p_{i,0} & \text{for } k = K+1 \\ n_k \cdot p_{i,k} & \text{for } k = 1, \ldots, K, \end{cases}$$

where $p_{i,0}, p_{i,1}, \ldots, p_{i,K}$ take on values from the following negative binomial distributions:

$$\begin{aligned} p_{i,0} &= \frac{\Gamma(S_i + \gamma_1)}{\Gamma(\gamma_1) S_i!} \left(\frac{\gamma_2}{\Theta + \gamma_2}\right)^a \left(\frac{\Theta}{\Theta + \gamma_2}\right)^{S_i} \\ p_{i,j} &= \frac{\Gamma(S_i + A_j + \gamma_1)}{\Gamma(A_j + \gamma_1) S_i!} \left(1 - \frac{\Theta}{n_j \cdot \Theta + \gamma_2}\right)^{A_j + \gamma_1} \left(\frac{\Theta}{n_j \cdot \Theta + \gamma_2}\right)^{S_i} \quad j = 1, \ldots, K, \end{aligned}$$

and $n_j = \sum_{i=1}^{L} \mathbb{I}_{z_i = j}$ and $A_j = \sum_{l: z_l = j, l \neq i} S_l$. This distribution corresponds to the posterior distribution shown in Step 2 of the MCMC sampler (Section 5).

Step 3 – Sampling the unique rate vector Since we use a gamma distribution as the DP base measure, the resulting conditional posterior distribution for the unique cluster-specific rates is as follows:

$$\begin{aligned} P(\phi_k|\boldsymbol{z}, \boldsymbol{S}, \boldsymbol{\theta}, \gamma_1, \gamma_2) &\propto P(\boldsymbol{S}_l, \quad l \in \{v : z_v = k\}|\phi_k, \boldsymbol{\theta}, \gamma_1, \gamma_2) \cdot P(\phi_k|\gamma_1, \gamma_2) \\ &\propto \phi_k^{B_k + \gamma_1 - 1} e^{-\phi_k \cdot (n_k \cdot \Theta + \gamma_2)} \end{aligned}$$

This has the form of a gamma distribution with parameters $B_k + \gamma_1$ and $n_k \cdot \Theta + \gamma_2$ where $B_k = \sum_{l \in \{v: z_v = k\}} S_l$ which is the distribution described in Step 3 of the MCMC sampler (Section 5).

Step 4 – Sampling the seasonal effect Let $R_t = \sum_{i=1}^{L} \epsilon_{i,t}$, then the conditional posterior distribution for the seasonal effect is:

$$\begin{aligned} P(\theta_j|\boldsymbol{\lambda}, \boldsymbol{z}, \boldsymbol{\epsilon}, \xi_1, \xi_2) &\propto P(\boldsymbol{\epsilon}_t, \quad l \in \{t : s(t) = j\}|\boldsymbol{\lambda}, \xi_1, \xi_2) \cdot P(\theta_j|\xi_1, \xi_2) \\ &\propto \theta_j^{\sum_{t: s(t) = j} R_t + \xi_1 - 1} e^{-\theta_j \cdot (m_j \cdot \sum_{l=1}^{L} \lambda_l' + \xi_2)} \end{aligned}$$

This is a gamma distribution with parameters $\sum_{t: s(t) = j} R_t + \xi_1$ and $m_j \cdot \sum_{l=1}^{L} \lambda_l' + \xi_2$ where $m_j = |t : s(t) = j|$.

Step 5 – Sampling the thinning value The prior distribution for the thinning value α_l is a beta distribution, resulting in the conditional posterior distribution:

$$P(\alpha_i|\boldsymbol{y}, \boldsymbol{\epsilon}, \eta_1, \eta_2) \quad \propto \quad \alpha^{\sum_{t=2}^{T} (y_{i,t} - \epsilon_{i,t}) + \eta_1 - 1} \cdot (1 - \alpha)^{\sum_{t=2}^{T} (y_{i,t-1} - (y_{i,t} - \epsilon_{i,t})) + \eta_2 - 1}$$

This is a beta distribution with parameters $\sum_{t=2}^{T} y_{i,t} - S_i + \eta_1$ and $\sum_{t=2}^{T}(y_{i,t-1} - y_{i,t}) + S_i + \eta_2$.

Step 6 – Sampling the concentration parameter

We follow **?** and use a gamma prior for the concentration parameter, τ. This stage requires first sampling an auxiliary variable κ which is then used to sample τ:

1. Sample $\kappa \sim \text{Beta}(\tau + 1, L)$.

2. Sample τ as the following mixture of two gammas:

$$\begin{aligned} \tau|\kappa, K &\sim \pi\text{Gamma}(a_\tau + K, b_\tau - \log(\kappa)) \\ &+ (1-\pi)\text{Gamma}(a_\tau + K - 1, b_\tau - \log(\kappa)), \end{aligned}$$

with weight π defined by $\pi/(1-\pi) = (a_\tau + K - 1)/(L \cdot [b_\tau - \log(\kappa)])$ where K is the number of unique clusters.

S2 Conditional least squares model

The PoINAR(1) model can be described in the following manner:

$$y_{t+1} = \alpha \circ y_t + \epsilon_{t+1} \quad t = 1, \ldots, T-1 \tag{S2.1}$$

$$\epsilon_t \sim \text{Poiss}(\lambda \cdot \theta_{s(t)}) \tag{S2.2}$$

The one-step-ahead conditional expected value for y_{t+1} is:

$$\hat{y}_{t+1} = \alpha \cdot y_t + \lambda \cdot \theta_{s(t)} \tag{S2.3}$$

The conditional least squares method estimates this model's parameters by solving the following equation

$$\min_{\lambda,\alpha,\theta_1,\ldots,\theta_{12}} \sum_{t=2}^{T} (y_t - \hat{y}_t)^2 = \min_{\lambda,\alpha,\theta_1,\ldots,\theta_{12}} \sum_{t=2}^{T} \left(y_t - \alpha \cdot y_t - \lambda \cdot \theta_{s(t)}\right)^2 \quad \text{s.t.} \sum_{s=1}^{12} \theta_s = 1. \tag{S2.4}$$

This is a nonlinear convex optimization problem. The Lagrangian method yields the following conditions:

$$\alpha = \frac{\sum_{t=2}^{T} y_t \cdot y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} - \lambda \cdot \frac{\sum_{t=2}^{T} \theta_{s(t)} y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} \tag{S2.5}$$

$$\lambda = \frac{\left(\sum_{t=2}^{T} y_{t-1}^2\right) \cdot \left(\sum_{t=2}^{T} y_t \cdot \theta_{s(t)}\right) - \left(\sum_{t=2}^{T} y_t \cdot y_{t-1}\right) \left(\sum_{t=2}^{T} y_{t-1} \theta_{s(t)}\right)}{\left(\sum_{t=2}^{T} y_{t-1}^2\right) \cdot \left(\sum_{t=2}^{T} \theta_{s(t)}^2\right) - \left(\sum_{t=2}^{T} y_{t-1}^2 \cdot \theta_{s(t)}\right)^2} \tag{S2.6}$$

$$\theta_i = \frac{2 \cdot \lambda \sum_{t:s(t)=i} (y_t - \alpha \cdot y_{t-1}) - C}{2\lambda^2 \cdot n_i} \quad i = 1, \ldots, 12 \tag{S2.7}$$

$$C = \frac{2 \cdot \lambda}{\sum_{i=1}^{12} \frac{1}{n_i}} \left(\sum_{i=1}^{12} \left[\frac{\sum_{t:s(t)=i} (y_t - \alpha \cdot y_{t-1})}{n_i} \right] - \lambda \right) \tag{S2.8}$$

Starting from a set of initial values, we can iterate between these equations and converge to a solution. The convergence is met within a few cycles.

S3 Simulation study

To assess the performance of our model, we simulate 9 different datasets from our multivariate PoINAR(1) process. Each dataset has $L = 100$ time series (locations) with $T = 208$ observations. The multiple time series are grouped into four equally sized clusters defined by a shared rate value, ϕ_k. The different data sets vary in the levels of separation between the cluster rates and the time series autocorrelation values, α_l. In this section, we evaluate the performance of our methods both in- and out-of-sample. The results show that our model can reasonably recover the ground-truth clusterings and also produce accurate out-of-sample forecasts under various settings. Our model also outperforms the conditional least-squares model (CLS), which is detailed in Section S2. One reasonable explanation for these results is that the CLS model does not allow for sharing of information between the time series and therefore is more prone to noise variation.

S3.1 Simulation settings

We have two main factors that we configure in each of the simulated data sets:

- The clusters' assigned rate values ϕ_k. We examine an "easy" setting in which the four cluster rate values are $1, 3, 6, 10$, a "medium setting" with values $0.01, 0.5, 1.2, 2$ and a "hard" setting with values $0.1, 0.2, 0.3, 0.6$. The rates values are well separated in the easy setting and become harder to distinguish as we move to the hard setting.

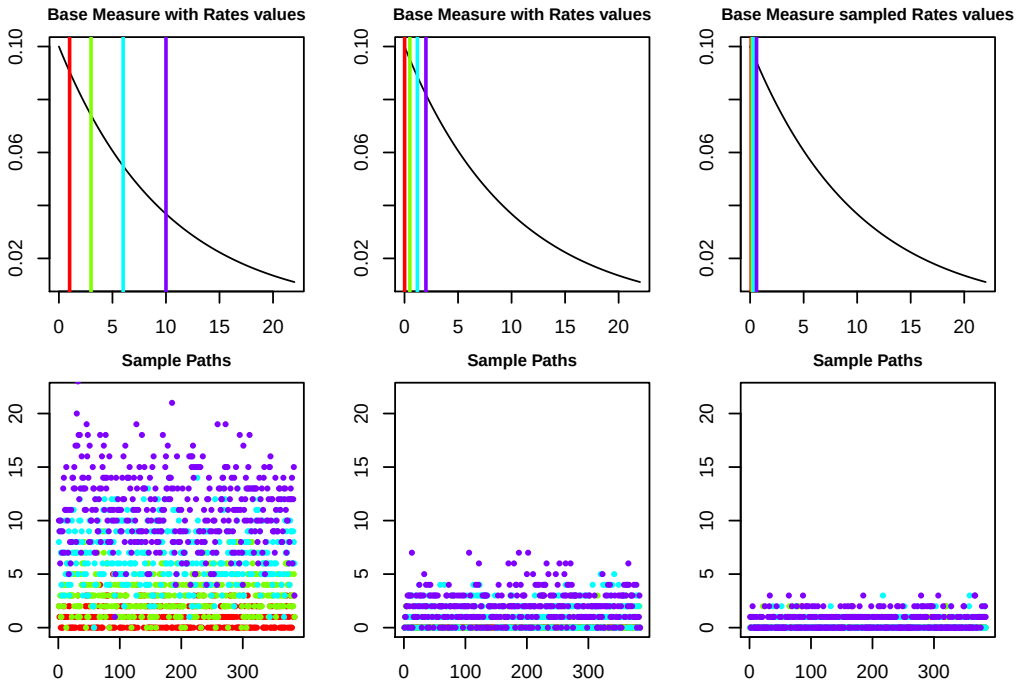

Figure 1: The top panel shows the rates values for the four different clusters, along with the prior distribution for the rates. The lower panel shows an example of 4 overlaid simulated time series. Each time series corresponds to a different cluster and is colored accordingly.

- The thinning value α_l, which directly relates to the autocorrelation values of the individual PoINAR(1) processes. We use three different thinning values shared between all locations: $0.1, 0.5, 0.9$.

Figure 1 illustrates examples of the simulated data for the three different rate scenarios using a thinning value of 0.3. In the "easy" setting, the tract means fall into four clear clusters. In the "hard" setting, it is much more difficult to distinguish between the four clusters solely based on the tract means. Furthermore, we can see that as the thinning value grows the tract means become larger and consequently it is easier to identify the clusters.

S3.2 Simulation results

Although we are primarily interested in the out-of-sample performance of our method, there are still two important measures that are useful to examine in-sample: (i) How many clusters does our method recover? (ii) How close is the recovered clustering assignment to the true assignment? By finding accurate clusterings, our method can borrow

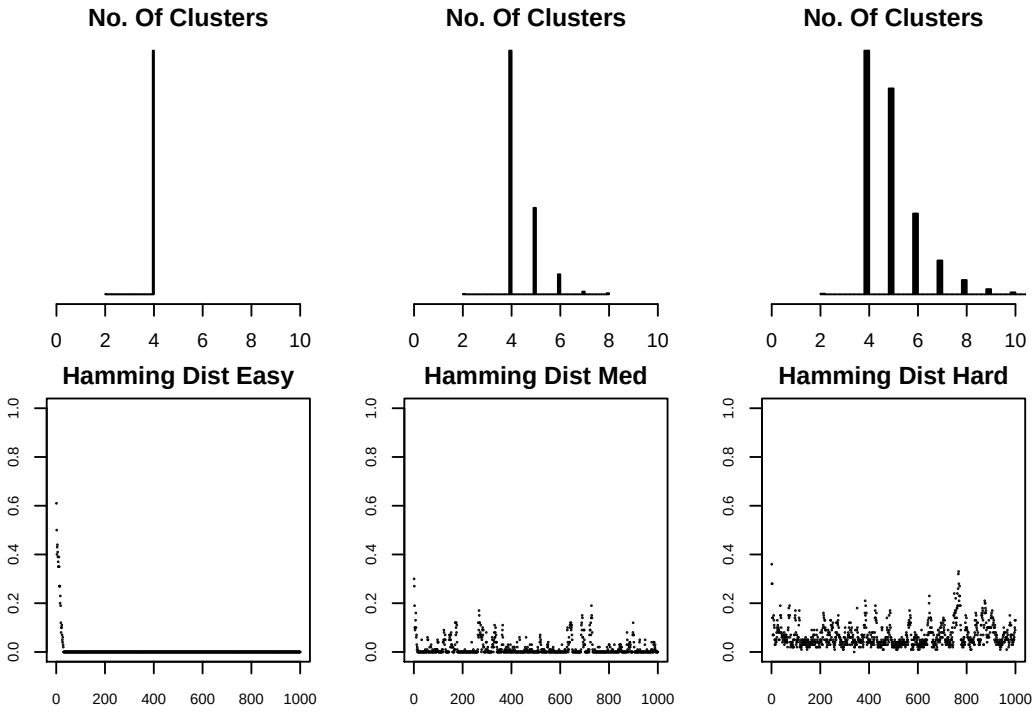

Figure 2: In-sample simulations results. The top panel displays the histogram for the number of clusters over the 1000 iterations. The bottom panel shows the Hamming distance errors between the estimated and true cluster assignments versus MCMC iteration.

information across the multiple time series yielding more accurate out-of-sample predictions.

We run our sampler for 1000 iterations for each of the 9 settings. Figure 2 displays histograms for the number of inferred clusters for each of the scenarios plotted in Figure 1. Figure 2 also shows the Hamming distances between the estimated and true clustering assignment labels. The distances are calculated by first choosing the optimal mapping of indices maximizing overlap between the true and estimated labels assignment sequences. As seen, the modal number of clusters is four in all of the three settings and, as expected, the method recovers the true clustering assignment more accurately for the easy setting than for the hard one. However, even for the hard setting, the Hamming distance errors are usually less than 10% indicating that most time series are correctly clustered. Although we only display the in-sample analysis for these three settings, these results generally hold for all 9 settings.

An interesting question is whether the methodology finds clustering structure when in fact all the time series belong to the same cluster. To examine this, we simulate a data set that has all of the time series grouped into a single cluster. Figure 3 shows the data and the results for the corresponding MCMC sampler. The model predominantly prefers to group all of the time series together, as we would hope in such a case.

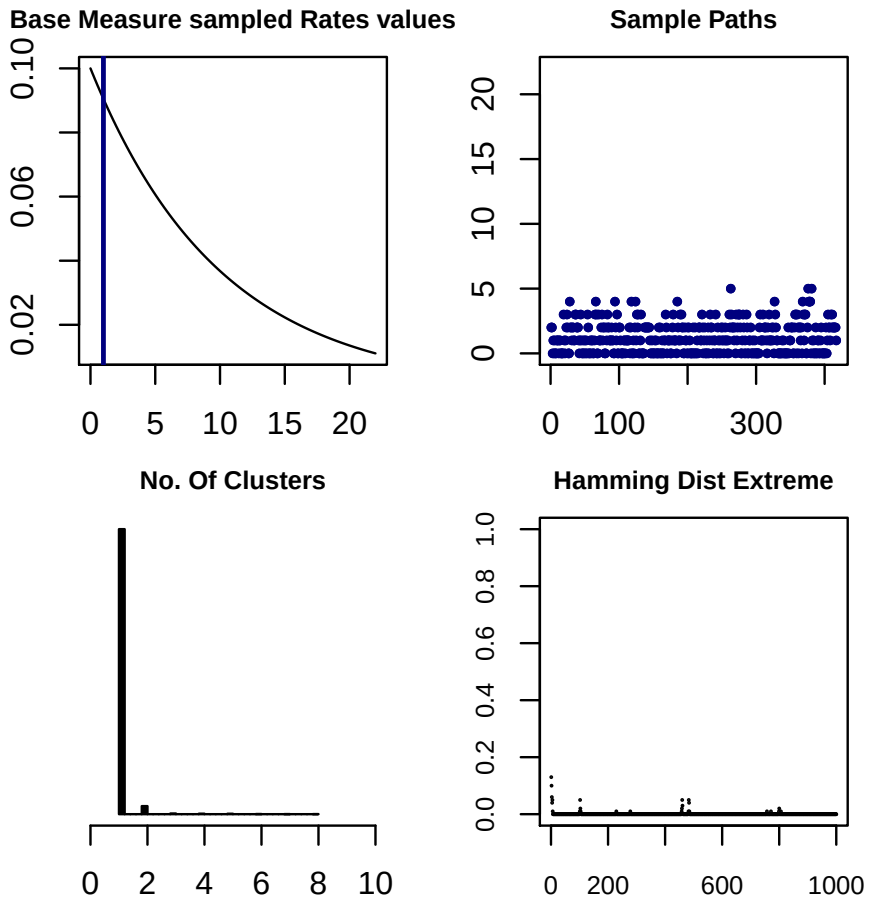

Figure 3: Simulation results for a case where there is a single cluster. The top panels display the single cluster rate value and an example of one of the time series' count data. The bottom panels display the histogram of the number of clusters our method finds and the corresponding Hamming distance errors versus MCMC iteration.

In order to evaluate our model's estimation performance, we compare the estimated one-step-ahead conditional expectation of the PoINAR(1) versus its corresponding true (simulated) population value. **?** showed that the h-step-ahead conditional expectation

of the PoINAR(1) model is:

$$\begin{aligned}\hat{y}_{T+h} &= \mathbb{E}(y_{T+h}|y_1,\ldots,y_T)\\ &= \mathbb{E}(\alpha^h \circ y_T + \sum_{j=1}^{h} \alpha^{h-j} \circ \epsilon_{T+j}|y_T)\\ &= \alpha^h \cdot y_T + \lambda \sum_{j=1}^{h} \alpha^{h-j} \cdot \theta_{s(T+j)}. \end{aligned} \tag{S3.9}$$

To use this predictor for our multivariate PoINAR(1) process, we need to produce estimators for the rates value, $\boldsymbol{\lambda}$, the seasonal effects values, $\boldsymbol{\theta}_s$, and the thinning value, $\boldsymbol{\alpha}$, for the L time series. To this end, we run the MCMC sampler for $m = 1000$ iterations and discard the first 100 of them as burn-in. We then thin every 5^{th} iteration which leaves us with 180 iterations from which to infer the parameters in our model. The one-step-ahead conditional expected value for the m^{th} iteration is:

$$\hat{y}_{T+1}|\lambda_i^{'(m)}, \alpha_i^{(m)}, \boldsymbol{\theta}^{(m)} = \alpha_i^{(m)} \cdot y_T + \lambda_i^{'(m)} \cdot \theta_{i,s(T+j)}^{(m)} \tag{S3.10}$$

For each time series (location), we now have samples from the posterior distribution of the conditional expected value which we average to produce the corresponding estimated value. We compare the performance of our method with two benchmark methods: the conditional least-squares (CLS) method and a simple Poisson process (SPP). Since the CLS method models the PoINAR(1) process for each time series separately, we estimate its parameters correspondingly. We then plug these estimators into (S3.9) to produce the corresponding one-step-ahead predicted value for each of the time series. For further details on the CLS method, the reader is referred to Section S2. The SPP assumes, for a single time series, the observed counts are independent identically distributed Poisson random variables with a constant rate value, λ. Therefore, we estimate λ for each time series using its corresponding counts average and then use this as the one-step-ahead predictor.

To evaluate the different methodologies we use root mean square error (RMSE) and absolute percentage error (APE) between the true population expected value and its corresponding estimated value based on the $L = 100$ time series. The results of this analysis are presented in Table 1. The analysis reveals that our method consistently yields more accurate results compared to the CLS method and the SPP. As expected, the better the separation between the cluster rate values, the easier it is for our method to estimate the parameters more accurately. In addition, generally higher autocorrelation values produce lower APE but higher RMSE. Intuitively because the stationary distribution mean value for the PoINAR(1) process is $\mathbb{E}(y) = \frac{\lambda}{1-\alpha}$, higher autocorrelation, α, yields a higher marginal mean value (or alternatively higher count values). This indicates a larger separation between the clusters counts values for the data sets with higher α. Therefore, higher autocorrelation helps our method identify the "true" clusters and yield more accurate estimators based on shrinking.

In conclusion, we believe that because the CLS and SPP methods consider each time series separately, they are more prone to over-fitting. The proposed Bayesian methodol-

Thin	0.1			0.5			0.9		
Rates	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard
SPP RMSE	0.477	0.113	0.005	1.674	0.880	0.293	6.128	1.155	0.552
CLS RMSE	0.306	0.080	0.035	0.284	0.114	0.057	0.343	0.118	0.055
BNP RMSE	**0.219**	**0.058**	**0.026**	**0.260**	**0.086**	**0.045**	**0.299**	**0.075**	**0.043**
SPP APE	0.067	0.159	0.1241	0.1958	0.3192	0.2013	0.1230	0.3372	0.2737
CLS APE	0.041	0.142	0.110	0.024	0.120	0.093	0.006	0.090	0.047
BNP APE	**0.033**	**0.041**	**0.072**	**0.019**	**0.033**	**0.044**	**0.005**	**0.046**	**0.022**
$E(y_{,T+1})$	5.383	1.001	0.317	9.861	1.848	0.591	52.161	9.908	3.0633

Table 1: Conditional mean estimation comparison between the CLS, SPP and our Bayesian nonparametric (BNP) method. The first four rows show the mean square error (MSE) and the absolute percentage error (APE) between the population (true) one-step-ahead conditional mean and its corresponding estimated value. The last row shows the average population (true) conditional expected value.

ogy allows the estimates to pool information from several time series resulting in more robust parameter estimates.

S4 Washington, D.C. population density map

Figure 4 shows the map of population density across the 188 Washington, D.C. census tracts.

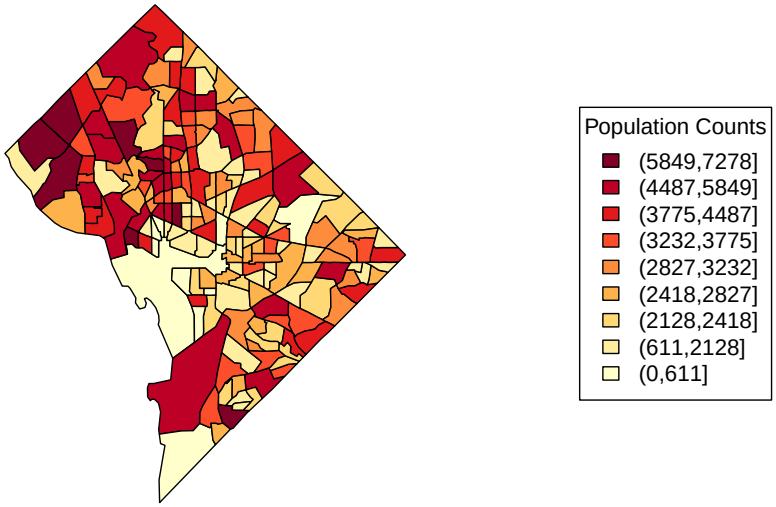

Figure 4: Washington, D.C. map of population density across the 188 census tracts.

S5 Washington D.C. crimes time series

Figure 5 shows the time series of weekly counts of violent crimes in four tracts between 2001 and 2008. We again see that some tracts can have very few occurrences whereas others have as many as 9 violent crimes per week. Also, since the counts are both discrete and small, it is hard to see clear seasonality within the weekly series.

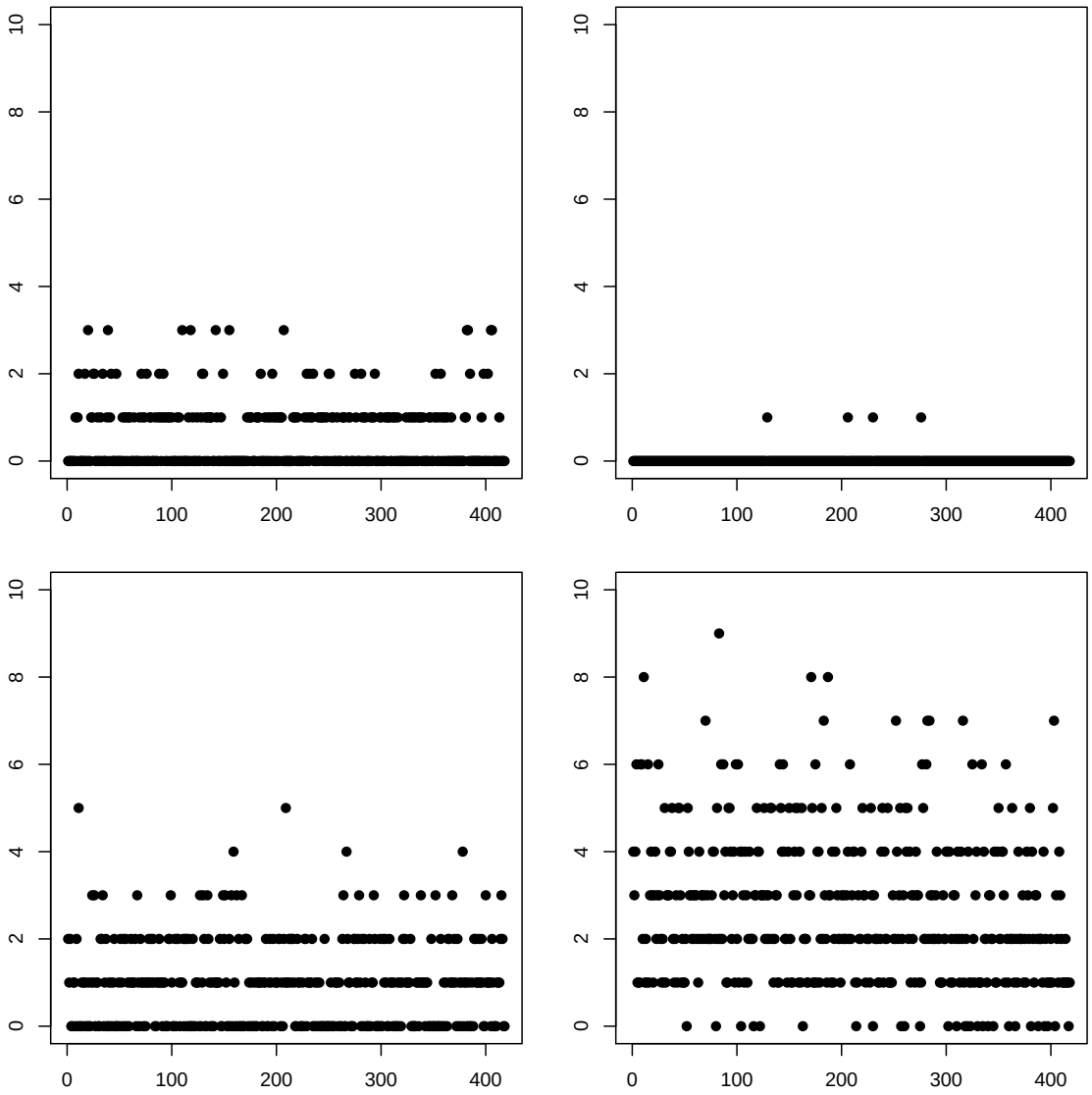

Figure 5: Weekly violent crime counts between 2001 and 2008 in 4 census tracts.

$y_{\cdot,T}$	0	1	2	3	4	Overall
SPP bias	0.0873	0.092	0.2080	0.4730	0.5308	0.126
	(0.0303)	(0.0458)	(0.095)	(0.1761)	(0.2434)	(0.0379)
CLS bias	-0.1625	-0.0834	**0.1310**	0.3287	0.3849	-0.073
	(0.0383)	(0.0441)	(0.0716)	(0.1276)	(0.3736)	(0.0405)
BNP bias	**-0.0234**	**0.0690**	0.2175	**0.2196**	**0.19262**	**0.0456**
	(0.0156)	(0.0393)	(0.0710)	(0.1143)	(0.3528)	(0.0232)
Frequency	0.5900	0.2340	0.1160	0.0400	0.0200	1

Table 2: One-step-ahead average bias as a function of the last observed value of $y_{\cdot,T}$.

S6 Bias analysis

In Table 2, we provide a summary of the average one-week-ahead bias for the Washington, D.C. crime data analysis. As we see from the results, in general, our method produces the smallest bias, but the differences between the methods are not significant except when the last observed count is zero.