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

Sivan Aldor-Noiman Lawrence D. Brown Emily B. Fox
Robert A. Stine

December 16, 2012

Abstract

To aid in the efficient tasking of police and other protective measures, there is significant interest in being able to predict regions in which crimes are likely to occur. Violent crimes often exhibit both temporal and spatial characteristics, though the spatial patterns do not vary smoothly across the map and instead we see spatially disjoint areas that exhibit similar crime behaviors. It is this indeterminate inter-region correlation structure along with the low-count discrete nature of the data that motivate our proposed forecasting tool. In particular, we propose to model the crime counts in each region using an integer-valued first order autoregressive process. We take a Bayesian nonparametric approach to flexibly discover a clustering of these region-specific time series. We also present methods for accounting for seasonality and covariates. We demonstrate our approach through an analysis of reported violent crime data in Washington D.C., collected between 2001-2008, and show that our forecasts outperform standard methods while additionally providing useful tools such as prediction intervals.

Keywords: Violet crime counts, Low-count time series, INAR, Bayesian nonparametric methods

1 Introduction

Violent crimes are a significant source of concern in major metropolitan areas across the United States. The impact of such violent crimes on the city are manyfold, ranging from effects on residents to tourism, and the ability to curb such crimes is of utmost

importance. In particular, there is significant interest in being able to predict regions in which crimes are likely to occur so that protective measures may be employed both in the short- and long-term. The violent crimes of interest often exhibit both temporal and spatial characteristics. From a temporal standpoint, violent crimes show seasonal behavior during a year. The rate of violent crimes rises during warmer months of the year [12]. From a spatial standpoint the behavior is more complex. One might expect that neighboring regions experience similar crime rates. However, there are often key geographic features that impede crime from varying smoothly across regions. Take for example, Washington D.C., which has Rock Creek Park located in the northeast corner of the city and the Anacostia River which runs along the southern part of the city. Both of these create natural obstacles for the spreading of crimes. Often such effects occur at a finer resolution (e.g. railroad tracks and highways). Such situations make it challenging to devise models that account for local spatial structure without over-smoothing between clearly distinct regions.

Of course, demographic characteristics also account for the variation in violent crime rates across a city. One measurement of such demographics, which we focus on in this paper, are the data collected by the Census Bureau. In particular, we consider the 188 census tracts in Washington D.C., a region that has consistently ranked among the top cities for rates of violent crimes in the United States. A census tract consists of a few street blocks and is designed to be homogeneous with respect to demographic features such as economic status and living conditions. Figure 1 (left) shows a Washington D.C. city map with the 188 census tracts boundaries layered on top. The spatial sizes of census tracts vary widely depending on the population density. Due to the demographic homogeneity of census tracks, it is not surprising that neighboring tracks can have quite different crime dynamics.

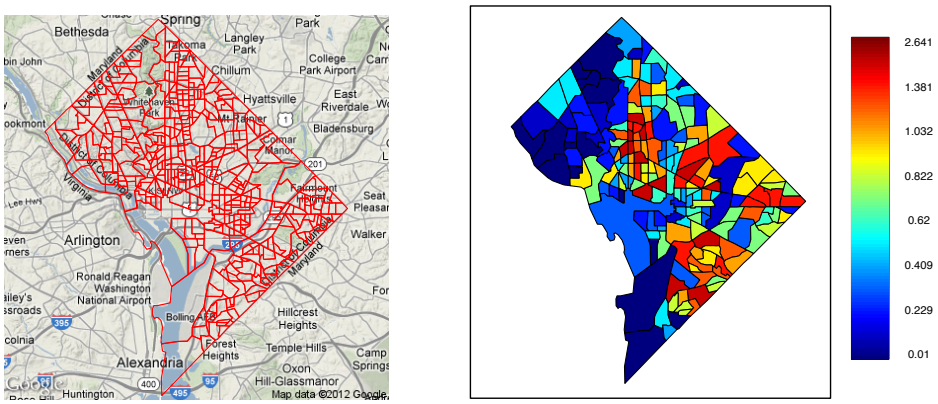

Figure 1: *Left:* Map of the 188 census tracts in Washington D.C.. *Right:* Weekly average violent crime counts across the 188 census tracts.

The combination of demographic features and natural formations contributes to the spatially diverse crime patterns across the census tracts. Figure 1 (right) shows the weekly average violent crime counts between 2001 and 2008. This plot reveals a few interesting features. First, the weekly average number of violent crimes are (fortunately) low. Second, tracts with similar average crime counts are often quite spatially disjoint. This suggests that crime, as hypothesized, does not vary smoothly across the map. When taking a closer look at the tracts over time, we see very different behaviors across the tracts. For example, Figure 2 shows four of the tracts' weekly violent crime counts that occurred between 2001 and 2008. We can see that some time series have very few occurrences while others have as many as 9 violent crimes per week. Also, since the counts are both discrete and low it is hard to see clear seasonality within the weekly series.

The features of these multiple time series, namely (i) low-count discrete data with (ii) an uncertain correlation structure, necessitate the development of new forecasting tools. In particular, we propose a novel methodology which models each time series as an integer-valued first-order autoregressive process (INAR(1)). An INAR(1) process

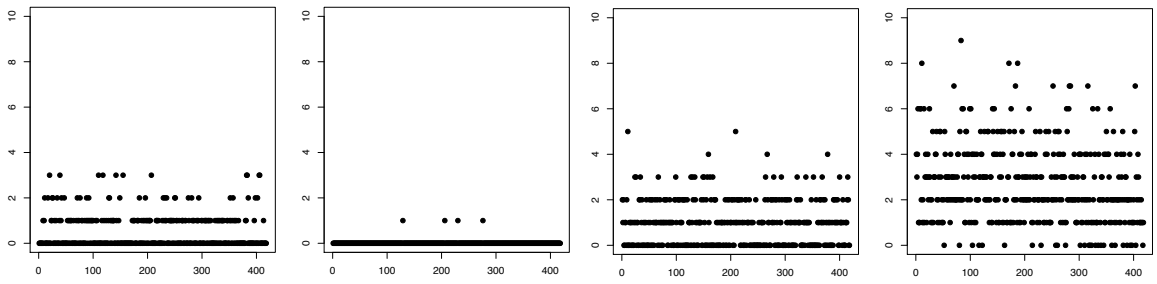

Figure 2: Weekly violent crime counts between 2001 and 2008 in 4 census tracts.

is a convolution between a binomial and a Poisson random variable [2, 13]. We induce correlation between the time series through the INAR(1) innovation processes. These innovation processes are decomposed into two latent factors: a common seasonal effect that is shared between the time series and a rate function which is tract specific. To cope with the high dimensionality of the data, we seek a method of multiple shrinkage. We take a Bayesian nonparametric approach and impose a Dirichlet process prior on the tract rates, which leads to a clustering of rates, thus efficiently sharing information between tracts in a flexible, data-driven manner.

We develop an efficient Markov chain Monte Carlo scheme and demonstrate that our proposed multivariate INAR(1) model produces accurate out-of-sample forecasts that outperform a conditional least-squares (CLS) counterpart both on simulations and the Washington D.C. crime data. One reasonable explanation as to why our model outperforms the CLS is because the number of clusters discovered is fairly small relative to the number of time series. Our model essentially shrinks all of the time series estimators in the same cluster toward a common mean yielding better out-of-sample forecasts. Another advantage, a byproduct of the Bayesian framework, is that we can provide the posterior distribution of the p-step-ahead forecasts. These distributions are important in the context of crime forecasting since crime distributions are

right-skewed and decision makers often care about preparing for the worst-case scenario. Producing the prediction intervals for violent crimes in each of the locations can help the police distinguish between an unusual rise in violent crimes that requires intervention and a random rise which is due to common variation in crime patterns.

Spatio-temporal count data applications present a natural candidate for multivariate Poisson-based models. For example, occurrences of earthquakes were modeled in [5] using a maximum likelihood approach to infer the model parameters of a multivariate INAR(1) process with Poisson innovations. (Note that the formulation does not maintain Poisson margins, as discussed in Section 2.) [17] employed Poisson processes to track the intensity of violent crimes in Cincinnati. The spatial Poisson rate is factored into a process density, modeled using Bayesian nonparametrics, and an overall intensity. Both were allowed to evolve in time. However, such a formulation assumes spatial smoothness to the crime rates. Additionally, the focus of [17] is on in-sample inference rather than on predicting future events. In contrast, our research focuses on providing methods to forecast multiple integer-valued low-count time series. We harness the efficient and elegant structure of INAR(1) processes and present a method for modeling multiple, correlated time series while maintaining Poisson margins. The correlations are induced via a Bayesian nonparametric clustering of the time series, and in doing so, we efficiently share information to produce better out-of-sample predictions. Bayesian nonparametric methods have previously been studied as tools for data-driven clustering analysis (cf., [19, 7, 9]). However, these studies either focused on clustering continuous-valued time series or Poisson counts which have no time component.

2 INAR(1) Background

We begin by introducing notation and background for the necessary building blocks to describe our proposed data generating process. Throughout, let $Y_{l,t}$ denote the number of violent crimes at location $l = 1, \ldots, L$ during week $t = 1, \ldots, T$.

2.1 Univariate INAR(1)

A univariate INAR(1) model is defined as follows [2]:

$$Y_{l,t+1} = \alpha \circ Y_{l,t} + \epsilon_{l,t+1} \quad \text{for } t = 0, \pm 1, \pm 2, \ldots. \tag{1}$$

For any nonnegative integer-valued random variable $Y_{l,t}$ and for any $\alpha \in [0, 1]$, the random variable $\alpha \circ Y_{l,t}$ is the result of a binomial thinning operator:

$$\alpha \circ Y_{l,t} = \sum_{i=1}^{Y_{l,t}} B_i(\alpha), \tag{2}$$

where $B_i(\alpha)$ are independent identically distributed Bernoulli random variables with success probability α. There is a limited class of initial distributions for $Y_{l,0}$ and innovation distributions for $\epsilon_{l,t}$ that yield a strongly stationary process [16]. One example is the Poisson distribution [21]. We refer to the resulting model as a *univariate PoINAR(1)*.

2.2 Multivariate INAR(1)

We proceed by defining $\boldsymbol{Y}_t := (Y_{1,t}, \ldots, Y_{L,t})$ as the multivariate violent crime counts during time t at locations $l = 1, \ldots, L$ or alternatively one can view $\boldsymbol{Y}_t$ as the map

of crime counts at time t. To model $\boldsymbol{Y}_t$ as a multivariate INAR(1) process, we define the multivariate binomial thinning operator at time t, $\boldsymbol{\alpha} \circ \boldsymbol{Y}_t$ where $\boldsymbol{\alpha}$ denotes the $L \times L$ dimensional thinning matrix with entries in $[0, 1]$. The result of this operator is a L-dimensional random vector, with i^{th} component

$$[\boldsymbol{\alpha} \circ \boldsymbol{Y}_t]_i := \sum_{l=1}^{L} \alpha_{i,l} \circ Y_{l,t}, \tag{3}$$

where $\alpha_{i,l} \circ Y_{l,t}$ is the binomial thinning operator as previously defined in Eq. (2). The individual binomial thinning operators are assumed to be independent of each other, i.e. $\alpha_{i,l} \circ Y_{l,t} \perp \alpha_{j,m} \circ Y_{m,\tau}$ for all i, j, l, m, t, τ.

In the multivariate setting, it is more complicated to define a process that maintains Poisson margins. Let $\boldsymbol{\epsilon}_t := (\epsilon_{1,t}, \epsilon_{2,t}, \ldots, \epsilon_{L,t})$ denote the innovations vector at time t. Simply taking $\epsilon_{l,t}$ to be independently Poisson distributed with rate function $\Lambda_{l,t}$, and likewise assuming $Y_{l,0}$ is Poisson distributed, does not ensure that $Y_{l,t}$ is marginally Poisson distributed. Actually, it is straightforward to prove that when the off-diagonal elements of the thinning matrix are non-zero, a stationary distribution exists but is no longer the Poisson distribution (see McKenzie [13], Pedeli and Karliss [14]). Such a multivariate INAR(1) was considered in [5].

3 Multivariate PoINAR(1)

In this section we propose a method of modeling multivariate INAR processes that maintain Poisson margins. We first introduce the basic model and then demonstrate how to induce inter-correlations by placing a Dirichlet process prior on the rate parameters of the Poisson innovation processes. We conclude this section by highlighting the

similarities and differences between the proposed model and the vector autoregressive process, which is the equivalent model with Gaussian margins.

3.1 A multivariate PoINAR(1) process

For parsimony and to maintain Poisson margins, we assume that the $\boldsymbol{\alpha}$ thinning matrix is a diagonal matrix, i.e. $\alpha_{i,j} = 0$ for $i \neq j$:

$$\begin{pmatrix} Y_{1,t+1} \\ Y_{2,t+1} \\ \vdots \\ Y_{L,t+1} \end{pmatrix} = \begin{pmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \alpha_L \end{pmatrix} \circ \begin{pmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{L,t} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \vdots \\ \epsilon_{L,t} \end{pmatrix}. \tag{4}$$

For notational convenience we will denote the diagonal elements by $\alpha_i := \alpha_{i,i}$. A diagonal thinning matrix implies that at time t, the l^{th} entry of the thinned random vector is only a function of the i^{th} location itself, i.e.

$$[\boldsymbol{\alpha} \circ \mathbf{Y}_t]_l = \alpha_l \circ Y_{l,t}. \tag{5}$$

For the innovations processes, we assume $\epsilon_{l,t}|\Lambda_{l,t} \sim \text{Poiss}(\Lambda_{l,t})$. That is, the innovations are independently Poisson distributed across time and space. The resulting multivariate INAR(1) yields a stationary Poisson distribution for each element in $\mathbf{Y}_t$. We refer to this process as the *multivariate PoINAR(1)*. We emphasize that the diagonal thinning matrix not only dramatically reduces the number of model parameters, but also enables defining such a stationary process with Poisson univariate margins.

Conditioning on the set of rate parameters $\{\Lambda_{l,t}\}$ yields L independent time series. Motivated by the structure and dimensionality of the crime data, we would like our

model to capture dependencies between the time series as well. In the next section, we show how we introduce such dependencies by using a Dirichlet process mixture model for the innovations processes.

3.2 Capturing dependencies

There are several ways to induce dependencies between the dimensions of the multivariate PoINAR(1) model. As previously mentioned, there are two sources of variation in the model, the multivariate binomial thinning operators and the innovation processes. We propose to generate the inter-correlations through the innovation processes and assume that the binomial thinning operators are independent across the time series (dimensions). Such a formulation shares information between tracts while allowing tract-dependent autocorrelations. Furthermore, focusing on the innovations processes allows for computational efficiencies, as described in Section 4.

Recall that the innovations are assumed to follow a Poisson distribution with a latent rate function, $\Lambda_{l,t}$. The rate parameter is a function of both the corresponding location l and the time period t of the specific innovation term, $\epsilon_{l,t}$. We decompose the rate function into two elements:

- Spatial component - The tract-specific rate value, λ_l.
- Temporal component - A seasonal monthly effect, $\theta_{s(t)}$, that is spatially homogenous. Here, $s(t)$ is a function that maps time t to its associated month.

The two components are multiplied together to create the rate values of the innovations processes at each location l during time t, i.e. $\Lambda_{l,t} = \lambda_l \cdot \theta_{s(t)}$. The resulting

model for the innovations can be written as follows:

$$\epsilon_{l,t} \sim \text{Poiss}(\lambda_l \cdot \theta_{s(t)}). \tag{6}$$

Since the temporal component is shared across the different time series (locations) it induces some dependence. However, the seasonal effect alone may not account for all the inter-location dependencies. Consequently, we impose a Dirichlet process (DP) prior on the tract-specific rates, λ_l. The Dirichlet process, denoted by $\text{DP}(\tau, G_0)$, provides a distribution over countably infinite probability measures

$$\text{G} = \sum_{k=1}^{\infty} \beta_k \cdot \delta_{\phi_k} \qquad \phi_k \sim G_0 \tag{7}$$

on a parameter space Θ. The weights are sampled via a *stick-breaking* construction Sethuraman [15]:

$$\beta_k = \nu_k \prod_{l=1}^{k-1}(1-\nu_l) \qquad \nu_k \sim \text{Beta}(1,\tau). \tag{8}$$

In effect, we have divided a unit-length stick into lengths given by the weights β_k: the k^{th} weight is a random proportion ν_k of the remaining stick after the first $(k-1)$ weights have been chosen. We denote this distribution by $\beta \sim$GEM(τ). The DP has proven useful in many applications due to its clustering properties (cf., Teh et al. [19]), which are clearly seen by examining the *predictive distribution* of draws $\lambda_l \sim G$. Because probability measures drawn from a DP are discrete, there is a strictly positive probability of multiple observations λ_l taking identical values within the set $\{\phi_k\}$, with ϕ_k defined as in Eq. (7). For each value λ_l, let z_l be the cluster membership indicator, i.e. z_i is a categorical random variable that identifies the unique value ϕ_k such that

$\lambda_l = \phi_{z_l}$. The predictive distribution on the membership variables can be written in the following manner:

$$P(Z_{N+1} = z | z_1, \ldots, z_N, \tau) = \frac{\tau}{N+\tau}\delta(z, K+1) + \frac{1}{N+\tau}\sum_{k=1}^{K} n_k \delta(z,k), \tag{9}$$

where the discrete Kronecker delta $\delta(z,k) = 1$ if $z = k$, and 0 otherwise. Also, n_k indicates the number of members belonging to the k^{th} group, i.e. $\sum_{i=1}^{N}\delta(z_i,k)$. $K+1$ is a new group which currently has no members. The distribution on partitions induced by the sequence of conditional distributions in Eq. (9) is commonly referred to as the *Chinese restaurant process* (CRP). Take l to be a customer entering a restaurant with infinitely many tables, each serving a unique dish ϕ_k. Each arriving customer chooses a table, indicated by z_l, in proportion to how many customers are currently sitting at that table. With a probability proportional to τ, the customer starts a new, previously unoccupied table $K+1$. From the CRP, we see that the DP has a reinforcement property that leads to clustering. It can also be shown that the expected number of clusters using the CRP is $O(\tau \cdot \log(L))$ where L is the number of observations (see Teh [18] for a detailed proof). This implies that on average the number of clusters is much smaller than the number of observations, L.

As previously mentioned, we impose a DP prior on the L tract-specific rates, λ_l. Note that here the number of observations from the DP is equal to the number of tracts, rather than the number of time points. The DP prior groups the time series according to their corresponding tract-specific rates into a few clusters. Members of the same cluster k share the same tract rate, ϕ_k. The grouping of the multiple time series into a small number of clusters provides a useful shrinkage tool that helps pool information across members of the same cluster, therefore yielding more accurate out-

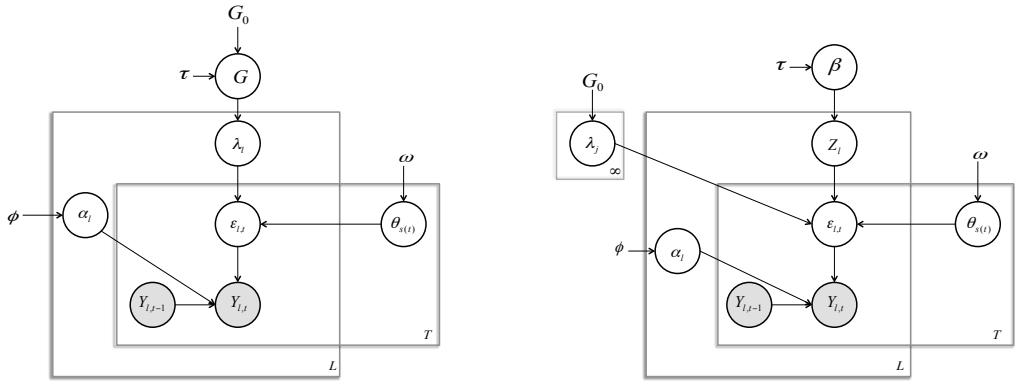

Figure 3: Graphical model of the multivariate dependent PoINAR(1) model. *Left:* The Dirichlet process-based innovations generating process of Eq. (11). *Right:* An equivalent representation using cluster indicator variables $z_1, \ldots, z_L$ as in Eq. (12).

of-sample predictions for the multiple time series. When we combine the tract-specific rates and the seasonal effects we get the following innovations generating process:

$$\begin{aligned} \epsilon_{l,t} &\sim \text{Poiss}(\lambda_l \cdot \theta_{s(t)}) \quad l = 1, \ldots, L \quad t = 1, \ldots, T \\ \theta_m &\sim \text{F}(\omega) \quad m = 1, \ldots, 12 \\ \lambda_l &\sim \text{G} \quad l = 1, \ldots, L \\ \text{G} &\sim \text{DP}(\tau, G_0). \end{aligned} \tag{10}$$

Figure 3 (left) shows a graphical representation of our dependent multivariate PoINAR(1) process. Alternatively, we can use an equivalent representation using the GEM distribution and the membership labels $z_1, \ldots, z_L$ (see Figure 3 (right)):

$$\begin{aligned} \epsilon_{l,t} &\sim \text{Poiss}(\phi_{z_l} \cdot \theta_{s(t)}) \quad l = 1, \ldots, L \quad t = 1, \ldots, T \\ \theta_m &\sim \text{F}(\omega) \quad m = 1, \ldots, 12 \\ z_l &\sim \text{Multi}(\beta) \quad l = 1, \ldots, L \\ \phi_j &\sim G_0 \qquad \beta \sim \text{GEM}(\tau) \quad j = 1, 2, \ldots. \end{aligned} \tag{11}$$

3.3 Prior specification

The dependent multiple PoINAR(1) requires estimation of three main components:

- thinning values,$(\alpha_1, \ldots, \alpha_L)$, for each of the time series.
- monthly seasonal effects, $(\theta_1, \ldots, \theta_{12})$.
- specific location rates, $[\lambda_1, \ldots, \lambda_L]$.

The Bayesian framework requires specification of prior distributions for these three elements. We choose priors which are both computationally convenient and weakly-informative. For the thinning values and monthly seasonal effects we specify:

$$\begin{aligned} \alpha_l &\overset{\text{iid}}{\sim} \text{Beta}(\eta_1, \eta_2) \quad \text{for } l = 1, \ldots, L \\ \theta_m &\overset{\text{iid}}{\sim} \text{Gamma}(\xi_1, \xi_2) \quad \text{for } m = 1, \ldots, 12. \end{aligned} \tag{12}$$

We have also explored the half-normal distribution as a prior for the seasonal effect and the analysis did not reveal any significant changes from the results presented in Section 6. The DP requires the specification of the base measure, G_0, and the concentration parameter, τ. We choose the base measure to be the Gamma(γ_1, γ_2) distribution, which is well suited not only because it is conjugate to the Poisson distribution but also because it has an easy interpretation in our context. Mainly, we have a prior belief that the violent crime weekly rates have low positive values with a few locations that might have a higher rate. Hence, a gamma distribution with shape and scale parameters $\gamma_1 = 1$ and $\gamma_2 = 0.1$ reflects these prior beliefs. For the concentration parameter, we specify $\tau \sim \text{Gamma}(a_\tau, b_\tau)$, as suggested by Escobar and West [8].

3.4 Relationship to VAR processes

The first-order vector autoregressive process (VAR(1)) is well-studied (cf., Tsay [20]) and often used to capture multivariate continuous-valued time-series. The VAR(1) model can be written in the following manner:

$$\begin{aligned} \boldsymbol{Y}_{t+1} &= \boldsymbol{\alpha} \cdot \boldsymbol{Y}_t + \boldsymbol{\epsilon}_t \quad t = 1, \ldots, T \\ \boldsymbol{\epsilon}_t | \Sigma &\overset{\text{iid}}{\sim} N(0, \Sigma) \\ Y_{l,0} | \mu_{l,0}, \sigma_{l,0} &\overset{\text{iid}}{\sim} N(\mu_{l,0}, \sigma_{l,0}) \quad l = 1, \ldots, L. \end{aligned} \tag{13}$$

Compare to the multivariate PoINAR(1):

$$\begin{aligned} \boldsymbol{Y}_{t+1} &= \boldsymbol{\alpha} \circ \boldsymbol{Y}_t + \boldsymbol{\epsilon}_t \quad t = 1, \ldots, T \\ \boldsymbol{\epsilon}_t | \Lambda_t &\overset{\text{iid}}{\sim} [\text{Poiss}(\Lambda_{1,t}), \text{Poiss}(\Lambda_{2,t}), \ldots, \text{Poiss}(\Lambda_{L,t})] \\ Y_{l,0} | \Lambda_{l,0} &\overset{\text{iid}}{\sim} \text{Poiss}(\Lambda_{l,0}) \quad l = 1, \ldots, L. \end{aligned} \tag{14}$$

These two models resemble each other not only notation-wise, but also in terms of some common characteristics:

- The marginal distribution of $Y_{l,t}$ – The PoINAR(1) has Poisson margins while the VAR(1) has Gaussian margins. If the parameters defining the processes are chosen carefully, these models can be shown to have a Poisson/Gaussian stationary distribution, respectively.

- The autoregressive parameter $\alpha_{l,l}$ – The parameter $\alpha_{l,l}$ plays the role of the first-order autocorrelation coefficient in both models, i.e. $\text{corr}(Y_{l,t+1}, Y_{l,t}) = \alpha_{l,l}$.

The similarities with the continuous VAR(1) make the discrete PoINAR(1) espe-

cially attractive and easy to interpret. However, the VAR(1) process has a single source of variation – the innovations process – while the PoINAR(1) process has two – the binomial thinning and innovations processes. This key difference will prove a complicating factor in inference for the PoINAR(1) model, which we address in Section 4.

4 The MCMC Sampler

As previously noted, the PoINAR(1) model is a combination of two latent processes: the binomial thinning process and the innovations process. Each of these processes has its own set of model parameters: the binomial thinning uses the thinning parameters $\{\alpha_l\}$ while the innovations process is a function of both the rates $\{\phi_k\}$ and the seasonal effects $\{\theta_m\}$. For posterior computations within our Bayesian framework, we employ an MCMC sampler. Intuitively, the idea is to sample a posterior latent innovations sequence and then condition on this sequence to sample both the latent DP clustering of census tracts and also the thinning parameters and seasonal effects. In contrast, in the corresponding Gaussian VAR(1) model there is no need to sample the innovations sequence since they are a deterministic function of the observations and the model parameters. Therefore, one would expect the multivariate PoINAR(1) model to be more computationally cumbersome compared to its VAR(1) counterpart. However, our proposed sampler harnesses computational advantages both from the low counts nature of the data as well as the fact that the clustering only relies on the sums of counts which are the Poisson sufficient statistic. We outline the resulting sampler below. For detailed derivations, see Section 1 of the Supplementary Material.

1. Sample the innovations, $\boldsymbol{\epsilon} := [\epsilon_1, \ldots, \epsilon_L]$ where $\epsilon_l := [\epsilon_{l,1}, \ldots, \epsilon_{l,T}]$ is the innova-

tions series for the l^{th} location. The posterior factorizes as:

$$P(\boldsymbol{\epsilon}|\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\theta}) \;=\; \prod_{l=1}^{L}\prod_{t=2}^{T} P(\epsilon_{l,t}|Y_{l,t-1}, Y_{l,t}, \alpha_l, \lambda_l, \boldsymbol{\theta}). \tag{15}$$

Given the observations $\boldsymbol{Y}$ and the multivariate PoINAR process parameters, the innovations can be sampled independently for each location and each time point. The possible values satisfy $\max\{0, Y_{l,t} - Y_{l,t-1}\} \leq \epsilon_{l,t} \leq Y_{l,t}$ with corresponding probabilities

$$P(\epsilon_{l,t}|Y_{l,t-1}, Y_{l,t}, \alpha_l, \lambda_l, \boldsymbol{\theta}) \propto \frac{1}{\epsilon_{l,t}!(Y_{l,t} - \epsilon_{l,t})!(Y_{l,t-1} - (Y_{l,t} - \epsilon_{l,t}))!} \left(\frac{\lambda_l \theta_{s(t)} \cdot (1 - \alpha_l)}{\alpha_l}\right)^{\epsilon_{l,t}}. \tag{16}$$

Although this is not a well-known discrete distribution, it is analytically tractable due to the low count nature of the data (i.e., $\max\{0, Y_{l,t} - Y_{l,t-1}\} \leq \epsilon_{l,t} \leq Y_{l,t}$ and $Y_{l,t}$ is assumed to be small. In the crime data the maximum value is 11.) Another important consideration that reduces the computational burden is that certain $\epsilon_{l,t}$ values can be deterministically set from the observations vector $\boldsymbol{Y}_t$: if $y_{l,t}$ is zero then $\epsilon_{l,t}$ must also be zero and if $y_{l,t-1}$ is zero then $\epsilon_{l,t}$ must equal the value of $y_{l,t}$. Since our crime data has many zero counts, these constraints substantially lower the computational cost of this sampling stage. If larger counts are observed, then one can use a Metropolis-Hastings step to sample from this distribution with a Poisson proposal distribution. Importantly, note that if the observed counts are large enough, a Gaussian approximation to the Poisson distribution can be used implying that a VAR(1) might be a good alternative to the PoINAR(1).

2. Sample the membership indicator vector, $\boldsymbol{z} := [z_1, \ldots, z_L]$. We harness the DP-induced Chinese restaurant process (CRP) and iteratively sample tract-specific cluster indicators:

$$P(z_l = k | \boldsymbol{z}_{/l}, \boldsymbol{\epsilon}, \Theta, \gamma_1, \gamma_2) \propto \begin{cases} \tau \cdot p_{l,0} & \text{for } k = K+1 \\ n_j \cdot p_{l,j} & \text{for } k = j \quad \text{where } j = 1, \ldots, K, \end{cases} \tag{17}$$

where $K+1$ identifies a previously unseen cluster, $\Theta = \sum_{t=1}^{T} \theta_{s(t)}$ and $\boldsymbol{z}_{/l}$ is the membership indicator vector not including the l^{th} term. The first terms, (τ, n_j), of Eq. (17) arise from the CRP prior of Eq. (9) and the exchangeability of the process such that each z_l can be treated as the last. The second terms, $(p_{l,0}, p_{l,j})$, correspond to the likelihood of the innovations $\boldsymbol{\epsilon}$ given the cluster assignments $(z_l = k, \boldsymbol{z}_{/l})$ and seasonal effects Θ, marginalizing the cluster-specific rates ϕ_k. The terms are given by the following negative binomial probability distribution functions:

$$\begin{aligned} p_{l,0} &= \frac{\Gamma(S_l + \gamma_1)}{\Gamma(\gamma_1) S_l!} \left(\frac{\gamma_2}{\Theta + \gamma_2}\right)^{\gamma_1} \left(\frac{\Theta}{\Theta + \gamma_2}\right)^{S_l} \\ p_{l,j} &= \frac{\Gamma(S_l + A_j + \gamma_1)}{\Gamma(A_j + \gamma_1) S_l!} \left(1 - \frac{\Theta}{n_j \cdot \Theta + \gamma_2}\right)^{A_j + \gamma_1} \left(\frac{\Theta}{n_j \cdot \Theta + \gamma_2}\right)^{S_l}, \end{aligned} \tag{18}$$

where $S_l = \sum_{t=1}^{T} \epsilon_{l,t}$ and $A_j = \sum_{i: z_i = j, i \neq l} S_i$. Note that $p_{l,0}$ and $p_{l,j}$ only rely on various sums of the innovations and the sum of seasonal effects. We also highlight that the conditional conjugacy of our formulation allows us to use the collapsed sampler of Eq. (17) for the z_l, marginalizing $\{\phi_k\}$.

3. Sample unique rates, ϕ_k. Although the unique rates are collapsed away in sampling the cluster indicators, z_l, they are relied upon in sampling the innovations

sequence (Step 1) and seasonal effects (Step 3). As such, we instantiate the unique rates as auxiliary variables for these steps, and then discard them. For each currently instantiated cluster, sample ϕ_k as:

$$\phi_k|\boldsymbol{\epsilon}, \boldsymbol{z}, \Theta, \gamma_1, \gamma_2 \;\sim\; \text{Gamma}(B_k + \gamma_1, n_k \cdot \Theta + \gamma_2) \tag{19}$$

where $B_k = \sum_{l \in \{v: z_v = k\}} S_l$. Again, we only rely on the innovations sum, S_l, to compute the posterior distribution.

4. Sample the seasonal effects vector, $[\theta_1, \ldots, \theta_{12}]$. The m^{th} element of this vector can be sampled as:

$$\theta_m|\boldsymbol{\epsilon}, \boldsymbol{\phi}, \xi_1, \xi_2 \;\sim\; \text{Gamma}\left(\sum_{l=1}^{L} \sum_{t:s(t)=m} \epsilon_{l,t} + \xi_1, q_m \cdot \sum_{l=1}^{L} \lambda_l + \xi_2\right), \tag{20}$$

where q_m counts the number cycles for the m^{th} month. Notice that for this step we need to sum the innovations over locations rather than time.

5. Sample the thinning values vector, $[\alpha_1, \ldots, \alpha_L]$. For location l,

$$\alpha_l|\boldsymbol{\epsilon}_l, \boldsymbol{Y}_l, \eta_1, \eta_2 \;\sim\; \text{Beta}(\sum_{t=2}^{T} Y_{l,t} - S_l + \eta_1, \sum_{t=2}^{T} (Y_{l,t-1} - Y_{l,t}) + S_l + \eta_2), \tag{21}$$

where S_l is defined as in Step 2.

6. Sample the concentration parameter, τ, for the Dirichlet process prior according to Escobar and West [8].

It is important to note that if the model did not include seasonal effects, then one could simply sample the sum of the innovations, S_l, instead of the vector of

innovations, ϵ_l. This would reduce the computational cost of the sampler since Step 1 is the most time consuming.

5 Simulation Examples

In order to assess the performance of our model, we simulate 9 different datasets from the multivariate PoINAR(1) process of Eq. (4). Each dataset has $L = 100$ dimensions (locations) with $T = 208$ observations. The multiple time series are grouped into four equally sized clusters defined by a shared rate value. The different data sets vary in:

- The levels of separation between the cluster rates, 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 rate values are well separated in the easy setting and become harder to distinguish as we move to the hard setting.

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

We evaluate the performance of our MCMC sampler both in- and out-of-sample. 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 simulation results show that our model can reasonably recover the ground-truth clusterings and also produce accurate out-of-sample forecasts under various settings. Table 1 presents the RMSE performance of our model compared to a simple Poisson process model

(SPP) and the conditional least-squares model (CLS), which are detailed in the Supplementary Material (see Section 2). This analysis shows that our model outperforms both of these models. As expected, the larger the separation between the cluster rate values, the easier it is for our method to estimate the parameters accurately. Also, higher autocorrelation helps our method identify the "true" clusters and yields more accurate estimators based on shrinking.

	Thin=0.1			Thin=0.5			Thin=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
MC RMSE	**0.219**	**0.058**	**0.026**	**0.260**	**0.086**	**0.045**	**0.299**	**0.075**	**0.043**
$E(Y_{\cdot,T+1})$	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 MCMC methods. The first four rows show the mean square error (MSE). The last row shows the average population (true) conditional expected value.

Another setting that we examine is one where there is a single cluster; that is, all of the time series are generated from a single process. As one would hope, under these conditions the sampler recovers a single cluster, further validating the methodology. For a more detailed description of the simulations and the results, see Section 3 of the Supplementary Material.

6 Violent Crime Data Analysis

In this section we examine both in- and out-of-sample results based on the Washington D.C. violent crime data described in Section 1. The data has $L = 188$ time series (census tracts) and contains $T = 418$ weeks of data between 2001-2008. We use the first 7 years worth of data to train our model and the last 52 weeks to evaluate out-of-sample forecasts. We ran 5 MCMC chains for 5000 iterations from different initial

values, each drawn from the following priors:

$$\begin{aligned} \theta_m &\sim \text{Gamma}(1,1) \quad \alpha_l \sim \text{Beta}(1,1) \\ \tau &\sim \text{Gamma}(2,4) \quad \phi_i \sim \text{Gamma}(1,1) \end{aligned} \tag{22}$$

We performed a sensitivity analysis for the hyperparameters during the simulation stage, but found no significant changes to the results. We discard the first 1000 iterations as burn-in and then thin the remaining 4000 samples by 50. Therefore, our inference for each of the model's parameters is based on the resulting $80 \cdot 5 = 400$ MCMC samples. We use the scale reduction factor (recommended by [11]) to monitor convergence across the chains.

We begin by looking at the distribution of the number of clusters over the 400 iterations in Figure 4. The mode is 17 clusters, which is a substantial reduction from the original $L = 188$ time series. A representative cluster assignment is presented in Figure 5 along with the posterior rate values for this assignment. The representative sample is selected as the cluster assignment that has the minimum average Hamming distance across the different iterations. See [10] for further details. An interesting phenomenon is that census tracts assigned to the same cluster are not spatially contiguous in general.

We further examine the posterior means for the rates, λ_l, of the 188 census tracts and their corresponding thinning values, α_l, across the MCMC samples. Figure 6 (left) shows the map of Washington D.C. census tract posterior mean rate values. We can see three hot-spot areas which have higher mean values corresponding to the southern part of Washington D.C., along 16th Street and Rhode Island Avenue. These areas exhibit high concentration of poverty, as noted in [1]. The results of Figure 6

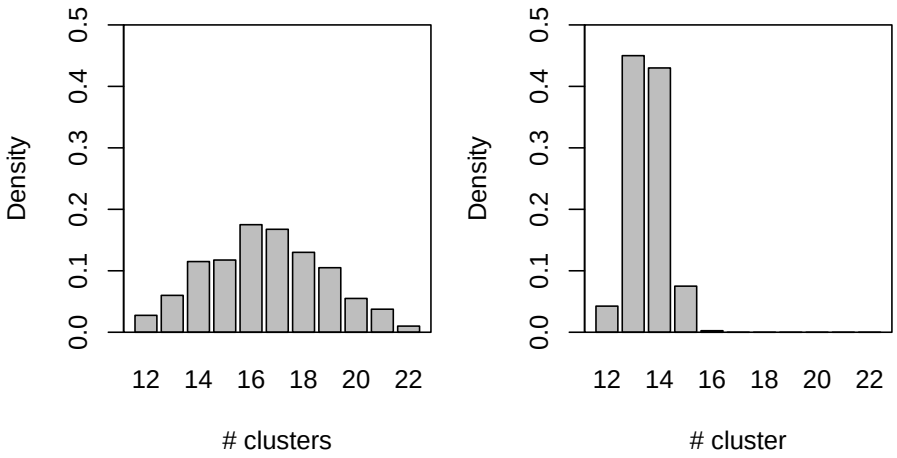

Figure 4: Histograms of the posterior number of clusters for the multivariate dependent PoINAR(1) described in Section 3.2 (left) and the population adjusted multivariate dependent PoINAR(1) model described in Section 7 (right).

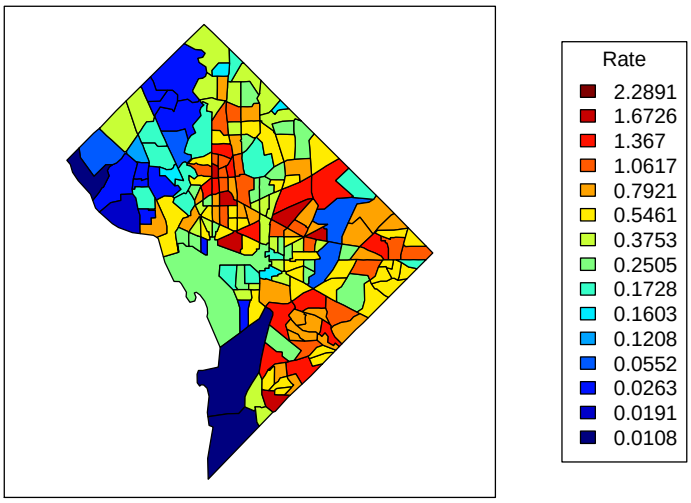

Figure 5: The minimum average Hamming distance cluster assignment along with the corresponding posterior rate values.

are also substantiated by our exploratory data analysis. In particular, compare to Figure 1 (right).

Figure 7 compares, for each census tract, the raw data autocorrelation values with the posterior mean thinning values. The raw data autocorrelation is calculated using the classical first order autocorrelation estimator for each time series separately (without adjusting for seasonality). As previously explained, the thinning values in our model represent the autocorrelation values for the INAR(1) time series. The comparison shows that the raw data autocorrelations vary on a wider range of values than their corresponding posterior mean values and some of these raw autocorrelations have small negative values. There are two reasons that can account for the differences between the two:

1. Our model only allows the thinning value to range between $[0, 1]$ and therefore cannot account for negative autocorrelation. We believe that the (small) negative raw autocorrelation values are probably due to noise variation and therefore we are less concerned about this phenomenon.

2. The posterior mean thinning values are adjusted for seasonal effects. Since the raw autocorrelations are not very large in magnitude after adjusting for the seasonality, we should expect them to be even smaller as these results indicate.

For the purpose of the out-of-sample evaluation we compare our method to the CLS and SPP. For both the CLS and our method, we use the estimated one-week-ahead conditional mean as the predictor for the crime counts in each locations l:

$$\hat{y}_{l,T+1} \quad = \quad \alpha_l \cdot y_{l,T} + \lambda_l \cdot \theta_{s(T+1)}. \tag{23}$$

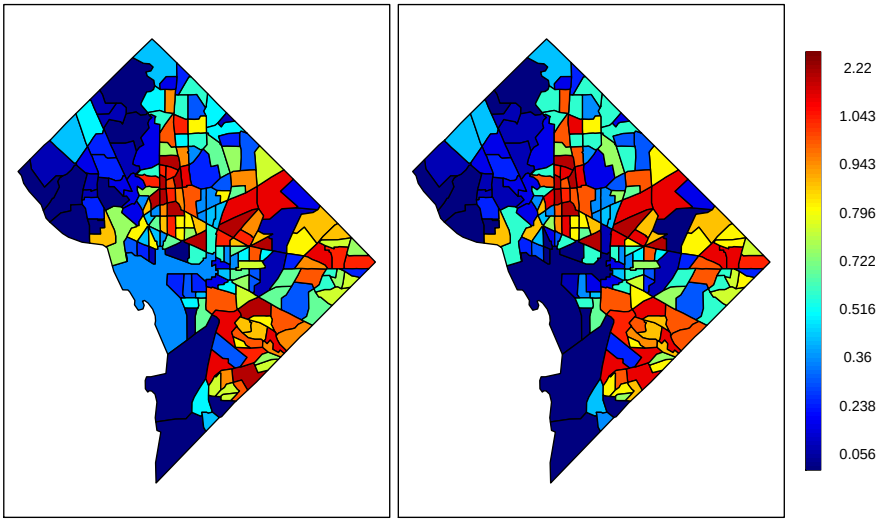

Figure 6: Map of posterior mean rates, λ_i, sampled from the multivariate dependent PoINAR(1) model described in Section 3.2 (left) and the population adjusted multivariate dependent PoINAR(1) model described in Section 7 (right).

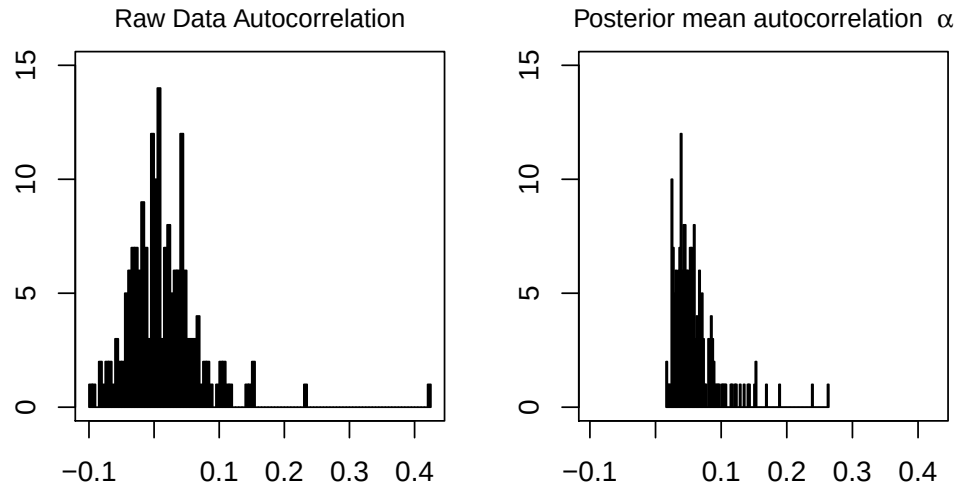

Figure 7: Histogram of raw data autocorrelations (left) and posterior mean autocorrelations α_i (right).

For the CLS method, we simply plug-in the estimates of α_l, λ_l and θ_m for each location. For our method, we compute an MCMC-based estimate by evaluating Eq. (23) for each of the 400 MCMC iterations and using the average of these values as the final predicted value (see Section 3.2 of the Supplementary Material for for further details). For the SPP, we simply use the average of past values as the predictor. We predict the one-week-ahead crime counts in each of the locations at the beginning of each month during 2008. All together there are 12 such time points. Table 2 shows the one-week-ahead predicted mean RMSE and corresponding standard errors as a function of the last observed value. The results indicate that for the most frequent values (0,1,2), our method produces lower RMSE. For the less frequent, higher values (3,4), the performances of all of the methods are (statistically) equivalent. This behavior is to be expected since our method shrinks the estimators toward the mean and therefore should perform better for the lower more frequent values and worse in the rarer cases. A summary of the average one-week-ahead bias value as a function of the last observed value is presented in Section 4 of the Supplementary Material. In general, our method produces the smallest bias values, but the differences between the methods are not significant except for the zero value.

The one-step-ahead conditional mean value is the best linear unbiased estimator under a quadratic loss function. Since the CLS method assumes a quadratic loss function, it is only natural to evaluate all three methods using the same loss function. However, Berk [3] proposed using a quantile loss function to reflect the sensitivity of the police department to forecasting errors. Under a τ-quantile loss function the predictor is just the predictive distribution's τ^{th} quantile. Using our method, one can easily sample from the following one-step-ahead predictive posterior distribution and

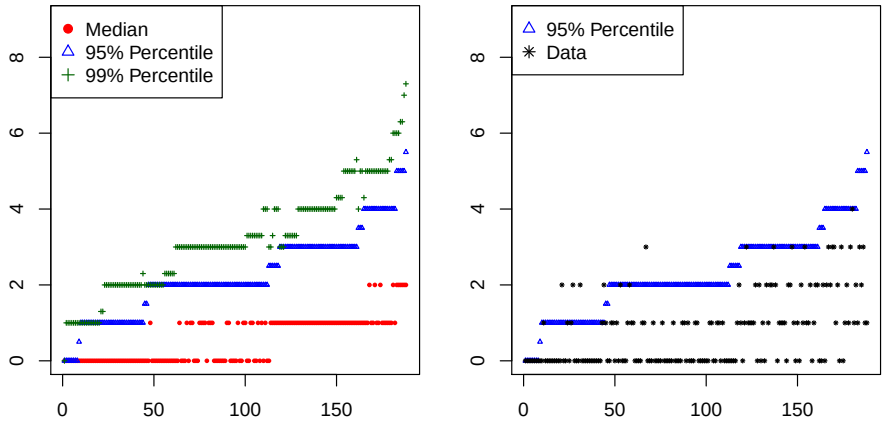

Figure 8: The predictive posterior distribution for each of the 188 locations. The red dots correspond to the median predicted number of violent crimes for each location. The blue triangles and green crosses correspond to the 95% and 99% percentiles of the predictive posterior distribution. The black stars corresponds to the test-set actual observed value of crimes.

evaluate any desired quantile:

$$P(Y_{l,t+1} = y_{l,t+1} | Y_{l,t} = y_{l,t}, \alpha_l^{(m)}, \boldsymbol{\theta}^{(m)}, \lambda_l^{(m)}) =$$
$$\sum_{r=0}^{\infty} \binom{y_{l,t-1}}{y_{l,t} - r} (\alpha_l^{(m)})^{y_{l,t}-r} \cdot (1 - \alpha_l^{(m)})^{y_{l,t-1}-(y_{l,t}-r)} \frac{e^{-\phi_l^{(m)} \cdot \theta_{s(t+1)}^{(m)}} \cdot (\phi_l^{(m)} \cdot \theta_{s(t+1)}^{(m)})^r}{r!}. \quad (24)$$

where $\lambda_l^{(m)} = \phi_{z_l^{(m)}}^{(m)}$, $\boldsymbol{\theta}^{(m)}$ and $\alpha_l^{(m)}$ are the rate, seasonal component and thinning value estimated during the m^{th} iteration of the MCMC sampler. Figure 8 shows the 95% and 99% quantiles for each of the 188 locations and the corresponding ones-step-ahead $y_{l,T+1}$ true value. The quantiles may also be used to provide prediction intervals for each of the locations. The police department can use these intervals along with the point estimate of the predicted value to distinguish between an unusual surge in crimes which requires allocation of more resources, and a small insignificant rise in crimes, which probably does not require any intervention.

$y_{\cdot,T}$	0	1	2	3	4	Overall
SPP RMSE	0.8373	0.966	1.1829	1.4722	1.4252	0.970
	(0.034)	(0.0311)	(0.0453)	(0.088)	(0.1631)	(0.0167)
CLS RMSE	0.7729	0.9501	1.0605	1.1370	1.3258	0.9235
	(0.0245)	(0.0430)	(0.0660)	(0.0982)	(0.1991)	(0.0368)
MCMC RMSE	**0.7222**	**0.9172**	**1.009**	**1.0225**	**1.1600**	**0.72168**
	(0.0135)	(0.0172)	(0.4336)	(0.0862)	(0.1782)	(0.0016)
Frequency	0.5900	0.2340	0.1160	0.0400	0.0200	1

Table 2: One-step-ahead average RMSE as a function of the last observed value of $y_{\cdot,T}$. We also provide the standard errors associated with the average RMSE.

7 Covariates Adjusted Dependent PoINAR(1)

As previous research suggests, crime behavior is correlated with demographic covariates. The census bureau provides various demographic indicators on each of the census tracts, and there are different methods by which to incorporate these covariates in our model. For example, we could model the tract-specific rate as a regression on these covariates and cluster the coefficients of the regression. The grouping of the coefficients may provide further insight into the relationships between crime forecasting and the demographic indicators and improve the accuracy of our predictions. Alternatively, we could incorporate the demographic covariates directly into the clustering mechanism as proposed in Blei and Frazier [4]. This might improve our model's forecasting ability, but it will not allow us to examine the impact of the covariates on the crime rates, making it harder to interpret. In the next section we look at the former method for incorporating covariates.

7.1 Adjusting for population

The main goal of this section is to demonstrate how to add covariates to our model and to explore the benefit of doing so. To this end, we look at the population sizes

in each of the census tracts as a possible explanatory variable.

Let X_l denote the population size of the l^{th} census tract based on the data collected by the 2000 census. We provide a map of the Washington D.C. population density in Section 5 of the Supplementary Material. To incorporate this variable in our model, we can simply redefine the tract-specific rate as a linear function of the population size, i.e. $\lambda_l = X_l \cdot \psi_l$, where ψ_l is the number of violent crimes per person in the l^{th} tract. We then place a DP prior directly on the rate per person parameter, ψ_l, yielding the following model:

$$\begin{aligned} \epsilon_{l,t} &\sim \text{Poiss}(X_l \cdot \psi_l \cdot \theta_{s(t)}) \quad l = 1, \ldots, L \quad t = 1, \ldots, T \\ \theta_m &\sim \text{F}(\omega) \quad m = 1, \ldots, 12 \\ \psi_l &\sim \text{G} \quad l = 1, \ldots, L \\ \text{G} &\sim \text{DP}(\tau, G_0). \end{aligned} \tag{25}$$

It is straightforward to adjust the MCMC sampler described in Section 4 to account for the population covariate, X_l. We change the base measure G_0 to Gamma(0.5,0.5) to reflect the adjustment for population sizes (and still remain weakly informative). After these few alterations, we run the sampler in a similar manner to the previously described sampler of Section 4.

7.2 Results analysis

Using the covariate-adjusted PoINAR(1) of Section 7.1, we again analyze the Washington D.C. crime data. Similar to the results presented in Section 6, we begin by looking at the posterior distribution of the number of clusters over the 400 MCMC

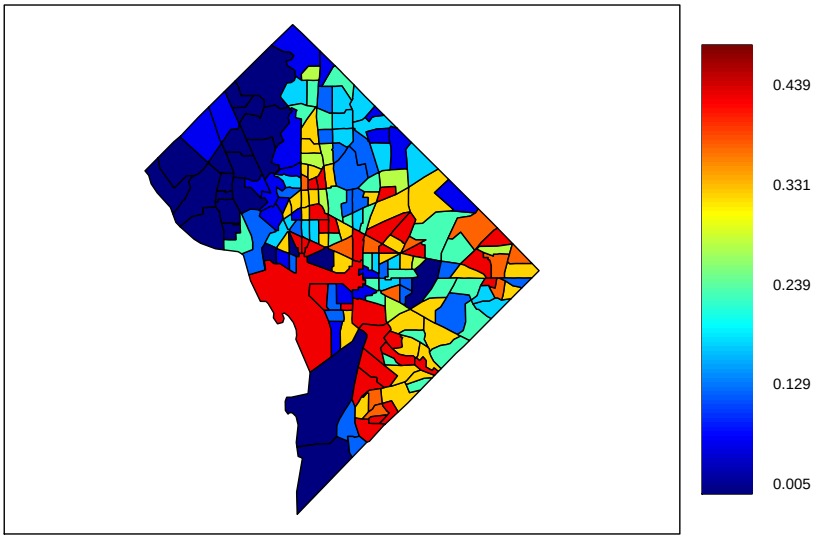

Figure 9: Map of the posterior mean values for crime rates per person, ψ_i.

iterations (again taken from 5 chains each run 5000 iterations). Figure 4 (right) indicates that the distribution is much narrower when we adjust for the population density, and has a mode of 14 clusters. Figure 9 shows the map of posterior mean values for crime rates per person, ψ_l. This map highlights two main features:

1. The city center has a high violent crime count per person.

2. The north-western area of the city has a very low violent crime count per person.

3. There are three hot-spots: in the city center and in the eastern and south-western parts of the city.

These results match the conclusions reported in [6] and the first two insights were harder to detect from the analysis using the unadjusted multivariate PoINAR(1) model (see Figure 6). .

In Table 3 we compare the one-week-ahead forecasting performance of the unadjusted model presented in Section 3.2 and the population adjusted model. Based on

this analysis, the two models yield similar results, with the unadjusted model performing mildly better. Even though adding the census tract population size covariate does not seem to improve predictive accuracy, the method still provides a useful tool since it reveals a more interpretable grouping of the the time series which better match domain knowledge (as reported in [6]). Of course, there are many more covariates that might help prediction accuracy such as poverty indices, median housing prices and median age for each census tract. We leave it for future research to explore the potential benefit of adding these covariates. In such cases, another interesting question is whether to consider a global clustering informed by all covariates or various collections of clustering informed by individual covariates.

$y_{\cdot,T}$	0	1	2	3	4	Overall
Unadj. RMSE	0.6956	0.8018	0.8218	1.7346	7.2442	0.9663
Pop.-adj. RMSE	0.6949	0.8246	0.8278	1.7354	7.2468	0.9713
Frequency	0.5691	0.2340	0.1489	0.0426	0.0053	1

Table 3: One-step-ahead MCMC-based estimate of RMSE comparing the unadjusted and population-adjusted multivariate PoINAR(1) models as a function of the last observed value of $y_{\cdot,T}$.

8 Discussion

In this paper we presented a method of forecasting multiple correlated low-count time series building on the univariate PoINAR framework. The model induces correlation between the different time series through two sources: an overall temporal seasonal effect and a clustering on individual rate parameters. The latter clustering is induced by a Dirichlet process, which encourages sparse representations in terms of a small number of clusters. The grouping of the different rates allows our inference method to pool strength across the different time series, shrinking the estimators to provide

better out-of-sample forecasts.

Our model assumes that there is some underlying clustering assignment of the multiple time series. Moreover, once these clusters are decided they remain constant throughout time. We may relax this assumption and allow for temporally evolving cluster assignments. There are a few ways to create such a mechanism, for example we might impose dependent Dirichlet process priors, such as in [17].

Finally, although our focus here was on violent crime data, this model is broadly applicable to many low-count spatio-temporal data sets, including the number of insurance claims across the U.S., earthquakes across the globe [5], wildfire across counties [22], and so forth.

References

[1] Wikipedia page on crimes in Washington D.C.

[2] Alzaid, A. and AlOsh, M. (1988). Firs Order IntegerValued autoregressive (INAR (1)) process: Distributional and regression properties. *Statistica Neerlandica*, 42(1):53–61.

[3] Berk, R. (2008). Forecasting methods in crime and justice. *Annual Review of Law and Social Science*, 4(1):219–238.

[4] Blei, D. M. and Frazier, P. I. (2009). Distance dependent chinese restaurant processes. *arXiv:0910.1022*.

[5] Boudreault, M. and Charpentier, A. (2011). Multivariate integer-valued autoregressive models applied to earthquake counts.

[6] Cahill, M. and Roman, J. K. (2010). Small number of blocks account for lots of crime in D.C.

[7] Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L., and Jordan, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics*, 64(2):635–644.

[8] Escobar, M. D. and West, M. (1994). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Assocaition*, 90:577—588.

[9] Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2010). Bayesian nonparametric methods for learning Markov switching processes. *IEEE Signal Processing Magazine*, 27(6):43–54.

[10] Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5(2):1020–1056.

[11] Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

[12] McDowall, D., Loftin, C., and Pate, M. (2011). Seasonal cycles in crime, and their variability. *Journal of Quantitative Criminology*.

[13] McKenzie, E. (2000). Discrete variate time series. *Simulation*, 21(August):1–34.

[14] Pedeli, X. and Karliss, D. (2011). A bivariate INAR(1) process with application. *Statistical Modelling*, 11(4):325 –349.

[15] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650.

[16] Steutel, F. W. and Harn, K. V. (1986). Discrete operator-selfdecomposabiuty and queueing networks. *Communications in Statistics. Stochastic Models*, 2(2):161–169.

[17] Taddy, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *Journal of the American Statistical Association*, 105(492):1403–1417.

[18] Teh, Y. W. (2011). Dirichlet process. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning*, pages 280–287. Springer US.

[19] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

[20] Tsay, R. S. (2001). *Analysis of Financial Time Series*. Wiley-Interscience, 1 edition.

[21] Wolpert, R. L. and Brown, L. D. (2011). Stationary infinitely-divisible Markov processes with non-negative integer values.

[22] Xu, H. (2011). Point process modeling of wildfire hazard in los angeles county, california. *The Annals of Applied Statistics*, 5(2):684–704.