Statistica Sinica **26** (2016), 1587-1610
doi:http://dx.doi.org/10.5705/ss.2014.217t

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 and* [3]*University of Washington*

Abstract: There is significant interest in being able to predict where crimes will happen, for example to aid in the efficient tasking of police and other protective measures. We aim to model both the temporal and spatial dependencies often exhibited by violent crimes in order to make such predictions. The temporal variation of crimes typically follows patterns familiar in time series analysis, but the spatial patterns are irregular and do not vary smoothly across the area. Instead we find that spatially disjoint regions exhibit correlated crime patterns. It is this indeterminate inter-region correlation structure along with the discrete nature of small counts of serious crimes that motivates 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 clusters of region-specific time series. We then describe how to account for covariates within this framework. Both approaches adjust for seasonality. We demonstrate our approach through an analysis of weekly reported violent crimes in Washington, D.C. between 2001-2008. Our forecasts outperform those of standard methods while additionally providing information from the posterior distribution of forecasts, such as prediction intervals.

Key words and phrases: Bayesian nonparametric methods, INAR, low-count time series, violent crime counts.

1. Introduction

Largely driven by reasons of computational tractability, a significant effort in time series modeling has historically focused on Gaussian-based models including the classical autoregressive (AR) and autoregressive moving average (ARMA) processes, the latter of which can be equated with linear, state-space models. With the advent of computational methods for simulation, interest has extended beyond this class of models to capture various empirical features such as heavy-tailed distributions or non-linear dynamics, though typically with a focus on continuous-valued observations. A wide variety of applications, however, concern counts, as in studies of infectious disease, sales and marketing, and demography. For large counts, log-Gaussian transformations are often appropriate, enabling classical time series models to once again be employed. Increasingly, we face

numerous collections of small counts produced by the ability to collect and record data at fine scales. Instead of aggregate statistics at the level of a company or region, interest lies in individual- or local-level inferences. Two important questions arise in such situations: (i) how should the dynamics of small counts be modeled, and (ii) how do we appropriately share information among related series when each individually provides only limited data? Echoing major facets of the work of George Tiao (such as Tiao and Box (1981), Hillmer and Tiao (1982), and Box and Tiao (2011)), we develop a Bayesian model for multiple time series that can capture seasonal and exogenous factors.

We take motivation in particular from a dataset of weekly violent crime counts in Washington D.C. recorded at the census-tract level. Police forces have significant interest in being able to predict regions in which crimes are likely to occur so that preventive measures may be employed in both the short- and long-term. Based on our focus on violent crimes occurring within a small region (census tract), the weekly counts are (fortunately) small, as illustrated in Figures 1 and 2. In such cases, for a given census tract one might propose log-Gaussian Cox models with a Poisson observation model and latent Gaussian process intensity. Such a formulation is problematic in our situation. For instance, such a model would not present an explicit recursive structure unless the latent Gaussian process were confined to the special case of a Gaussian autoregression. More relevant to our application, these models do not maintain Poisson margins, a characteristic of our dataset (see Figure 2. To capture this aspect of the crime counts, we instead employ an approach more akin to standard AR models, namely Poisson, integer-valued AR (PoINAR) processes (Alzaid and Al-Osh (1988), McKenzie (2000)). Such processes yield Poisson margins and a simple recursive structure. This family has received relatively little attention in time series modeling and a significant focus has been on univariate modeling.

Returning to our goal of modeling large collections of relatively short, low-count time series, we are in a "large p, small n" scenario and seek a method of sharing information among series. Although our counts of violent crimes are indexed spatially, the trends do not vary smoothly across the region. Although one might expect neighboring regions to experience similar crime rates, both geographic features (e.g., Rock Creek Park in the northeast and the Anacostia River in the south of Washington, D.C.) and transportation systems (railroad tracks and highways) impede the spread of crime. Likewise, the combination of demographic homogeneity within census tracts with heterogeneity among tracts implies that neighboring tracts often have rather different crime dynamics. This spatial heterogeneity is clear in the map of Figure 1 (right). The combination of these attributes makes it challenging to borrow strength among adjacent regions without over-smoothing. To avoid this problem, we consider the series as an

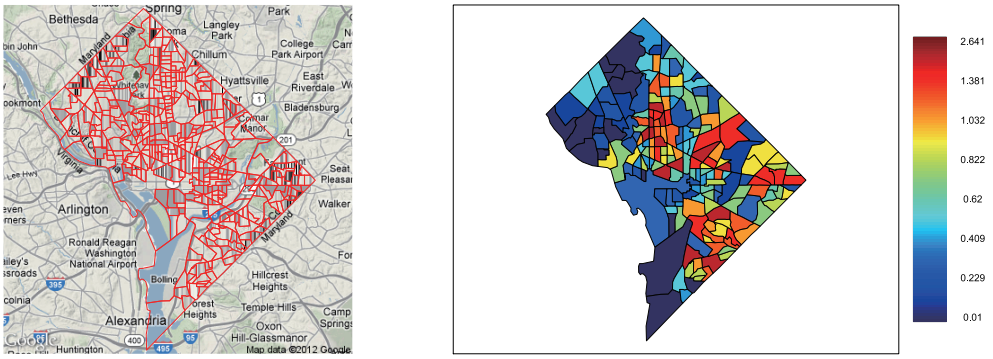

Figure 1. *Left:* Map of the 188 census tracts in Washington, D.C.. *Right:* Weekly average violent crime counts across the 188 census tracts.

indexed collection and ignore the explicit spatial structure. Our goal becomes developing a type of Bayesian multiple shrinkage to cope with the large number of series. To this end, we induce correlation between the spatial time series through the innovations of multiple, series-specific PoINAR processes. We obtain multiple shrinkage adaptively by using a Bayesian nonparametric approach that imposes a Dirichlet process prior on the series-specific innovation rates. The Dirichlet process leads to a clustering of rates and thus efficient sharing of the information among the spatial time series in a flexible, data-driven manner. Our model essentially shrinks the estimators for the time series that share a rate toward a common mean, thereby yielding better out-of-sample forecasts.

We develop an efficient Markov chain Monte Carlo (MCMC) scheme for fitting these models. The results on both simulated data and counts of violent crimes in Washington, D.C. demonstrate that our proposed Bayesian multiple PoINAR model produces out-of-sample forecasts that are more accurate than a model treating each series independently via a conditional least squares (CLS) PoINAR fit. These results can be attributed to the fact that our approach discovers a small number of clusters relative to the number of census tracts. This demonstrates the importance of multiple shrinkage to modeling large collections of low-count time series. Another advantage, a byproduct of the Bayesian framework, is that our model provides posterior distributions of the p-step-ahead forecasts. These distributions are important in the context of forecasting crime because the distribution of crime is right-skewed and decision makers often care about preparing for worst-case scenarios. Prediction intervals for the number of violent crimes in each region can help the police distinguish between an unusual rise in violent crimes that requires intervention and a rise which is due to random variation. Such trends are common in many applications, including e-commerce sales.

Our paper is structured as follows. In Section 3, we provide background on univariate PoINAR processes. We then present our method for correlating multiple PoINAR processes while maintaining Poisson margins in Section 4. The associated posterior computations via MCMC are detailed in Section 5. In Section 6, we briefly summarize a simulation and in Section 7, we analyze the crime data of interest. Finally, in Section 8 we describe a method for accounting for covariates, and in particular consider population size as a predictor.

2. Modeling Violent Crime

A s an example of our methodology, we consider rates of violent crimes in the 188 census tracts in Washington, D.C.. The nation's capital has consistently ranked among the top cities for rates of violent crimes in the United States. Violent crimes in our data consist of rape, robbery, arson and aggravated assault. Along with murder, these types of crimes define the FBI part 1 violent crimes list. Part 1 crimes are considered serious and are directly reported to the police (as opposed to other law agencies such as the IRS). Indeed, the Washington, D.C. police department keeps a record of all reported type 1 violent crimes and makes it publicly available through their website (`http://crimemap.dc.gov/CrimeMapSearch.aspx`.

We begin with a purely geographic approach that is aligned with police procedures and work solely with raw crime counts within census tracts. We then examine how to account for covariates, in particular population size, in Section 8.

Figure 1 (left) shows a map of Washington, D.C. with boundaries of the census tracts superimposed. A census tract consists of adjacent street blocks selected to be homogeneous with respect to demographic features such as economic status and living conditions. The sizes of the census tracts vary widely depending on population density. According to the 2,000 Census, tracts in Washington, D.C. average 3,043 residents, ranging from 149 to 7,278.

The variation of the counts of violent crimes within a tract is well approximated by a Poisson model. Figure 2 (top-left) shows the frequencies (on a log scale) of all weekly counts, combined over tracts and years. Most often, no violent crime happens in a tract; more than half of the 78,208 weekly counts are zero, and zero is the most common count for 90 of the 188 tracts. As an example of the preponderance of zeros, Figure 2 (top-right) shows the sequence of weekly counts for the tract with the median rate of violent crimes. (Section 6 of the Supplementary Material displays the counts for other individual tracts.) Note that the near-linear decay of the log frequencies in Figure 2 (top-right) is not typical of a Poisson distribution, but recall that these frequencies mix counts from tracts with very low rates with counts from tracts with higher rates. To motivate our use of Poisson models for the very small counts in individual

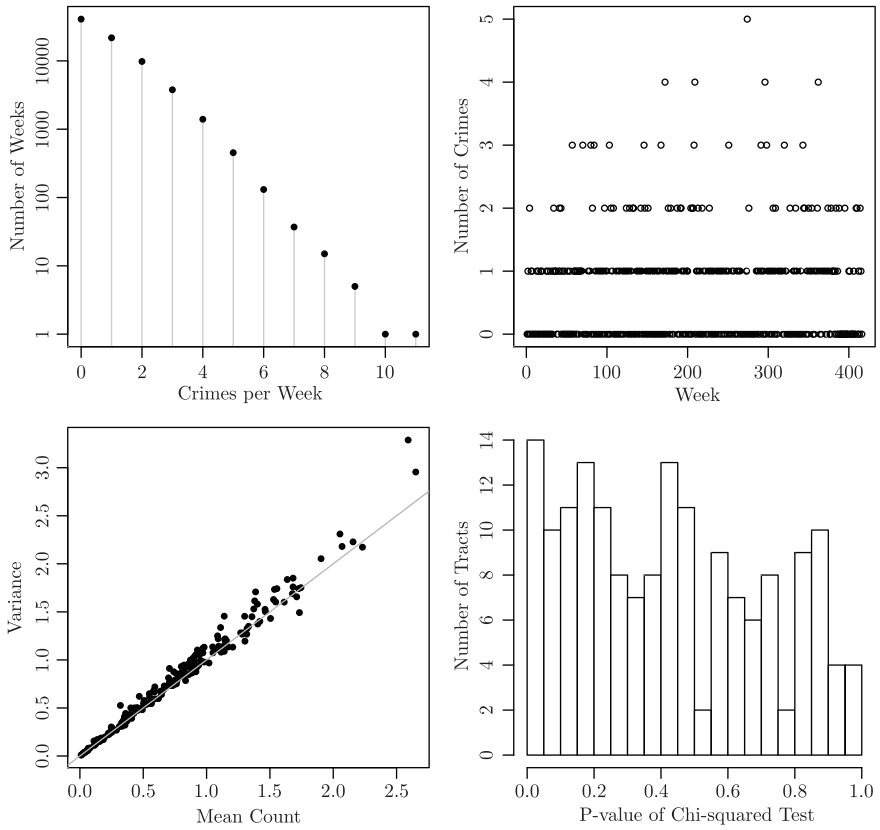

Figure 2. *Top Left:* Histogram of weekly counts of violent crime in the 188 census tracts of Washington, D.C. *Top Right:* Counts of violent crimes in the tract with median rate. *Bottom left:* Weekly variance versus mean for the tracts. *Bottom right:* Histogram of p-values of chi-squared test of Poisson variation.

tracts, Figure 2 (bottom-left) plots the variances of the weekly counts by tract versus the corresponding mean, with a diagonal line for comparison. The data are slightly over-dispersed, but close to matching this characteristic of Poisson variation. Within tracts, we also computed the p-values of the chi-squared test of goodness-of-fit for tract-level Poisson models. For each tract, we compared the frequency of counts to the expected counts from a Poisson model with mean set to the tract average. The histogram of these p-values shown in Figure 2 (bottom-right) is roughly uniform. (We excluded 21 tracts whose frequencies were too small for this test.)

2.1. Related approaches

Multivariate Poisson-based models are a natural match to multivariate time

series of counts, and this structure has been employed in various prior applications. For example, Boudreault and Charpentier (2011) model the occurrences of earthquakes using a maximum likelihood approach to infer the parameters of a multivariate INAR(1) process with Poisson innovations. (This formulation does not maintain Poisson margins, as discussed in Section 3. Taddy (2010) employed Poisson processes to track the intensity of violent crimes in Cincinnati and treats these as point processes. Taddy (2010) factors the spatial Poisson rate into a process density, modeled using Bayesian nonparametrics, and an overall intensity. Both were allowed to evolve in time. Such a formulation, however, assumes spatial smoothness of the crime rates. Additionally, Taddy (2010) focuses on in-sample inference rather than predicting future events. In contrast, our research focuses on models for areal data and provides methods to forecast these as 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 more accurate out-of-sample predictions. Bayesian nonparametric methods have previously been studied as tools for data-driven clustering analysis (cf., Teh et al. (2006), Dorazio et al. (2008), Fox et al. (2010)). These studies, however, focus on clustering either continuous-valued time series or Poisson counts which have no time component.

3. Univariate PoINAR(1) Background

A univariate PoINAR(1) model is defined as follows (Alzaid and Al-Osh (1988)):

$$Y_{t+1} = \alpha \circ Y_t + \epsilon_{t+1} \quad \text{for} \quad t = 0, 1, 2, \dots , \tag{3.1}$$

where the innovations $\{\epsilon_t\}$ are iid Poisson. The operator $\circ$ denotes binomial thinning. For any nonnegative integer-valued random variable X and for any $\alpha \in [0, 1]$, the random variable $\alpha \circ X$ is defined

$$\alpha \circ X = \sum_{i=1}^{X} B_i(\alpha), \tag{3.2}$$

where $B_i(\alpha)$ are independent, identically distributed Bernoulli random variables with success probability α. Given a Poisson distribution on the initial state Y_0 with finite mean $\mu = E\, Y_0$ and independence between Y_t and $\epsilon_t \sim \text{Poisson}((1 - \alpha)\mu)$, the construction (3.1) yields a strongly stationary process. In essence, to obtain a stationary Poisson marginal distribution from (3.1), the innovations must also be Poisson (Alzaid and Al-Osh (1988), Steutel and Van Harn (1986), Wolpert and Brown (2011)).

4. Multivariate PoINAR(1)

In this section we define a multivariate INAR process that retains Poisson margins. We first introduce the basic model and then demonstrate how to induce correlations among the component series by placing a Dirichlet process prior on the rate parameters of the Poisson innovations. We conclude this section by highlighting the similarities and differences between the proposed model and the vector autoregressive process, which is the corresponding model with Gaussian margins.

Throughout, let $Y_{l,t}$ denote the number of violent crimes at tract $l = 1, \ldots, L$ during week $t = 1, \ldots, T$. Furthermore, let $\boldsymbol{Y}_t := (Y_{1,t}, \ldots, Y_{L,t})$ denote a vector of crime counts at time t and $\boldsymbol{\epsilon}_t := (\epsilon_{1,t}, \epsilon_{2,t}, \ldots, \epsilon_{L,t})$ the vector of innovations.

4.1. The multiple PoINAR(1) process

One might imagine employing a multivariate PoINAR(1) analogue of a vector autoregressive process by considering:

$$\boldsymbol{Y}_{t+1} = \boldsymbol{\alpha} \circ \boldsymbol{Y}_t + \boldsymbol{\epsilon}_{t+1} \ ,$$

where $\boldsymbol{\alpha}$ is an $L \times L$ matrix with entries $0 \leq \alpha_{i,l} \leq 1$, and $\boldsymbol{\alpha} \circ \boldsymbol{Y}_t$ is defined as:

$$[\boldsymbol{\alpha} \circ \boldsymbol{Y}]_{i,t} := \sum_{l=1}^{L} \alpha_{i,l} \circ Y_{l,t}, \tag{4.1}$$

with $\alpha_{i,l} \circ Y_{l,t}$ defined by the binomial thinning operator defined in (3.2). Even in the simplest scenario of $\epsilon_{i,t}$ being independent Poisson innovations, however, the resulting margins are in general not Poisson. In fact, it is straightforward to prove that when the off-diagonal elements of the thinning matrix $\boldsymbol{\alpha}$ are non-zero, a stationary distribution exists but is no longer the Poisson distribution McKenzie (2000), Pedeli and Karliss (2011)). Such a multivariate INAR(1) was considered in Boudreault and Charpentier (2011).

The one scenario that preserves Poisson margins occurs if $\boldsymbol{\alpha}$ is diagonal. The model is

$$\begin{pmatrix} Y_{1,t+1} \\ Y_{2,t+1} \\ \vdots \\ Y_{L,t+1} \end{pmatrix} = \begin{pmatrix} \alpha_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & \alpha_L \end{pmatrix} \circ \begin{pmatrix} Y_{1,t} \\ Y_{2,t} \\ \vdots \\ Y_{L,t} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t+1} \\ \epsilon_{2,t+1} \\ \vdots \\ \epsilon_{L,t+1} \end{pmatrix}. \tag{4.2}$$

We refer to this process as the *multiple PoINAR(1)*. For notational convenience we denote the diagonal elements by $\alpha_l := \alpha_{l,l}$. The diagonal thinning matrix implies that at time t, the lth entry of the thinned random vector is only a function of the lth location:

$$[\boldsymbol{\alpha} \circ \mathbf{Y}_t]_l = \alpha_l \circ Y_{l,t}. \tag{4.3}$$

For the innovations processes, we assume $\epsilon_{l,t}|\Lambda_{l,t} \sim \text{Poisson}(\Lambda_{l,t})$. Conditional on the rate parameters $\Lambda_{l,t}$, the innovations are independently Poisson distributed across time and space. The resulting multiple INAR(1) yields a marginal Poisson distribution for each element in $\mathbf{Y}_t$. We emphasize that restricting $\boldsymbol{\alpha}$ to a diagonal thinning matrix not only dramatically reduces the number of model parameters, but also produces a process with Poisson margins.

Conditioning on the rate parameters $\{\Lambda_{l,t}\}$ yields L independent time series. To allow such models to capture dependence among the time series, we introduce a Dirichlet process mixture model for the innovations.

4.2. Capturing dependence

There are several ways to induce dependence among the elements of the multiple PoINAR(1) process. The model has two sources of variation: the multivariate binomial thinning operator and the innovation process. We propose to generate the dependence through the innovations and assume that the binomial thinning operators are independent across the time series to maintain Poisson margins. This formulation shares information between tracts while allowing tract-dependent autocorrelations. Furthermore, this focus on the innovations provides computational efficiencies as described in Section 5.

The innovations are assumed to follow a Poisson distribution with rates $\Lambda_{l,t}$. The rate is a function of both the location l and the time period t of the specific innovation $\epsilon_{l,t}$. We decompose the rate $\Lambda_{l,t}$ into a product of spatial and temporal components:

$$\Lambda_{l,t} = \lambda_l \, \theta_{s(t)}, \tag{4.4}$$

where the summands are a location-specific rate, λ_l, and a seasonal monthly rate, $\theta_{s(t)}$, that is spatially homogeneous. Crime rates often vary seasonally, with a higher rate during warmer months of the year (McDowall, Loftin and Pate (2012)). Here, $s(t)$ is a function that maps week t to its associated month. That is, we assume a constant seasonal effect within months and model this effect with parameters $\theta_1, \ldots, \theta_{12}$. The resulting model for the innovations can be written as follows:

$$\epsilon_{l,t} \sim \text{Poisson}(\lambda_l \, \theta_{s(t)}). \tag{4.5}$$

The temporal component induces some dependence across tracts because it is shared across the different time series. A Dirichlet process (DP) prior on the rates, λ_l, provides the balance of the dependence across tracts, but in a spatially heterogeneous manner by inducing a clustering on the rate parameters, as detailed below.

The Dirichlet process, denoted $\mathrm{DP}(\tau, G_0)$, provides a distribution over probability measures with a countably infinite number of atoms. Here, G_0 denotes a base measure on some space Ω, which in our application represents the positive real line on which the rates λ_l are defined. The concentration parameter $\tau > 0$ controls the distribution of the atomic masses, and thus the induced clustering properties of the process. A draw from a DP can be constructed as:

$$\mathrm{G} = \sum_{k=1}^{\infty} \beta_k \, \delta_{\phi_k}, \qquad \phi_k \overset{iid}{\sim} G_0, \tag{4.6}$$

where the weights β_k are obtained via the *stick-breaking* process (Sethuraman (1994)):

$$\beta_k = \nu_k \prod_{l=1}^{k-1} (1 - \nu_l) \qquad \nu_k \sim \mathrm{Beta}(1, \tau). \tag{4.7}$$

The process sequentially partitions the unit interval: the kth weight is a random proportion ν_k of the segment that remains after the first $k-1$ weights have been chosen. We denote this distribution by $\beta \sim \mathrm{GEM}(\tau)$.

The DP has proven useful in many applications due to its clustering properties (cf., Teh et al. (2006)). The *predictive distribution* of draws $\lambda_l \sim \mathrm{G}$ shows why the DP produces clusters. 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 in (4.6). For each sampled observation λ_l, let z_l index the corresponding unique parameter ϕ_k such that $\lambda_l = \phi_{z_l}$. The predictive distribution on the membership variables can be written as

$$Z_{L+1} | (z_1, \ldots, z_L, \tau) = \begin{cases} K+1 \text{ w.p. } \frac{\tau}{L+\tau}, \\ k \qquad \text{w.p. } \frac{n_k}{L+\tau} \text{ for } k = 1, \ldots, K, \end{cases} \tag{4.8}$$

where n_k indicates the number of members taking value k,and K identifies the number of distinct values observed through the first L samples. The distribution on partitions induced by the sequence of conditional distributions in (4.8) is commonly referred to as the *Chinese restaurant process* (CRP). The CRP provides an alternative representation to the DP (Pitman (2006)). This representation emphasizes the reinforcement property of the DP that leads to its clustering properties. It can be shown that the expected number of clusters using the CRP grows as $O(\tau \log(L))$ where L is the number of observations (see Teh (2011) for a detailed proof). This implies that the average number of clusters is much smaller than the number of observations.

In our model for crime rates, we impose a DP prior on the L tract-specific rates, λ_l. The number of observations from the DP is equal to the number of

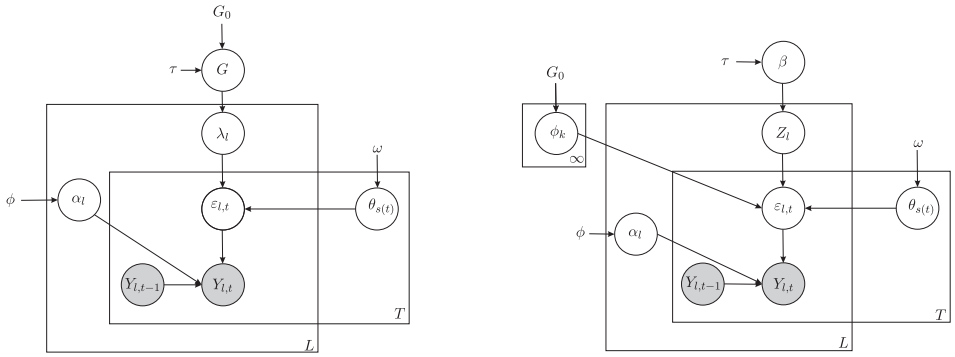

Figure 3. Graphical model of the multiple PoINAR(1) model. *Left:* Innovations generating from a Dirichlet process mixture model as in (4.9). *Right:* An equivalent representation using cluster indicator variables $z_1, \ldots, z_L$ as in (4.10).

tracts, rather than the number of time points. The DP prior thus groups the time series according to their corresponding tract-specific rates into a few clusters; tracts in the kth cluster share a common rate ϕ_k. The grouping of the time series into a small number of clusters provides useful shrinkage that pools information across the cluster, thereby yielding more accurate out-of-sample predictions for the multiple time series. When we combine the tract-specific rates and the seasonal effects, we obtain a generating process for the innovations:

$$\begin{aligned} \epsilon_{l,t} &\sim \text{Poisson}(\lambda_l \, \theta_{s(t)}) \quad l = 1, \ldots, L \quad t = 1, \ldots, T, \\ \theta_m &\sim \text{F} \quad m = 1, \ldots, 12, \\ \lambda_l &\sim \text{G} \quad l = 1, \ldots, L, \\ \text{G} &\sim \text{DP}(\tau, G_0) \; . \end{aligned} \tag{4.9}$$

For our application, we choose F to be a gamma distribution (see (5.6)).

Figure 3 (left) shows a graphical representation of our dependent multiple PoINAR(1) process. For details on interpreting such graphical representations, the reader is referred to Jordan (2004). Alternatively, we can employ an equivalent representation using the GEM distribution, membership indicators $z_1, \ldots, z_L$, and unique rate parameters ϕ_k (see Figure 3 (right)):

$$\begin{aligned} \epsilon_{l,t} &\sim \text{Poisson}(\phi_{z_l} \, \theta_{s(t)}) \quad l = 1, \ldots, L \quad t = 1, \ldots, T, \\ \theta_m &\sim \text{F} \quad m = 1, \ldots, 12, \\ z_l &\sim \beta \quad l = 1, \ldots, L, \\ \phi_k &\sim G_0 \quad \beta \sim \text{GEM}(\tau) \quad k = 1, 2, \ldots . \end{aligned} \tag{4.10}$$

4.3. Prior specification

The multiple PoINAR(1) requires estimation of three main components: thinning values $(\alpha_1, \ldots, \alpha_L)$, one for each tract-specific time series monthly seasonal effects, $(\theta_1, \ldots, \theta_{12})$, and rates for each tract, $[\lambda_1, \ldots, \lambda_L]$. The Bayesian framework places prior distributions on each of these three elements. Our priors are both computationally convenient and weakly-informative. For the thinning values and monthly seasonal effects we specify:

$$\begin{aligned} \alpha_l &\overset{\text{i.i.d.}}{\sim} \text{Beta}(\eta_1, \eta_2) \quad \text{for } l = 1, \ldots, L, \\ \theta_m &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(\xi_1, \xi_2) \quad \text{for } m = 1, \ldots, 12. \end{aligned} \tag{4.11}$$

We also explored the half-normal distribution as a prior for the seasonal effect, and we found the choice did not produce material changes from the results presented in Section 7. The DP prior on the tract-specific rates λ_l outlined in Section 4.2 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 to our model because not only is it conjugate to the Poisson distribution, but it also provides a natural interpretation. In particular, we have a prior belief that weekly rates of violent crime are typically small, but a few tracts have higher rates. 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 (1994).

4.4. Relationship to the vector AR(1) process

The continuous counterpart to the multiple PoINAR(1) is the Gaussian first-order vector autoregressive process, denoted VAR(1). This process is composed of L possibly dependent AR(1) processes and can be formulated as

$$\begin{aligned} \boldsymbol{Y}_{t+1} &= \boldsymbol{A}\,\boldsymbol{Y}_t + \boldsymbol{\epsilon}_{t+1} \quad t = 1, \ldots, T, \\ \boldsymbol{\epsilon}_t | \Sigma &\overset{\text{i.i.d.}}{\sim} N(0, \Sigma), \\ Y_{l,0} | \mu_{l,0}, \sigma_{l,0} &\overset{\text{i.i.d.}}{\sim} N(\mu_{l,0}, \sigma_{l,0}) \quad l = 1, \ldots, L, \end{aligned} \tag{4.12}$$

where $\boldsymbol{A}$ denotes an $L \times L$ matrix whose maximum eigenvalue has modulus less than 1, $[\Sigma]_{i,j} = 0$ for $i \neq j$ and $[\Sigma]_{i,i} = \sigma_i^2$. Compare this specification to the multiple PoINAR(1):

$$\begin{aligned} \boldsymbol{Y}_{t+1} &= \boldsymbol{\alpha} \circ \boldsymbol{Y}_t + \boldsymbol{\epsilon}_{t+1} \quad t = 1, \ldots, T, \\ \boldsymbol{\epsilon}_t | \Lambda_t &\overset{\text{i.i.d.}}{\sim} [\text{Poisson}(\Lambda_{1,t}), \text{Poisson}(\Lambda_{2,t}), \ldots, \text{Poisson}(\Lambda_{L,t})], \\ Y_{l,0} | \Lambda_{l,0} &\overset{\text{i.i.d.}}{\sim} \text{Poisson}(\Lambda_{l,0}) \quad l = 1, \ldots, L. \end{aligned} \tag{4.13}$$

These two models not only share similar notation, but also possess two common characteristics:

- The distributions of the innovations match the marginal distributions of $Y_{l,t}$. The multiple PoINAR(1) with diagonal $\boldsymbol{\alpha}$ ($\alpha_i \in [0,1]$) has Poisson marginal distributions while the VAR(1) has Gaussian marginals for any suitable matrix $\boldsymbol{A}$. With appropriate initialization and innovations, the resulting processes have a Poisson or Gaussian stationary distribution, respectively.
- If $\boldsymbol{A}$ is diagonal, the autocorrelation coefficient in both models, $\text{corr}(Y_{l,t+1}, Y_{l,t})$, is the diagonal element of its coefficient matrix, $\alpha_{l,l}$ or $A_{l,l}$.

These similarities to the continuous VAR(1) make the discrete PoINAR(1) especially attractive and easy to interpret. The VAR(1) process, however, 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 complicates inference for the PoINAR(1) model that we address in Section 5.

5. The MCMC Sampler

The PoINAR(1) model combines two underlying processes: the binomial thinning process and the innovations process. Each of these processes has its own parameters: binomial thinning uses the thinning parameters $\{\alpha_l\}$ and the innovations process uses the rates $\{\phi_k\}$ and 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 VAR(1) model there is no need to sample the innovations sequence because they are uniquely determined by the observations and model parameters. Therefore, one would expect the multiple PoINAR(1) model to be computationally cumbersome compared to its VAR(1) counterpart. However, our proposed sampler harnesses computational advantages from small observed counts in our crime data and sufficient statistics implied by the Poisson model. We outline the resulting sampler below. For detailed derivations, see 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 innovations series for the lth tract. The full conditional distribution factors are

$$P(\boldsymbol{\epsilon}|\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\theta}) = \prod_{l=1}^{L} \prod_{t=2}^{T} P(\epsilon_{l,t}|Y_{l,t-1}, Y_{l,t}, \alpha_l, \lambda_l, \boldsymbol{\theta}). \tag{5.1}$$

Given the observations $\boldsymbol{Y}$ and the parameters of the multiple PoINAR process, the innovations can be sampled independently for each tract 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

$$P(\epsilon_{l,t}|Y_{l,t-1}, Y_{l,t}, \alpha_l, \lambda_l, \boldsymbol{\theta}) \propto \frac{1}{\epsilon_{l,t}!(Y_{l,t} - \epsilon_{l,t})!(Y_{l,t-1} - (Y_{l,t} - \epsilon_{l,t}))!} \left(\frac{\lambda_l \theta_{s(t)} (1 - \alpha_l)}{\alpha_l}\right)^{\epsilon_{l,t}}. \quad (5.2)$$

Although this expression does not define a well-known discrete distribution, it is analytically tractable because of the small counts in the data ($\max\{0, Y_{l,t} - Y_{l,t-1}\} \leq \epsilon_{l,t} \leq Y_{l,t}$ and $Y_{l,t}$ is assumed to be small). In the crime data for Washington, D.C., $\max Y_{l,t} = 11$. 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 $y_{l,t} = 0$ then $\epsilon_{l,t} = 0$ and if $y_{l,t-1} = 0$ then $\epsilon_{l,t} = y_{l,t}$. Since our crime data has many zero counts, these constraints substantially lower the computational cost of this portion of the sampling. If larger counts are observed, then one can use a Metropolis-Hastings step to sample from this distribution with a Poisson proposal distribution. Importantly, if the observed counts are large enough, one can apply a stabilizing transformation such as the square root and model the resulting process as a VAR. This strategy has been shown by Brown et al. (2005) to yield satisfactory results in the univariate case when λ_l exceeds (roughly) 5.

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

$$P(z_l = k|\boldsymbol{z}_{/l}, \boldsymbol{\epsilon}, \Theta, \gamma_1, \gamma_2, \tau) \propto \begin{cases} \tau\, p_{l,0} & \text{for } k = K+1, \\ n_k\, p_{l,k} & \text{for } k = 1, \ldots, K, \end{cases} \quad (5.3)$$

where $K + 1$ identifies a previously unseen cluster, $\Theta = \sum_{t=1}^{T} \theta_{s(t)}$, and $\boldsymbol{z}_{/l}$ is the vector of membership indicators, not including the lth term. The first terms, (τ, n_j), of (5.3) arise from the CRP prior of (4.8) 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 negative binomial distributions

$$p_{l,0} = \frac{\Gamma(S_l + \gamma_1)}{\Gamma(\gamma_1) S_l!} \left(\frac{\gamma_2}{\Theta + \gamma_2}\right)^{\gamma_1} \left(\frac{\Theta}{\Theta + \gamma_2}\right)^{S_l}, \quad (5.4)$$

$$p_{l,j} = \frac{\Gamma(S_l + A_j + \gamma_1)}{\Gamma(A_j + \gamma_1) S_l!} \left(1 - \frac{\Theta}{n_j\, \Theta + \gamma_2}\right)^{A_j + \gamma_1} \left(\frac{\Theta}{n_j\, \Theta + \gamma_2}\right)^{S_l},$$

where $S_l = \sum_{t=1}^{T} \epsilon_{l,t}$ and $A_j = \sum_{i:z_i=j, i\neq l} S_i$. Here $p_{l,0}$ and $p_{l,j}$ only rely on 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 (5.3) for the z_l, marginalizing $\{\phi_k\}$.

3. Sample unique rates, ϕ_k. Although the rates collapse away in sampling the cluster indicators, z_l, they are needed for sampling the innovations sequence (Step 1) and seasonal effects (Step 3). As such, we sample the unique rates as auxiliary variables for these steps, and then discard them. For each currently specified cluster, sample ϕ_k as

$$\phi_k|\boldsymbol{\epsilon}, \boldsymbol{z}, \Theta, \gamma_1, \gamma_2 \sim \text{Gamma}(B_k + \gamma_1, n_k\,\Theta + \gamma_2), \tag{5.5}$$

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

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

$$\theta_m|\boldsymbol{\epsilon}, \boldsymbol{\phi}, \xi_1, \xi_2 \sim \text{Gamma}\Big(\sum_{l=1}^{L}\sum_{t:s(t)=m} \epsilon_{l,t} + \xi_1, q_m \sum_{l=1}^{L}\lambda_l + \xi_2\Big), \tag{5.6}$$

where q_m counts the number of occurrences of the mth month in the data. Notice that for this step we sum the innovations over tracts rather than time.

5. Sample the vector of thinning parameters, $[\alpha_1, \ldots, \alpha_L]$. For tract l,

$$\alpha_l|\boldsymbol{\epsilon}_l, \boldsymbol{Y}_l, \eta_1, \eta_2 \sim \text{Beta}(\sum_{t=2}^{T} Y_{l,t} - S_l + \eta_1, \sum_{t=2}^{T}(Y_{l,t-1} - Y_{l,t}) + S_l + \eta_2), \tag{5.7}$$

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 (1994).

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.

6. Simulation Examples

To demonstrate the performance of our model, we simulated datasets from 9 different multiple PoINAR(1) processes of (4.2). Each dataset had $L = 100$ time series (tracts) with $T = 208$ observations that correspond to 4 years of weekly data. We grouped the multiple time series into four equally sized clusters that each share a common rate. The data sets varied in the choice of:

Table 1. RMSE of estimates of the conditional mean obtained by the CLS, SPP and our Bayesian nonparametric method. The last row shows the expected (true) conditional expected value.

	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
BNP 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

1. The separation between the cluster rates, ϕ_k. We examined an "easy" setting in which the four cluster rates were well separated at $1, 3, 6, 10$, a "medium setting" with less distinct rates $0.01, 0.5, 1.2, 2$, and a "hard" setting with rates $0.1, 0.2, 0.3, 0.6$.
2. The thinning values, α_l, which determine the autocorrelation of the individual PoINAR(1) processes. The examples used a common choice for α_l for all tracts in a dataset, chosen from $\alpha_l = 0.1, 0.5, 0.9$.

We evaluated the root mean square error (RMSE) and absolute percentage error (APE) of our MCMC sampler both in- and out-of-sample. These metrics measure the distance between the true population expected value and its corresponding estimate based on the observed $L = 100$ time series. The simulation results show that our model produces accurate out-of-sample forecasts under various configurations. Table 1 presents the RMSE of our Bayesian nonparametric model compared to the RMSE of a simple Poisson process model (SPP) and the conditional least-squares model (CLS). (The Supplementary Material, Section 2, details these.) The results in Table 1 show that our model outperforms these alternatives. As expected, the larger the separation between the cluster rates, the easier it is for our method to identify the true clusters and yield better estimates for their parameters. Also, higher autocorrelation helps our method produce more accurate estimators.

These simulation results indicate that the sampler finds clusters when they exist. It is also important to demonstrate that the model does not spuriously spawn clusters when the data are homogeneous. As part of our simulations, we examined the performance of these methods in the situation in which a single process (cluster) generates all of the time series. The findings are presented in Section 3 of the Supplementary Material, which also contains a more detailed description of the simulations and results presented in this section. As one would hope, under these conditions we identify a single cluster, further validating our methodology.

7. Analysis of Violent Crimes

We examined both in- and out-of-sample results using the reported counts of violent crimes in Washington, D.C., as described in Section 1. The data consist of $L = 188$ time series (census tracts) with $T = 418$ weeks of counts in 2001 through 2008. We used the first 7 years of data to train our model and the last 52 weeks to evaluate its out-of-sample forecasts. We ran 5 MCMC chains for $5,000$ iterations from different initial values, each drawn from the priors

$$\begin{aligned} \theta_m &\sim \text{Gamma}(1,1) \quad \alpha_l \sim \text{Beta}(1,1) \\ \tau &\sim \text{Gamma}(2,4) \quad \phi_i \sim \text{Gamma}(1,1). \end{aligned} \tag{7.1}$$

We performed a sensitivity analysis for the hyperparameters during the simulation stage, but found no significant changes to the results. We discarded the first 1,000 iterations as burn-in and then thinned the remaining $4,000$ samples. While thinning was not absolutely necessary, we found it computationally convenient to thin our MCMC output by retaining the full output of every 50th iteration. Some recent discussion about thinning can be found in Owen (2015). Therefore, our inference for each parameter of the model is based on the resulting $80 \cdot 5 = 400$ MCMC samples. We used the scale reduction factor recommended by Gelman and Rubin (1992) to monitor convergence across the chains.

We begin by looking at the distribution of the number of clusters over the 400 samples in the left panel of Figure 4. The mode is 17 clusters, which is a substantial reduction from the original $L = 188$ time series. Figure 5 presents a representative cluster assignment along with the posterior rates for this assignment. This cluster assignment was selected as the assignment that had the minimum average Hamming distance across the different iterations (see Fox et al. (2011) for further details). An interesting phenomenon was that census tracts assigned to the same cluster were frequently spatially separated.

We further examined 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) maps the posterior mean rates for the census tracts in Washington, D.C.. We can see certain regions that exhibit higher rates (e.g., tracts that correspond to a southern portion of the city, a central portion along 16th Street, and an east-central portion along Rhode Island Avenue.) The results of Figure 6 are also substantiated by Figure 1 (right). Figure 7 compares the sample autocorrelation of the counts for each tract with the posterior mean thinning values. The sample autocorrelation was 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 determine the autocorrelation for each INAR(1) time series. The comparison shows that the

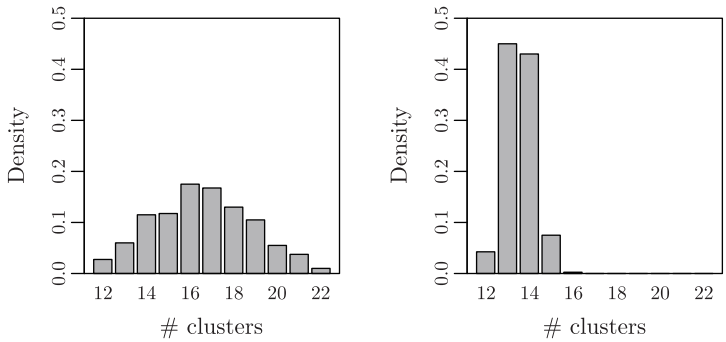

Figure 4. Histograms of the posterior number of clusters for the multiple dependent PoINAR(1) described in Section 4.2 (left) and the population adjusted multiple dependent PoINAR(1) model described in Section 8 (right).

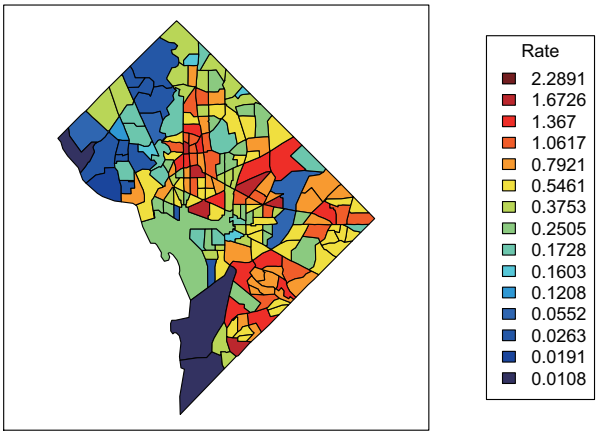

Figure 5. The minimum average Hamming distance cluster assignment along with the corresponding posterior rate values.

raw data autocorrelations vary over a wider range than their corresponding posterior mean values and some of these raw autocorrelations are slightly negative. Two reasons 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 autocorrelations are probably due to noise variation and therefore we are less concerned about this phenomenon. The standard error of an estimated first-order autocorrelation for white noise is approximately $1/\sqrt{T} = 1/\sqrt{418} \approx 0.05$; hence the bulk of raw autocorrelations are within about 2 standard errors of zero.
2. The posterior mean thinning values are adjusted for seasonal effects, whereas the raw autocorrelations are not. The values would be smaller in magnitude

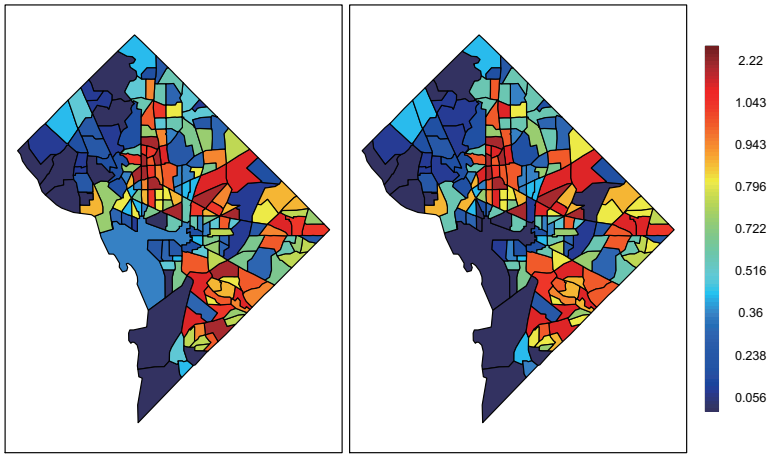

Figure 6. Map of posterior mean rates, λ_l, sampled from the multiple PoINAR(1) model described in Section 4.2 (left) and the population adjusted multiple PoINAR(1) model described in Section 8 (right).

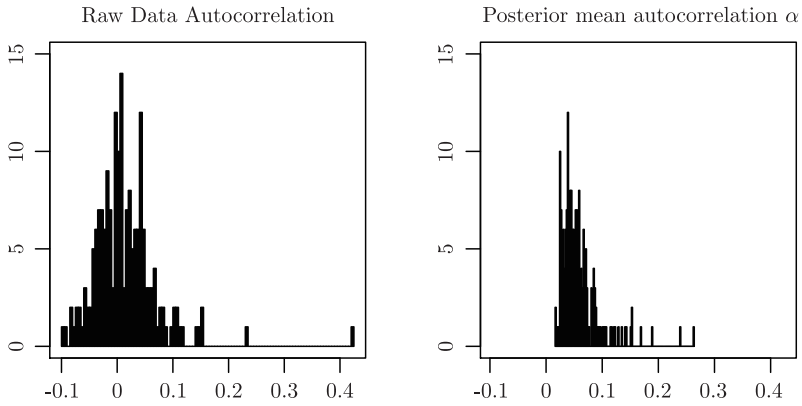

Figure 7. Histogram of raw data autocorrelations (left) and posterior mean autocorrelations α_l (right).

after adjusting for the seasonality, as our results suggest.

For the out-of-sample evaluation we compared our MCMC method to the CLS and SPP. For both the CLS and our method, we used the mean of the forecast distribution (posterior predictive distribution) one-week-ahead as the predictor for the crime counts in each tract,

$$\hat{y}_{l,T+1} = \alpha_l \, y_{l,T} + \lambda_l \, \theta_{s(T+1)}. \tag{7.2}$$

For CLS, we simply plugged-in the estimates of α_l, λ_l, and θ_m for each tract. For our method, we computed an MCMC-based estimate by evaluating (7.2) for each

of the 400 MCMC iterations and used the average of these as the final predicted value (see Section 3 of the Supplementary Material for further details). For the SPP, we averaged the past values as the predictor.

We predicted the one-week-ahead number of crimes in each tract for the first week of each month during 2008. Table 2 shows the one-week-ahead predicted mean RMSE and corresponding standard errors, conditional on the last observed value. The results indicate that when the last observed count is one of the most frequent values (0,1,2), our method produces lower RMSE. For the less frequent, higher counts (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 lower, more frequent counts and worse in the rare cases of high counts. A summary of the average one-week-ahead bias is presented in Section 4 of the Supplementary Material. 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.

The one-step-ahead conditional mean value is the best linear unbiased estimator under quadratic loss. Since the CLS method minimizes the observed squared error, it is only natural to evaluate all three methods using the same loss function. Alternatively, Berk (2008) proposed a quantile loss function that reflects the sensitivity of the police department to forecasting errors. Under a υ-quantile loss function, the predictor is just the predictive distribution's υ quantile. Using our method, one can easily sample from the following one-step-ahead predictive posterior distribution and evaluate any desired quantile:

$$\begin{aligned} &P(Y_{l,t+1} = y_{l,t+1} | Y_{l,t} = y_{l,t}, \alpha_l^{(m)}, \boldsymbol{\theta}^{(m)}, \lambda_l^{(m)}) \\ &= \sum_{r=0}^{\infty} \binom{y_{l,t-1}}{y_{l,t}-r} (\alpha_l^{(m)})^{y_{l,t}-r} (1-\alpha_l^{(m)})^{y_{l,t-1}-(y_{l,t}-r)} \frac{e^{-\phi_l^{(m)} \theta_{s(t+1)}^{(m)}} (\phi_l^{(m)} \theta_{s(t+1)}^{(m)})^r}{r!}, \end{aligned} \tag{7.3}$$

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 mth iteration of the MCMC sampler. Figure 8 shows the 95% and 99% quantiles for each of the 188 tracts and the corresponding one-step-ahead true value, $y_{l,T+1}$. The quantiles may also be used to provide prediction intervals for each tract. A police department can use these intervals along with the point estimate to distinguish between an unusual surge in crimes which requires allocation of more resources, and a random rise in crimes, which would not benefit from an intervention.

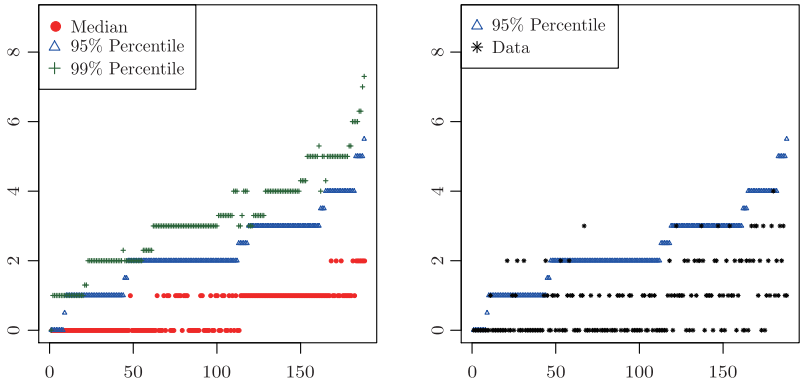

Figure 8. The predictive posterior distribution for each of the 188 tracts. The red dots correspond to the median predicted number of violent crimes for each tract. The blue triangles and green crosses correspond to the 95% and 99% percentiles of the predictive posterior distribution, respectively. The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

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.

$y_{\cdot,T}$	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)
Dependent PoINAR 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

8. Multiple PoINAR(1) with Covariates

Previous research has shown that crime rates are associated with demographic covariates, and we have several ways to incorporate such features in our Bayesian model. For example, Blei and Frazier (2010) incorporate covariates directly into the clustering mechanism. This approach might improve the accuracy of forecasts, but it would provide less in the way of interpretation, such as how the various covariates associate with crime rates. Instead, we take a more direct approach that offers the advantages of clustering as well as interpretation. We model the tract-specific rate λ_l as a linear function of covariates and cluster the coefficients of the equation. The clusters of coefficients may provide further insight into the relationships between crime and demographic characteristics.

8.1. Adjusting for population

The main goal of this section is to demonstrate how to add covariates to our

model and to explore the benefits of doing so. To this end, we looked at the population sizes in each of the census tracts as a possible explanatory variable.

Let X_l denote the population of the lth census tract. (We obtained the populations from the 2,000 census. Section 5 of the Supplementary Material shows a map of the population density in Washington, D.C..) To incorporate population into our model, we redefined the tract-specific rate as a linear function of population, $\lambda_l = X_l\,\psi_l$, where ψ_l is the number of violent crimes per person in the lth tract. We then placed a DP prior directly on the rate per person parameter, ψ_l, yielding the model

$$\begin{aligned} \epsilon_{l,t} &\sim \text{Poisson}(X_l\,\psi_l\,\theta_{s(t)}) \quad l = 1,\ldots,L \quad t = 1,\ldots,T, \\ \theta_m &\sim \text{F}(\omega) \quad m = 1,\ldots,12, \\ \psi_l &\sim \text{G} \quad l = 1,\ldots,L, \\ \text{G} &\sim \text{DP}(\tau, G_0). \end{aligned} \tag{8.1}$$

It is straightforward to adjust the MCMC sampler described in Section 5 to incorporate 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 while remaining weakly informative. After these simple modifications, we run the sampler in the manner previously described in Section 5.

8.2. Analysis of results

Using the covariate-adjusted multiple PoINAR(1) of Section 8.1, we again analyzed the counts of violent crimes in Washington. As in Section 7, we begin by showing the posterior distribution of the number of clusters over the 400 MCMC iterations (again taken from 5 chains, each run for 5,000 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. This suggests that population can account for a significant amount of the spatial heterogeneity in crime. Figure 9 maps the posterior means of the crime rate per person, ψ_l. This map highlights three main features: the center of Washington, D.C. has a high count of violent crimes per person; the northwest portion of the city has very few crimes per person; and the city has three hot-spots: in the center of the city and in the eastern and southwestern portions of the city.

These insights, also highlighted by Cahill and Roman (2010), differ from the conclusions one would make by simply looking at the mean values displayed in Figure 1 or from our previous analysis. The results emphasize tracts which exhibit high crime rates per person as opposed to high crime counts and can help police make future planning decisions, such as where to place a new station. These outcomes, important as they may be, are merely a byproduct of

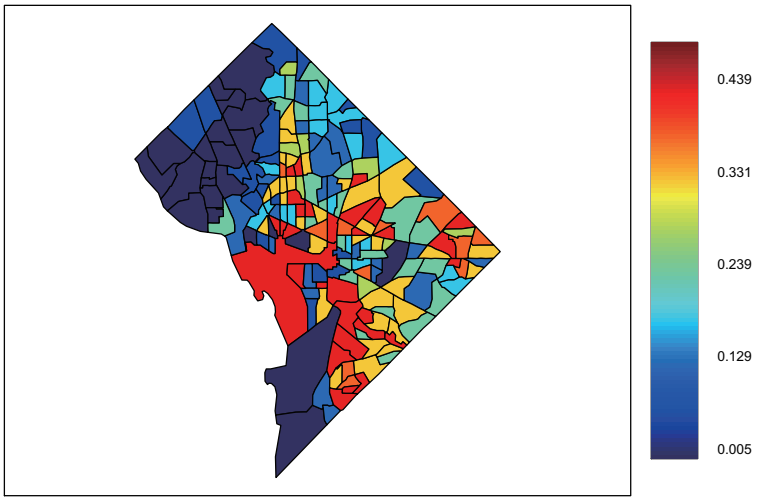

Figure 9. Map of the posterior mean values for crime rates per person, ψ_i.

our estimation method. The more interesting research question is whether the population covariate can improve the prediction abilities when compared to the unadjusted model.

We performed a one-week-ahead forecast for the last week in the series using both the unadjusted model of Section 4.2 and the adjusted model accounting for population. The overall RMSE of the unadjusted model was 0.9663 whereas the adjusted model was 0.9713. These results suggest that adding the population of the tract to the model does not necessarily improve predictive accuracy. However, a more extensive analysis would be needed to confidently settle such issues. Although adding the population of the tract to the model may not improve predictive accuracy, adding the covariate seems to provide a useful benefit in that the revised model reveals a more interpretable grouping of the time series. Of course, there are many other covariates that one could consider, for example measures of poverty, housing characteristics, etc.

9. Discussion

In this paper we have 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 model to borrow 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 identified, they remain fixed throughout time. One can relax this assumption and allow 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 those examined by Taddy (2010).

Finally, although our focus here is on counts of violent crimes, 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 (Boudreault and Charpentier (2011)), wildfires across counties (Xu (2011)), and so forth.

Supplementary Materials

The Supplementary Material provides further details on our MCMC sampler and the baseline models to which we compare. We also include an expanded discussion and set of results for our simulated data and Washington, D.C. crime data analyses.

Acknowledgements

This work was supported in part by NSF CAREER Award IIS-1350133.

References

Alzaid, A. and Al-Osh, M. (1988). First order integer valued autoregressive process: Distributional and regression properties. *Statist. Neerlandica* **42**, 53-61.

Berk, R. (2008). Forecasting methods in crime and justice. *Ann. Rev. Law Soc. Sci.* **4**, 219-238.

Blei, D. M. and Frazier, P. I. (2010). Distance dependent Chinese restaurant processes. *J. Machine Learn. Res.* **12**, 2461-2488.

Boudreault, M. and Charpentier, A. (2011). Multivariate integer-valued autoregressive models applied to earthquake counts. eprint arXiv:1112.0929.

Box, G. E. P. and Tiao, G. C. (2011). *Bayesian Inference in Statistical Analysis*, vol. 40. Wiley.

Brännäs, K. (1993). *Estimation and Testing in Integer-valued AR(1) Models*. University of Umeå.

Brown, L. D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyna, S. and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing science perspective. *Statist. Sci.* (Special Issue on Bayesian Statistics) **100**, 36-50.

Cahill, M. and Roman, J. K. (2010). Small number of blocks account for lots of crime in D.C. `http://www.dccrimepolicy.org/Briefs/images/DCPIBrief_CrimeByBlockFINAL_2.pdf`.

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**, 635-644.

Escobar, M. D. and West, M. (1994). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577-588.

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**, 43-54.

Fox, E. B., Sudderth, E. B., Jordan, M. I. and Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *Ann. Appl. Statist.* **5**, 1020-1056.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7**, 457-472.

Hillmer, S. C. and Tiao, G. C. (1982). An arima-model-based approach to seasonal adjustment. *J. Amer. Statist. Assoc.* **77**, 63-70.

Jordan, M. I. (2004). Graphical models. *Statist. Sci.* (Special Issue on Bayesian Statistics) **19**, 140-155.

McDowall, D., Loftin, C. and Pate, M. (2012). Seasonal cycles in crime, and their variability. *J. Quantitative Criminology* **28**, 389-410.

McKenzie, E. (2000). Discrete variate time series. *Simulation* **21**, 1-34.

Owen, A. B. (2015). Statistically efficient thinning of a Markov chain sampler. eprint arXiv:1510.07727.

Pedeli, X. and Karliss, D. (2011). A bivariate INAR(1) process with application. *Statist. Model.* **11**, 325-349.

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Springer-Verlag. Lecture Notes for St. Flour Summer School.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4**, 639-650.

Steutel, F. W. and Van Harn, K. (1986). Discrete operator-self decomposability and queueing networks. *Comm. Statist. Stochastic Models* **2**, 161-169.

Taddy, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *J. Amer. Statist. Assoc.* **105**, 1403-1417.

Teh, Y. W. (2011). Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 280-287.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566-1581.

Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications. *J. Amer. Statist. Assoc.* **76**, 802-816.

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

Xu, H. (2011). Point process modeling of wildfire hazard in Los Angeles County, California. *Ann. Appl. Statist.* **5**, 684-704.

The Climate Corporation, 201 Third Street, Suite 1100 San Francisco, CA 94103, USA.

E-mail: sivan@climate.com

Department of Statistics, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, USA.

E-mail: lbrown@wharton.upenn.edu

Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322, USA.

E-mail: ebfox@stat.washington.edu

Department of Statistics, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, USA.

E-mail: stine@wharton.upenn.edu

(Received June 2014; accepted March 2016)