

# Bayesian Dependent Functional Mixture Estimation for Area and Time-Indexed Data: An Application for the Prediction of Monthly County Employment

Terrance D. Savitsky\* and Matthew R. Williams†

**Abstract.** The U.S. Bureau of Labor Statistics (BLS) publishes employment totals for all U.S. counties on a monthly basis. BLS use the Quarterly Census of Employment and Wages, where responses are received on a 6–7 month lagged basis and aggregated to county, and apply a time series forecast model to each county and project forward to the current month, which ignores the dependence among counties. Our approach treats these by-county employment time series as a collection of area indexed noisy functions that we co-model. Our model includes predictor, trend and seasonality terms indexed by county. This application is among the first in the U.S. Federal Statistical System to address the joint modeling of a collection of time series expressing heterogeneous seasonality patterns between them. We demonstrate that use of a Fourier basis to model seasonality outperforms a locally-adaptive, intrinsic conditional autoregressive construction on our collection of time series where the degree of expressed seasonality varies. County-indexed parameters of the 3 terms are drawn from a dependent Dirichlet process (DDP) prior to allow the borrowing of information. We show that employment of both spatial and industry concentration predictors into the prior probabilities for co-clustering among the counties produces better prediction accuracy. Our DDP prior accounts for the possibility that nearby counties may express distinct underlying economic structures. A feature of our joint modeling framework is that it computes efficiently to support the monthly BLS production cycle. We compare the performances of alternative formulations for the dependent Dirichlet process prior on monthly, county employment data from 2002–2016.

**Keywords:** spatio-temporal modeling, Gaussian Markov Random Field, Dirichlet process, Bayesian hierarchical models, latent models, nonparametric statistics, functional data estimation, survey sampling.

**MSC2020 subject classifications:** 62F15, 62M10, 62-08, 62P20.

## 1 Introduction

The Local Area Unemployment Survey (LAUS) program of the U.S. Bureau of Labor Statistics (BLS) publishes employment and unemployment totals for local areas, in-

---

\*U. S. Bureau of Labor Statistics, Office of Survey Methods Research, Washington, DC 20212, USA, [Savitsky.Terrance@bls.gov](mailto:Savitsky.Terrance@bls.gov)

†National Center for Science and Engineering Statistics, National Science Foundation, Alexandria, VA 20212, USA, [mrwillia@nsf.gov](mailto:mrwillia@nsf.gov)

cluding counties, across all states in the U.S. in coordination and partnership with the states. BLS administers the Current Employment Statistics (CES) survey, on a monthly basis, to business establishments in relatively large metropolitan statistical areas. While covering much of the U.S. population, the CES survey *excludes* 1751 relatively sparsely populated counties of the total 3108 counties in the continental U.S. The LAUS program utilizes the Quarterly Census of Employment and Wages (QCEW), a census instrument administered to *all* business establishments by BLS, for the purpose of collecting workplace employment information for the 1751 counties not covered by the CES, which we label “non-CES counties”. There is, however, a 6–7 month availability lag for county estimates from the QCEW due to reporting requirements for establishments and subsequent internal BLS processing to flag and correct errors in these responses. As a result, the LAUS program applies a time series forecasting model to each non-CES county, separately, to project forward the QCEW employment to the current month. This approach ignores the dependence among the collection of county time series induced by similarities in their economic structures.

We devise a single model for all of the county time series and use this model to perform a 7-month prediction for all 3108 counties, simultaneously, in a fashion that allows the data to learn a dependence structure among the county time series.

Fixing a county, the time series for that county is modeled as the addition of trend, seasonality and predictor functional terms. We configure a nonparametric conditional autoregressive (CAR) (Rue and Held, 2005) prior for each of the trend and seasonality terms. The CAR prior is an equivalent to a smoothing spline model that utilizes local nearest neighbors to perform the smoothing, rather than shrinking to a global mean. The latent trend term is allowed to express a non-linear shape and the magnitude and shape of the seasonality may vary over the months. The noise is composed of measurement errors induced in the by-establishment reporting process.

The generating prior hyperparameters of each of the trend, seasonality and predictor coefficients are indexed by county. We allow the data to discover dependence among the counties by performing probabilistic clustering of these county-indexed parameters to achieve a non-parametric mixture distribution for each of the time series terms where we mix over counties. We employ a *dependent* Dirichlet process (DDP) prior framework of Müller et al. (2011) that uses a subset of predictors to construct a prior formulation for county co-clustering that increases the *prior* probability for the assignment of two counties to the same cluster to the extent that they express similar predictor values relative to the Dirichlet process (DP).

## 1.1 Introducing the QCEW Employment Data

Figure 1 renders time series for 6 randomly-selected counties over  $T = 180$  months, from January, 2002–December, 2016. Years in 5-year increments are labeled on the horizontal axis (displayed in 2-digits), where we observe some level of pause or decline in employment around the period of 2008 during the so-called, “great recession”. We expect the patterns expressed across the county-indexed collection of employment time series to demonstrate correlation with one another primarily based on their similarities

in economic composition or structure and the overall health of their economies, rather than based on spatial adjacency. Our modeling goal is to capture these dependencies among the collection of non-CES county time series in order to improve the bias and efficiency properties of predictions for the missing 6–7 months as compared to the modeling of each county in a separate model. Simultaneously, we wish to construct our joint model for the collection of county time series to allow sufficient estimation flexibility to capture features that take place for a subset of months and a subset of counties, like the great recession. We utilize the full set of  $n = 3108$  counties in our modeling, rather than the 1751 non-CES counties, in order to facilitate the borrowing of information for prediction of by-county employment totals. Our modeling approach must computationally scale with sufficiently fast turnaround to support the BLS monthly production cycle. The modeling of 3108 county times series, simultaneously, is facilitated by our use of probabilistic clustering over county-indexed model parameters described in the sequel. The clustering reduces the effective dimensionality of the 3108 sets of county-indexed parameters to the total number of estimated clusters.

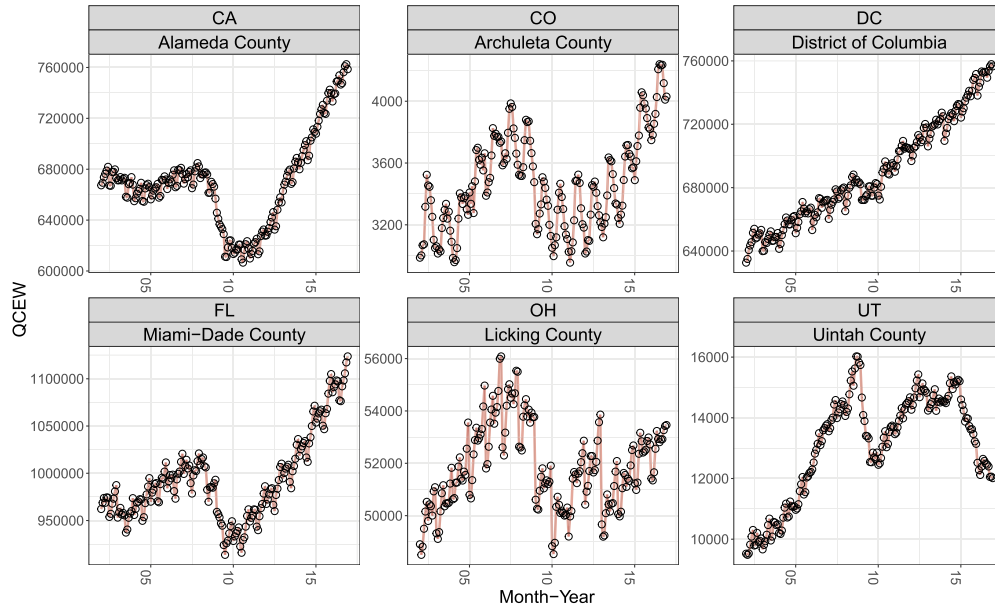


Figure 1: Randomly-selected by-county time series of employment total, from Jan, 2002–Dec, 2016. The horizontal axis denotes year in 5-year increments (labeled with the last 2-digits of the year). Employment totals reflect the reported number of employees for employers in each county, which is a place-of-work, rather than a place-of-residence measure.

The QCEW data are characterized by both evolving trends and 12-month seasonal patterns. Figure 2 displays the time series for Albany County, NY, where we observe both a long-term trend, including the downturn and recovery from the great recession, as well as a recurring 12-month seasonality.

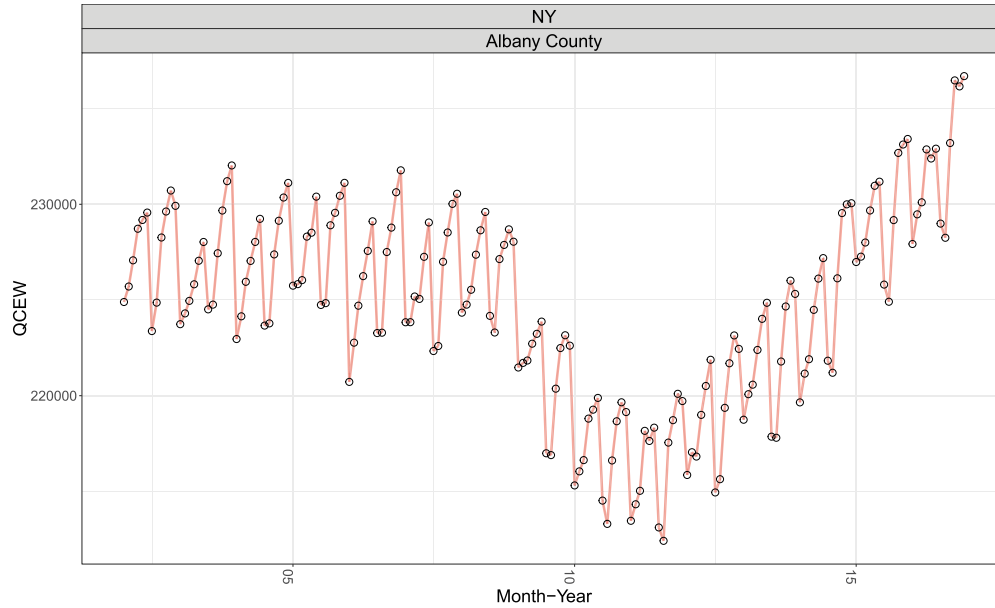


Figure 2: A single county time series of the employment total, from Jan, 2002–Dec, 2016. The horizontal axis denotes year in 5-year increments (labeled with the last 2-digits of the year). Employment totals reflect the reported number of employees for employers in each county, which is a place-of-work, rather than a place-of-residence measure.

The QCEW by-county data includes predictors that we use in both our fixed effects term and for our predictor-dependent clustering. The key predictor type that drives the prior distribution of co-clustering of county-indexed parameters is the “location quotient” (lq), defined as an index  $\in [0, 1]$  that quantifies the relative employment concentration of a particular industry in a county as compared to the national average. We use location quotients for industries constructed from the first 2-digits of the 6-digit detailed industry code assigned to each establishment under the North American Industry Classification System (NAICS) (which we denote as “economic sectors”). Economic sectors include; Construction, Transportation, Services, Leisure, Public, Mining, Manufacturing, Information, Education. We assemble the unemployment insurance (UI) claims in each month for each county from data provided by the states. The trend of the UI claims reveals provides a measure of the changes in the relative economic health of a county. We also utilize longitude and latitude as spatial predictors in the fixed effects term and also for our predictor-dependent clustering to capture residual spatial dependence not accounted for by the lq predictors.

We introduce our probability model in Section 2 that employs terms to capture seasonality, trend and incorporate predictors into county-indexed regression models, tied together in a predictor-dependent Dirichlet process mixture. We outline computations for the full conditional posterior distributions in Section 3, where we introduce the two

fit statistics we use to compare performances of our suite of models applied to prediction of QCEW in Section 4. We conclude with a discussion in Section 5.

## 2 Model for Collection of Time Series

We construct our data matrix of county employment totals as  $n \times T$ ,  $\mathbf{Y}$ , which we treat as a collection of county-indexed latent functions of time with additive noise. Our approach contrasts with a model that defines a spatially-indexed process that may evolve over time. We select the former modeling approach due to the presence of seasonality, which is more difficult to capture with the latter approach in a flexible, nonparametric construction. We do, however, later compare our approach to one implementation of a time evolving spatial process.

### 2.1 Model for Single County Time Series

We begin by fixing a county in some row of  $\mathbf{Y}$  and constructing our additive functional priors specification for the  $1 \times T$  county time series.

Our likelihood is formulated as,

$$y_j | \mathbf{x}_j, \boldsymbol{\beta}, \gamma_{1j}, \gamma_{2j} \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_j \boldsymbol{\beta} + \gamma_{1j} + \gamma_{2j}, \tau_y^{-1}), \tag{1}$$

$$\tau_y \sim \mathcal{G}(1, 1), \tag{2}$$

where  $\mathcal{N}(\cdot)$  and  $\mathcal{G}(\cdot)$  denote the normal and gamma distributions, respectively, with the latter parameterized with (shape, rate) hyperparameters. The first term,  $\mathbf{x}_j \boldsymbol{\beta}$ , includes the influence of predictors for the county estimated function for month,  $j \in (1, \dots, T)$ . The 2 random effect terms,  $(\gamma_{1j}, \gamma_{2j})$ , capture residual trend and seasonality patterns (not explained by the predictor term), respectively.

### 2.2 Multivariate Gaussian Mixture for Fixed Effects Term

The  $p \times 1$  vector of predictors,  $\mathbf{x}_j$ , includes latitude and longitude, by-industry location quotients, and the level of unemployment insurance claims for each county and month. Regression coefficients,  $\boldsymbol{\beta}$ , are, in turn, drawn from,

$$\boldsymbol{\beta} | \boldsymbol{\mu}, \Lambda \stackrel{\text{ind}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \Lambda^{-1}). \tag{3}$$

### 2.3 Conditional Autoregressive Prior for Random Effects Terms

We implement a conditional prior formulation for the trend and seasonal random effects that induces local smoothing with,

$$\gamma_{mj} | \boldsymbol{\gamma}_{m,-j}, \tau_{\gamma,m} \sim \mathcal{N}(\bar{\gamma}_{mij}, [d_{mj} \tau_{\gamma,m}]^{-1}), \tag{4}$$

$$\gamma_{m,-j} = (\gamma_{m1}, \gamma_{m2}, \dots, \gamma_{m(j-1)}, \gamma_{m(j+1)}, \dots, \gamma_{mT})', \quad (5)$$

$$\bar{\gamma}_{mj} = \frac{1}{d_{mj}} \sum_{r \sim j} \omega_{mjr} \gamma_{mr}. \quad (6)$$

Fixing a random effects term,  $m \in (1, 2)$ , the prior specification for  $(\gamma_{mj})_{j=1, \dots, T}$  in (4) is intrinsic conditional autoregressive (Rue and Held, 2005) over the  $T$  months to capture the autocorrelation in the time series for each county. The intrinsic conditional autoregressive prior (ICAR) encodes a sparse precision matrix through a  $T \times T$  adjacency matrix,  $\Omega_m = (\omega_{mjr})$ , where  $\omega_{mjr} = 1$  indicates  $j \sim r$ , (which is shorthand for  $j$  is a “neighbor of” or “communicates with”,  $r$ ), for random effect term,  $m \in (1, 2)$ ; otherwise,  $\omega_{mjr} = 0$ . The diagonal entries of  $\Omega_m$  are set to 0; e.g.,  $\omega_{mjj} = 0, \forall j \in (1, \dots, T)$ . The mean in the conditional distribution of  $\gamma_{mj}$  is proportional to an neighbor-weighted average of the  $(\gamma_{mr})$  over all  $r \sim j$  for each month,  $j$ , and the average is normalized by  $d_{mj} = \sum_{r \sim j} \omega_{mjr}$ , which denotes the number of neighbors of  $\gamma_{mj}$ . Since  $\omega_{jj} = 0, \forall j \in (1, \dots, T)$  and  $\omega_{jr} = 0, \forall r \not\sim j$ , we may re-express  $\bar{\gamma}_{mj}$  in (6) as the inner product,  $\frac{1}{d_{mj}} \omega'_{mj} \gamma_m^{T \times 1}$ . The  $T \times T$  matrix,  $C_m := D_m^{-1} \Omega_m$ , provides a shorthand notation for the weighted average,  $\bar{\gamma}_{mj} = C_{mj}^{1 \times T} \gamma_m^{T \times 1}$  where  $T \times T$ ,  $D_m = \text{diag}(d_{m1}, \dots, d_{mT})$ .

We note that our prior constructions for  $(\gamma_m)$  are *a priori* and *a posteriori* independent over terms,  $m = 1, 2$ , given the parameters of the conditional autoregressive models.

We utilize 2 random effects terms: 1. A trend term; 2. A 12-month seasonality term. Both are defined on a line of  $T = 180$  months. The intrinsic conditional autoregressive prior for the seasonal term constructs,  $\Omega_2$ , with seasonal increments,  $(\gamma_{2j}, \gamma_{2(j+1)}, \dots, \gamma_{2(j+O-1)})$ , for  $j = 1, \dots, T - O + 1$ , which produces a joint density for  $T \times 1$ ,  $\gamma_2$  proportional to,

$$p(\gamma_2 | \tau_{\gamma, 2}) \propto \tau_{\gamma, 2}^{\frac{T-O+1}{2}} \exp \left( -\frac{\tau_{\gamma, 2}}{2} \sum_{j=1}^{T-O+1} (\gamma_{2j} + \gamma_{2(j+1)} + \dots + \gamma_{2(j+O-1)})^2 \right), \quad (7)$$

where  $O = 12$  denotes the period of seasonality. The bandwidth of seasonal adjacency matrix,  $\Omega_2$ , is  $O$  (Rue and Held, 2005). Our construction for seasonality is nonparametric where the data are able to learn a varying or evolving seasonal pattern over time.

The trend term ( $m = 1$ ) employs a so-called “random walk of order 1” (RW1) construction, with increments,  $\gamma_{1j} - \gamma_{1(j-1)}$  and  $\gamma_{1j} - \gamma_{1(j+1)}$ , where for each month  $j$ ,  $\Omega_1$  is constructed such that the month before,  $\gamma_{1,(j-1)}$  and the month following,  $\gamma_{1,(j+1)}$  are deemed neighbors of  $\gamma_{1j}$ . The beginning and last months of the time series communicate with only a single neighbor. Both conditional autoregressive terms induce probabilistic local smoothing within increments defined by first differences for the trend term and by the sum of  $O$  months for the seasonal term. The structure of the  $(\Omega_m)$  are band diagonal with widths proportional to the lengths of the increments. (See Rue and Held (2005) for more details.)

The implied joint distribution for term  $m \in (1, 2)$ ,  $\gamma_m = (\gamma_{m1}, \dots, \gamma_{mT})' = \mathcal{N}_T(\mathbf{0}, Q_m^{-1} = [\tau_m(D_m - \Omega_m)]^{-1})$ , where precision matrix,  $Q_m$ , is rank deficient such

that the joint distribution is improper. Under our RW1 prior for the trend term,  $\gamma_1$ , the overall mean level is undefined, since the prior is constructed on increments specified as differences of random effect values for neighboring time periods (on a line), which produces rank  $T - 1$  for  $Q_1$ . Similarly, the  $O = 12$  order of seasonality term produces  $Q_2$  to be of rank  $T - O + 1$  since the joint prior is defined up to any repeated seasonal pattern,  $(c_1, \dots, c_O)$ , that sums to 0 ( $\sum_{o=1}^O c_o = 0$ ) in each sequence of  $O$  values. The rank deficiencies arise because the ICAR prior is defined locally on neighboring values in a moving average construction, rather than globally across the  $T$  months for each county. Even though the ICAR prior produces an improper joint distribution, the conditional distributions are, nevertheless, proper.

A general Bayesian dynamic linear model (DLM) is specified with  $y_t = x_t + \nu_t$ ,  $x_t = F_t' \gamma_t$  and  $\gamma_t = G_t \gamma_{t-1} + \epsilon_t$ , where  $F_t$  is a predictor matrix and  $G_t$  denotes an evolution matrix and  $\nu_t$  and  $\epsilon_t$  are independent noise processes with global scale parameters (West, 2013; Prado and West, 2010). For a trend term, the  $G_t$  is diagonal with each parameter bounded  $\in (-1, 1)$ , which induces a global (as contrasted with a local neighborhood of time points) shrinking of all  $\gamma_t$ . Our trend term under the CAR prior is equivalent to setting the diagonals of  $G_t$  to 1, which induces a nonparametric random walk neighborhood smoother. Our choice of the first difference under our CAR formulation uses time points immediately following and after each focus time point for smoothing  $\gamma_t$ . Seasonality under a DLM specifies  $G_t = \phi H_a$ , where  $H_a$  is a sinusoidal term for angle,  $a \in (0, 2\pi)$  and  $\phi \in (0, 1)$ .  $H_a$  encodes a parametric, fixed seasonal smoothing pattern. Our CAR prior for seasonality is more flexible as the seasonality pattern is permitted to evolve over time, though we see in the sequel that this flexibility confounds the clustering of seasonal term patterns over counties and induces undesired spiky artifacts. In contrast to local seasonal smoothing, our use of the Fourier basis for modeling seasonality provides a semi-parametric global estimate of seasonality that we later introduce as an alternative to the CAR prior.

We adjust the ICAR prior for the modeling of the seasonality term,  $\gamma_2$ , by including a global parameter,  $\rho_m \in [0, 1)$ , that we intend to allow better estimation of low seasonality expressed as a dampened (longer) wavelength or length scale.

$$\bar{\gamma}_{mj} = \rho_m \times \frac{1}{d_{mj}} \sum_{r \sim j} \omega_{mjr} \gamma_{mr} \tag{8}$$

which produces the joint distribution for  $\gamma = (\gamma_{m1}, \dots, \gamma_{mT})' = \mathcal{N}_T(\mathbf{0}, Q_m^{-1} = [\tau_m(D_m - \rho_m \Omega_m)]^{-1})$ , where precision matrix,  $Q_m$ , is of full rank for  $\rho_m \in [0, 1)$ . Fixing  $\rho_m = 1$  parameterizes the rank-deficient improper joint distribution of the ICAR prior. The parameter,  $\rho_m$ , is often interpreted as a strength-of-correlation parameter (Banerjee et al., 2003) because the dependencies indexed in  $\Omega_m$  are included in proportion to the value of  $\rho_m$ . Another way that we may interpret  $\rho_m$  is as a length scale adjustment parameter that determines the degree of smoothness in the estimated  $T \times 1$  function,  $\gamma_m$  (Mukherjee et al., 2011). The maximum length scale is hard-coded based in the bandwidth of  $T \times T$   $\Omega_m$  and  $\rho_m$  allows the data to adjust the length scale in the  $T \times 1$  function to be larger. Values of  $\rho_2$  closer to 1 will produce the minimum length scale that estimates a higher level of seasonality. The introduction of  $\rho_m$ , however, induces

a global behavior over the  $T \times 1$  time series for joint shrinking of all of the  $(\gamma_{2t})_{t=1}^T$ , which is a similar behavior to the DLM where for a uni-dimensional  $\gamma_t$ ,  $\rho_2$  is equivalent to  $\phi$ , though the seasonality pattern is still more flexible than for a DLM. We label this model as a proper conditional autoregressive model (PCAR) in order to distinguish it from the ICAR model. Note that we also use  $\rho_1$  for prior on trend term,  $\gamma_1$ , under the PCAR option for ease of implementation, though the data estimate  $\rho_1 \approx 1$ , so it makes no difference in the modeling of the trend term.

## 2.4 Fourier Basis Alternative for the Seasonal Effects Term

The PCAR alternative dampens the expression of seasonality in the predictions for low-seasonality counties, but we later show that it also tends to dampen estimated seasonality for high-seasonality counties, which is an undesired effect. The length scale adjustment parameter,  $\rho_m$ , is difficult to estimate because each  $\rho_m$  parameterizes a precision matrix and the degree of estimated seasonality is very sensitive to its value. By contrast, the ICAR prior does the opposite by over-amplifying seasonality in the predictions for low-seasonality counties due to its local smoothing properties that use a subset of time points to estimate each monthly seasonal coefficient. While the modeling of seasonality is well-understood in application to a single time series, there is little guidance on prediction of seasonalities for a collection of time series in a single model.

We are motivated to construct a sinusoidal basis to augment our predictors to provide a more stable estimator for seasonality that can accommodate both high and low expressions of seasonality. The Fourier basis employs predictors,  $\mathbf{z}_j$ , constructed as an orthogonal trigonometric basis,

$$\begin{matrix} (O-1) \times 1 \\ \mathbf{z}_j \end{matrix} = \left\{ \cos\left(\frac{2\pi k_1 j}{O}\right), \sin\left(\frac{2\pi k_2 j}{O}\right) \right\}'_{k_1=1, \dots, O/2, k_2=1, \dots, (O/2-1)} \quad (9)$$

for an even value of  $O$ , the periodicity of our repeated pattern, which is  $O = 12$  months for our QCEW time series. This set-up constructs a basis of oscillating functions that, together with a global intercept, spans any repeated pattern of periodicity with order,  $O$ .

Wei (2006) propose the use of Fourier functions as an orthogonal basis in the time domain for modeling of periodic functions (e.g., seasonality) that contrasts with the usual use of Fourier functions to transform from the time to the frequency domain (which we do not do). Montesinos-López et al. (2018) compare B-spline and Fourier bases for the modeling of seasonality in functional data regression and conclude that the Fourier basis performs best.

Our analysis suggests to us that best results are achieved by incorporating the Fourier basis,  $\mathbf{z}_j$ , into an augmented predictor vector,  $\mathbf{x}_j \leftarrow (\mathbf{x}_j, \mathbf{z}_j)$  (in contrast to creating a separate term for Fourier basis), and modeling them as part of the fixed effects term in (3). The Fourier basis construction estimates seasonality, globally in time, across each county time series. While less flexible than the ICAR prior and more similar to a DLM, we expect this approach to better adapt to variations in the amplitude of seasonality across counties. We retain the ICAR parameterization for the trend effects,  $\gamma_1$ , but drop the  $\gamma_2$ , under this alternative Fourier basis set-up.



Prediction of seasonality in the presence of our predictors in an augmented  $\mathbf{x}_j$  induces a moderate amount of over-shrinkage in the prediction of seasonality. The over-shrinkage is completely eliminated by replacing the predictor matrix with its principal components where we discard any redundant components (with zero eigenvalues). This step is similar to the Moran's I basis construction (Hughes and Haran, 2013) used for the modeling of spatial association (which we discuss, in detail, in the sequel as a competitor for our modeling framework). Both approaches employ a spectral rotation.

We have defined 3 *alternatives* for the modeling of seasonality among the collection of time series: 1. ICAR prior on random effects,  $\gamma_2$ ; 2. PCAR prior on random effects,  $\gamma_2$ ; Fourier basis (FB),  $\mathbf{z}_j$ , incorporated into the fixed effects term with coefficients  $\beta$ , in lieu of  $\gamma_2$ .

## 2.5 Dependent Dirichlet Process

We now expand our model from a single county model of (1), to the collection of 3108 counties by indexing our observed response,  $y_j$  and the fixed effects,  $\beta_j$ , and our trend and seasonality terms,  $(\gamma_{1j}, \gamma_{2j})$ , by county,  $i \in (1, \dots, n)$  in the expanded likelihood,

$$y_{ij} | \mathbf{x}_{ij}^{p \times 1}, \beta_i, \gamma_{1ij}, \gamma_{2ij} \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_{ij} \beta_i + \gamma_{1ij} + \gamma_{2ij}, \tau_y^{-1}). \quad (10)$$

We collect the county-indexed *generating* hyperparameters for the predictor, trend and seasonality terms into  $\theta_i := (\mu_i, \Lambda_i, (\tau_{\gamma, 1i}, \rho_{1i}))$  (in the case we use a Fourier basis to model seasonality). The indexing of  $\theta_i$  by county,  $i \in (1, \dots, n)$  induces a mixture formulation over county-indexed  $\theta_i$  for each of the random effect terms. We cluster the county-indexed generating distribution hyperparameters by using a nonparametric mixing measure. We index the generating parameters for all of the Fourier basis seasonality, fixed effects coefficients and trend terms by county because there is a substantial heterogeneity among the expressed time series patterns across the 3108 counties of the U.S., as we see in the illustrations of Figure 1, such that we seek a joint prior distribution that provides a large support over time series patterns. (In the case we use an ICAR or PCAR prior for seasonality in lieu of the Fourier basis, their generating patterns are also indexed by county.) To make this choice intuitive, we note that if we fail to index the generating parameters of the seasonality term (for any of the Fourier basis, PCAR or ICAR alternatives) we would fail to account for the massive heterogeneity in the presence of and patterns for expressed seasonality over the counties.

The predictors in the fixed effects term may capture some dependence among the counties induced by similarities in industries, spatial location or economic conditions. The use of location quotients may be especially useful to avoid over-smoothing based on spatial location; for example, two nearby counties (one rural and the other urban) may express very different economic structures, despite their spatial adjacency. Trend and seasonality components are expressed to varying degrees among the counties. We index the 3 terms by county,  $i \in (1, \dots, n)$ , in order to capture heterogeneity among the counties, regulated by a partition model (clustering) prior that allows the data to learn additional dependence among the collection of county time series. We show in

the sequel that using the lq and spatial predictors to influence the prior probabilities of co-clustering among the counties, in addition to their use in the fixed effects term, substantially improves prediction accuracies. Each time series,  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$ , is scaled to variance 1 in order to allow the basis of clustering to be the shape of the functions, rather than their magnitudes.

Our use of a nonparametric prior to be next introduced for clustering county-indexed functional generating parameters is of a similar idea to that of Gelfand et al. (2005), but they impose the Dirichlet process prior directly on the function parameters (e.g., on  $(\beta_i, \gamma_{1i}, \gamma_{2i})$ ), so that co-clustered functions must be *exactly* the same, which we find to be too restrictive for working with our QCEW employment data application. Instead, we impose our nonparametric clustering prior on the generating parameters,  $\theta_i$ , of the prior distributions for the functions such that functions which are co-clustered are drawn from the same *distribution*, but are not required to be exactly equal.

Our mixing measure,  $F$ , is used to draw  $\theta_1, \dots, \theta_n | F \sim F$ .  $F$  is not subscripted by a parameter to indicate that we will, in turn, sample the unknown distribution,  $F$ , from a *nonparametric* Dirichlet process (DP) prior,  $F | \alpha, F_0 \sim DP(\alpha F_0)$ . The DP may be specified by expressing  $F$  in a discrete, stick-breaking construction (Sethuraman, 1994),

$$F = \sum_{k=1}^{\infty} p_k \delta_{\theta_k^*}, \quad (11)$$

where  $F$  is a countably infinite mixture of weighted point masses (or “spikes”) with “locations”,  $\theta_1^*, \dots, \theta_K^*$ , in the support of  $F$  indexing the unique values for the  $(\theta_i)$ , where  $K \leq n$ . We record cluster memberships of counties with  $\mathbf{s} = (s_1, \dots, s_n)$  where  $s_i = k$  denotes  $\theta_i = \theta_k^*$  so that  $\{\mathbf{s}, (\theta_k^*)\}$  provides an equivalent parameterization to  $(\theta_i)$  and we recover  $\theta_i = \theta_{s_i}^*$ . The DP is induced by placing priors on the weights,  $p_k$ , and the locations,  $\theta_k^*$ . Each location is drawn from  $\theta_k^* \sim F_0$ . Under the alternative using the Fourier basis,  $F_0$  is constructed as:

$$\begin{aligned} F_0 = & \text{Huang-Wand}(\Lambda_i | \delta, a_1, \dots, a_p) \times \\ & \mathcal{N}_p(\boldsymbol{\mu}_i | \mathbf{m}, P_1^{-1}) \times \\ & \mathcal{G}(\tau_{\gamma, 1i} | a = 1, b = 1) \times \mathcal{U}(\rho_{1i} | 0, 1), \end{aligned} \quad (12)$$

where Huang-Wand  $(\Lambda | \delta, a_1, \dots, a_p)$  denotes the prior of Huang and Wand (2013) defined for covariance matrices that generalizes the inverse Wishart prior to produce marginally noninformative half  $t$ -distributed priors with  $\delta$  degrees of freedom for the variances. Choosing  $\delta = 2$ , produces  $\mathcal{U}(-1, 1)$  priors for the correlations. The prior for a precision matrix is constructed by forming  $Q = 2\delta \text{diag}(a_1, \dots, a_p)$  with  $a_\ell \sim \mathcal{G}(1/2, 1)$ ,  $\ell = 1, \dots, p$ . We proceed to draw  $\Lambda \sim \mathcal{W}_p(\nu, Q)$  where  $\nu = \delta + p - 1$ . By contrast, the usual Wishart prior would fix the diagonals of  $Q$ . We set  $\delta = 2$  to produce a weakly informative prior for the variances.

The weight,  $p_k \in (0, 1)$  is composed as  $p_k = v_k \prod_{h=1}^{k-1} (1 - v_h)$  where  $v_k$  is drawn from the beta distribution,  $\mathcal{B}e(1, \alpha)$ . This construction provides a prior penalty on the number of mixture components, but we also see that a higher value for  $\alpha$  will produce

more clusters (unique locations). Since the locations are drawn from  $F_0$ , as the number of unique locations increases, the estimated  $F$  approaches the base distribution,  $F_0$ . Under this construction,  $\mathbb{E}(F) = F_0$ .

The joint prior for cluster assignments under the DP prior is stated with,

$$f(s_1, \dots, s_n) \propto \alpha^{K-1} \prod_{k=1}^K (n_k - 1)!, \tag{13}$$

after marginalizing out the random measure,  $F$ , where,  $n_k = \sum_{\ell=1}^n \mathbb{I}(s_\ell = k)$  denotes the number of counties assigned to cluster,  $k$ . We note that the DP prior assigns counties,  $i \in 1, \dots, n$  to cluster,  $k \in (1, \dots, K)$  proportionally to the popularity of cluster,  $k$ , as measured by its size,  $n_k$ .

We, next, generalize the above-specified DP formulation. We induce predictor-assisted clustering by, first, jointly modeling our response variable,  $y_{ij}$ , along with a subset of selected predictors,  $\boldsymbol{\xi}_{ij} \subset \mathbf{x}_{ij}$ . We specify the additional likelihood,

$$\boldsymbol{\xi}_{ij}^{q \times 1} | \boldsymbol{\mu}_{\xi,i}, \Lambda_{\xi,i} \stackrel{\text{ind}}{\sim} \mathcal{N}_q \left( \boldsymbol{\mu}_{\xi,i}, \left( \Lambda_{\xi,i} \right)^{-1} \right), \tag{14}$$

and, second, include the parameters of this likelihood in an augmented,  $\boldsymbol{\theta}_i := (\boldsymbol{\mu}_i, \Lambda_i, (\tau_{\gamma,1i}, \rho_{1i}), \boldsymbol{\mu}_{\xi,i}, \Lambda_{\xi,i})$  on which we, third, impose prior  $F$  that, in turn, receives a DP prior. Our intent is to increase the *a priori* probability for co-clustering of counties,  $i$  and  $\ell$ , proportionally to how similar are the values for the  $q \times T$ , predictors,  $\Xi_i = (\xi_{i1}^{q \times 1}, \dots, \xi_{iT})$  and  $\Xi_\ell$ . Augmenting  $\boldsymbol{\theta}_i$  to include the mixing parameters of (14) adjusts (13) to *incorporate* information about the predictors that produces,

$$f \left( s_1, \dots, s_n | \Xi_1^{q \times T}, \dots, \Xi_n \right) \propto \alpha^{K-1} \prod_{k=1}^K g(\Xi_k^*) (n_k - 1)!, \tag{15}$$

where,

$$g(\Xi_k^*) = \int \prod_{i:s_i=k} \prod_{j=1}^T f(\boldsymbol{\xi}_{ij} | \boldsymbol{\mu}_{\xi,k}^*, \Lambda_{\xi,k}^*) f(\boldsymbol{\mu}_{\xi,k}^*, \Lambda_{\xi,k}^*) d\boldsymbol{\mu}_{\xi,k}^* d\Lambda_{\xi,k}^*, \tag{16}$$

such that  $g(\Xi_k^*)$  constructs the similarity function of Müller et al. (2011). If counties  $i$  and  $\ell$  share similar values for their predictors then there is a higher prior probability that they will be co-clustered and  $(\boldsymbol{\mu}_{\xi,i}, \Lambda_{\xi,i}) = (\boldsymbol{\mu}_{\xi,\ell}, \Lambda_{\xi,\ell}) = (\boldsymbol{\mu}_{\xi,k}^*, \Lambda_{\xi,k}^*)$ . The integration outlined in (16) is numerically performed in our Markov Chain Monte Carlo (MCMC) posterior sampling algorithm.

The integration of (16) and incorporation of  $g(\Xi_k^*)$  into (15) is accomplished by simply including  $(\boldsymbol{\mu}_{\xi,i}, \Lambda_{\xi,i}) \in \boldsymbol{\theta}_i$ , under the DP prior construction. Augmenting  $\boldsymbol{\theta}_i$  in this way may be viewed as a computational device to produce (15). We *don't* believe that the predictors,  $\Xi_i$ , are random, but treat them as such in order to insert the similarity functions,  $(g(\Xi_k^*))_{k=1, \dots, K}$  into the prior for  $\mathbf{s}$ . By doing so, we have estimated a complicated predictor-dependent Dirichlet process mixture model with a much simpler Dirichlet process model through imposing a probability model on our target predictors.

The insertion of predictor information into the prior for cluster assignments can reduce prior uncertainty for the cluster assignments of establishments who share similar predictor values with a large number of other establishments; nevertheless, as the number of predictors grows, the prior uncertainty will grow because it is defined on the space of predictor values. So care is warranted in the choice of which and how many predictors to use for indexing the prior for cluster assignments.

Augmenting  $\theta_i$  to include the generating parameters,  $(\mu_{\xi,i}, \Lambda_{\xi,i})$ , extends  $F_0$  to,

$$\begin{aligned} F_0 = & \text{Huang-Wand}(\Lambda_i | \delta, a_{11}, \dots, a_{1p}) \times \\ & \mathcal{N}_p(\mu_i | \mathbf{m}, P_1^{-1}) \times \\ & \mathcal{G}(\tau_{\gamma,1i} | a = 1, b = 1) \times \mathcal{U}(\rho_{1i} | 0, 1) \times \\ & \text{Huang-Wand}(\Lambda_{\xi,i} | \delta, a_{31}, \dots, a_{3q}) \times \mathcal{N}_q(\mu_{\xi,i} | \mathbf{0}, P_3^{-1}). \end{aligned} \quad (17)$$

There are similar approaches to ours that construct spatial- or predictor-dependent DP priors that could be used to induce clustering over counties; for example, the probit stick breaking construction of Rodríguez and Dunson (2011) specifies a stick-breaking construction as do we for  $F$  in (11) with weights,  $p_k(\mathbf{s}) = \Phi(\alpha_k(\mathbf{s})) \prod_{r < k} (1 - \Phi(\alpha_r(\mathbf{s})))$ , where  $\Phi(\cdot)$  denotes the cdf of a Gaussian distribution, for each spatial location,  $s$ . They then suggest using a Gaussian process prior over the  $\{\alpha_k(\mathbf{s}_i)\}_{i=1}^n$  to encode a spatial dependence (over counties). Hossain et al. (2013) similarly utilize the stick breaking construction to induce a dependence among the weights with  $p_{ki} = q_{ki} h_{ki} \prod_{\ell < k} (1 - q_{\ell i} h_{\ell i})$ , where  $h_{\ell i} \sim \beta(1, \alpha)$  and  $q_{ki} = \exp(\mathbf{x}_{ki} \delta_{ki}) / \sum_{\ell=1}^K \exp(\mathbf{x}_{\ell i} \delta_{\ell i})$ , where  $\mathbf{x}_{ki}$  are general predictors and  $\delta_{ki}$  are regression coefficients. Rodríguez and Dunson (2011) note that the incorporation of by-county dependence through the weights, rather than the atoms/locations in the stick-breaking construction (as we do in our method), has the notable advantage that it places some prior mass on an independent prior (whereas our formulation does not in that there is always some non-zero probability of co-clustering, no matter how far apart are predictor values). Since Hossain et al. (2013) allow incorporation of any predictors, not just spatial locations, it could accommodate our use of the lq predictors (which we know from experience are more relevant than are spatial locations for our class of economic data). Yet, we do not pursue these weight-dependent approaches because they are simply computationally infeasible for an application of our size. The elegance of our approach is that while it constructs a predictor-dependent Dirichlet process, it is estimated as an ordinary DP simply by augmenting  $\theta_i$  with predictor model hyperparameters of  $\Xi_i$ . So we are able to utilize efficient DP estimation algorithms, such as Algorithm 8 of Neal (2000b).

Another class of alternatives model a smooth spatial dependencies among the collection of county time series. Nobre et al. (2011) generalize the DLM setup with  $x_t(\mathbf{s}) = \sum_{\ell=1}^p \phi_{\ell}(\mathbf{s}) x_{t-\ell} + \epsilon_t(\mathbf{s})$  for spatial location,  $\mathbf{s}$  where  $|\phi_{\ell}(\mathbf{s})| < 1$  where here  $p$  denotes the autoregressive (AR) order. The restriction on  $\phi_{\ell}(\mathbf{s})$  makes estimation difficult, so they reparameterize the set-up using the characteristic polynomial,  $H_{\ell}(\mathbf{s})$ , and the back-shift operator,  $B$ , to achieve,  $\prod_{\ell=1}^p (1 - BH_{\ell})(\mathbf{s}) x_t(\mathbf{s}) = \epsilon_t(\mathbf{s})$ . While they accomplish efficient estimation, their focus is confined to modeling spatial associations among the time series, which is insufficient for our application to economic, employment data. Our

industry-indexed lq predictors are much more powerful for this purpose in predicting both association and heterogeneity. Our computation is also expected to be even more efficient than is theirs because clustering induces a dimension reduction over county-indexed parameters.

## 2.6 Summary of our Joint Model over Counties

We summarize the probability model for the joint likelihood  $(y_{ij}, \xi_{ij})$  for our response and subset of predictors,  $\xi_{ij} \subset \mathbf{x}_{ij}$ . We co-model  $\xi_{ij}$  in order to extract the county-indexed parameters and include them in the *prior* probability for cluster assignments of counties to induce a higher prior probability of co-clustering among counties sharing similar predictor values. This procedure is used to estimate the conditional distribution for  $y_{ij}|\mathbf{x}_{ij}$  in a fashion that utilizes predictors in both the fixed effects term and in the prior probabilities of co-clustering.

**Likelihood for response,  $(y_{ij})_{i,j}$** , with  $i = 1, \dots, (n = 3108)$  counties and  $j = 1, \dots, (T = 180)$  months.

$$y_{ij} | \mathbf{x}_{ij}, \beta_i, \gamma_{1ij}, \gamma_{2ij} \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_{ij}\beta_i + \gamma_{1ij} + \gamma_{2ij}, \tau_y^{-1}), \tag{18}$$

$$\tau_y \sim \mathcal{G}(1, 1) \tag{19}$$

**Likelihood for predictors used for clustering,  $\xi_{ij} \subseteq \mathbf{x}_{ij}$**

$$\xi_{ij}^{q \times 1} | \mu_{\xi,i}, \Lambda_{\xi,i} \stackrel{\text{ind}}{\sim} \mathcal{N}_q\left(\mu_{\xi,i}, \left(\Lambda_{\xi,i}\right)^{-1}\right) \tag{20}$$

**Gaussian Mixture prior for fixed effects coefficients (may include FB),  $\beta_i$**

$$\beta_i^{p \times 1} | \mu_i, \Lambda_i \stackrel{\text{ind}}{\sim} \mathbf{N}_P(\mu_i, \Lambda_i^{-1}) \tag{21}$$

**Mixture formulations for random effects terms,  $(\gamma_{mi})$ ,  $m = (1, 2)$**

**Proper CAR prior on both Trend and Seasonality terms**

**Employ seasonal term,  $(\gamma_{2i})$ , in lieu of Fourier basis in augmented  $\mathbf{x}_{ij}$**

**The ICAR prior alternative is achieved from PCAR by removing  $\rho_{mi}$**

$$\gamma_{mij} | \gamma_{mi,-j}, \tau_{\gamma,mi}, \rho_{mi} \sim \mathcal{N}(\bar{\gamma}_{mij}, [d_{mj}\tau_{\gamma,mi}]^{-1}), \tag{22a}$$

$$\gamma_{mi,-j} = (\gamma_{mi1}, \gamma_{mi2}, \dots, \gamma_{mi(j-1)}, \gamma_{mi(j+1)}, \dots, \gamma_{miT})', \tag{22b}$$

$$\bar{\gamma}_{mij} = \rho_{mi} \times \frac{1}{d_{mj}} \sum_{r \sim j} \omega_{mjr} \gamma_{mir}, \tag{22c}$$

$$= \frac{\rho_{mi}}{d_{mj}} \omega_{mj}' \gamma_{mi}, \tag{22d}$$

$$= \rho_{mi} \times C_{mj} \gamma_{mi} \tag{22e}$$

Dependent Dirichlet process (DDP) prior for measure,  $F$

Construction of  $\theta_i$  under PCAR prior on Seasonal term

$$\theta_i := \left( \boldsymbol{\mu}_i, \Lambda_i, (\tau_{\gamma, mi}, \rho_{mi})_{m=1, \dots, 2}, \boldsymbol{\mu}_{\xi, i}, \Lambda_{\xi, i} \right) \quad (23)$$

Construction of  $\theta_i$  under Fourier Basis construction of Seasonal term

$$\theta_i := \left( \boldsymbol{\mu}_i, \Lambda_i, (\tau_{\gamma, 1i}, \rho_{1i}), \boldsymbol{\mu}_{\xi, i}, \Lambda_{\xi, i} \right) \quad (24)$$

DDP prior

$$\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n | F \sim F, \quad (25a)$$

$$F | \alpha, F_0 \sim \text{DP}(\alpha, F_0), \quad (25b)$$

$$\begin{aligned} F_0 = & \text{Huang-Wand}(\Lambda_i | \delta, a_{11}, \dots, a_{1p}) \times \\ & \mathcal{N}_p(\boldsymbol{\mu}_i | \mathbf{m}, P_1^{-1}) \times \\ & \mathcal{G}(\tau_{\gamma, 1i} | a = 1, b = 1) \times \mathcal{U}(\rho_{1i} | 0, 1) \times \\ & \text{Huang-Wand}(\Lambda_{\xi, i} | \delta, a_{21}, \dots, a_{2q}) \times \mathcal{N}_q(\boldsymbol{\mu}_{\xi, i} | \mathbf{0}, P_2^{-1}) \end{aligned} \quad (25c)$$

### 3 Computation

We utilize a sequential Gibbs scan to construct our MCMC (posterior) sampler for all of our models, under which each parameter set is sampled from its full conditional posterior distribution. Random effects,  $(\gamma_{mij})$ , are drawn from conjugate univariate normal conditional distributions over the  $T$  time points for each county,  $i \in 1, \dots, n$ . Coefficients,  $(\boldsymbol{\beta}_i)$ , of the predictor term (including the Fourier basis) are sampled from conjugate multivariate Gaussian conditional distributions. The  $n \times 1$  cluster assignment parameter vector,  $\mathbf{s}$ , is sampled from a nonconjugate posterior conditional distribution using the well-known algorithm 8 of Neal (2000b), which nominates new cluster locations ahead of sampled observations that may be selected on each posterior draw in the cluster assignment,  $s_i$ , for each county,  $i = 1, \dots, n$ . The posterior probability for cluster assignment of each county multiplies the prior distributions for each of the parameters in  $\theta_i$  into the prior probability for cluster assignment. Parameter locations,  $(\Lambda_{\xi, k}^*, \boldsymbol{\mu}_{\xi, k}^*)$ , associated with predictors,  $(\boldsymbol{\xi}_{ij})$ , use the predictor likelihood for (14) to achieve our predictor dependent Dirichlet process construction. Conditioned on the  $(a_{\ell q})$  parameters of the Huang-Wand priors, the precision matrix locations,  $(\Lambda_k^*, \Lambda_{\xi, k}^*)$ , are sampled from Wishart distributions. The locations for length scale parameters of the PCAR model,  $(\rho_{mk}^*) \in [0, 1)$ , are sampled from the univariate slice sampler of Neal (2000a). The remaining parameters are precision parameters and are sampled from conjugate univariate gamma conditional distributions.

All of our models are implemented in R Core Team (2014) using C++ (for fast computation) wrapped into easy-to-use R functions, which we will make available on request. Each model is run for 50,000 iterations and half are discarded as burn-in. Convergence of the MCMC sampler is assessed by employing a fixed width estimator with Monte Carlo standard errors (MCSE) computed using the consistent batch means (CBM) method (Jones et al., 2006).

The Fourier basis DPP model (DDP-FB) takes approximately 63 hours or 2.6 days to complete the 50,000 iterations over our  $n = 3108$  counties, each observed for  $T = 180$  months, in a single thread of an Intel Core i7 processor. The ICAR DDP (DDP-ICAR) and PCAR DDP (DDP-PCAR) models take about 10% longer than DDP-FB. This computation time readily supports the BLS monthly production cycle and we note in the sequel that our chosen model provides high quality uncertainty quantification, which is important information that the LAUS program provides to the States to accompany the predictions.

We employ two prediction error criteria to compare the relative performances of our models: 1. The root mean-squared prediction error (RMSPE) based on the squared difference between the predicted values (obtained from the model posterior predictive distribution) and true values for  $y_{ij}$  for the 7 predicted months over all 3108 counties; 2. A mean absolute prediction error percentage (MAPE-C), separately computed for *each* county, and then averaged over all of the counties. This criterion equally weights a 5% error in a tiny county, like Loving County, TX (with population of about 133) with a 5% error in Harris County, TX (which contains the city of Houston). On the one hand, accurate estimation of employment totals in *every* county is important to the States and it is for this reason that the LAUS program chose this criterion for assessing prediction accuracy. On the other hand, perhaps a state would accept a relatively small increment in error for a relatively less populated county in exchange for a decrement in the predicted error for a relatively more populated county. This criterion intends to highlight the uniform prediction performance of the models across all of the counties estimated by the LAUS program. The LAUS survey program restricts computation of the MAPE-C to the 1751 subset of non-CES counties, which are smaller and more heterogeneous than the larger CES counties. They make this restriction to non-CES counties for computing MAPE-C because they use the available survey estimates for the CES counties, which are much more stably estimated due to their relatively large sample sizes. The computations of MAPE-C for the model alternatives in this paper are performed by the LAUS survey program.

## 4 Application to Employment Prediction for QCEW

We apply our dependent functional mixture model alternatives to the prediction of county employment from the QCEW during the lag period. QCEW employment levels from January, 2002–December, 2016 are used for this purpose, where we exclude the last 7 months (June–December, inclusive) for model estimation and subsequently conduct predictions for the excluded, last 7 months of 2016 by drawing  $(y_{ij})_{i=1,\dots,n;j=174,\dots,180}$  from its posterior predictive distribution (which marginalizes over the posterior samples for the model parameters). The prediction for each county and month is formulated as the mean its posterior predictive distribution. Our fixed effects,  $\mathbf{x}_j$  include the 9 industry-indexed lq predictors, UI claims and latitude and longitude, the latter two predictors capture residual spatial dependence. The subset of predictors,  $\xi_{ij}$ , use for our predictor-assisted DDP formulation are the location quotients and the latitude and longitude. Using location quotients in the DDP assigns a higher probability for co-clustering to two counties where the underlying economic structures are similar.

We begin by comparing the relative performances in the predictions of *seasonality* for both counties that well express seasonality (high-seasonality counties) and counties that express little seasonality (low-seasonality counties) because this comparison determined our model selection. The green, short-dashed line in Figure 3 during the 7-month prediction period corresponds to predictions using the ICAR prior (which sets the  $(\rho_{2i}) = 1$ ). We label this model as “DDP-ICAR”, where “DDP” indicates that the ICAR prior is embedded in the predictor-dependent Dirichlet process prior. The DDP-ICAR induces spikiness in many of the counties where seasonality is lightly and irregularly expressed due to the local, moving average manner in which seasonality is estimated where only a subset of months is used to estimate each seasonality coefficient,  $\gamma_{2ij}$ . Most low-seasonality counties are highly irregular in their expressions of seasonality. The clustering appears to confound the estimation because co-clustered counties that express even small differences in seasonality amplitudes may induce a spike in the county whose data expresses a smaller seasonality amplitude. For example, the observed time series counties like Uintah County, UT, displayed in Figure 1 express highly irregular patterns. There is essentially too much flexibility in the estimation of seasonality under the DDP-ICAR model with the result that misclustering induces large local (for a subset of months) spikes. Though most of response values in the prediction interval are well-predicted, the spike induces a large error on one or two data points.

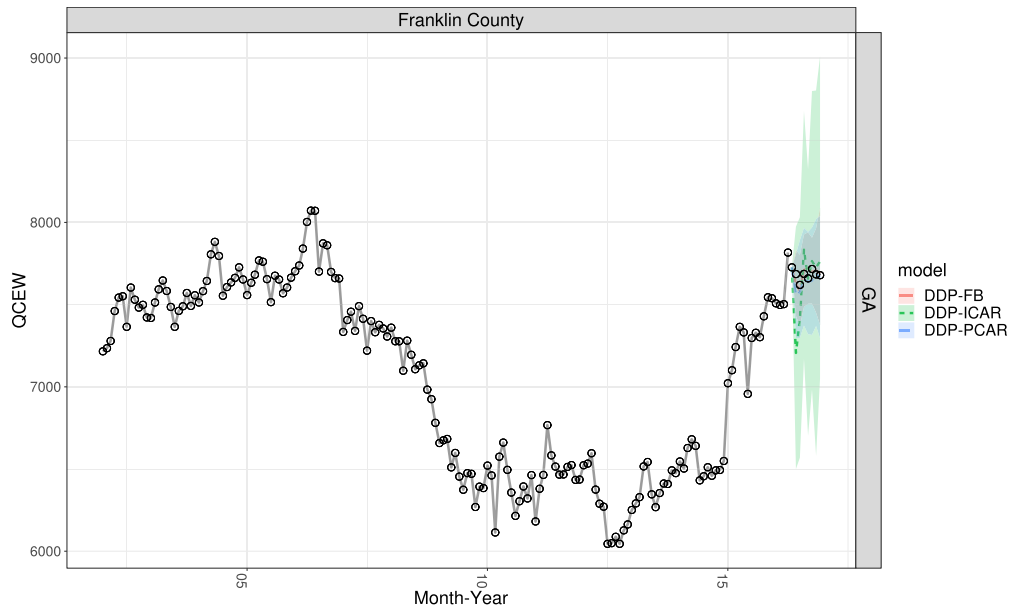


Figure 3: Comparing the prediction performance of the ICAR prior (green, short-dashed line) to a PCAR prior (blue, long-dashed line) and to the Fourier basis prior (pink, solid line) for a county where seasonality is *less* expressed for the 7-month prediction period. The hollow circles represent the data points. The shading represents 95% credible intervals.



As we expected, allowing the data to estimate the  $(\rho_{2i}) < 1$  in the PCAR construction allocated the low-seasonality counties to clusters of with  $\rho_{2k}^* \approx 0.7$ , which produced more smoothing and eliminated the spike as seen in the blue, long-dashed line. We label this model as “DDP-PCAR”. There is, however, a bit of oversmoothing of the seasonal term under the PCAR model relative to the simpler FB model, shown in the pink, solid line. The degree of estimated seasonality is particularly sensitive to  $\rho_{2i}$  because  $\rho_{2i}$  multiplies by a sparse adjacency matrix that fully determines the seasonal amplitude and pattern. While we achieve good mixing in our MCMC for  $\rho_{2i}$  using a single parameter to determine the degree of seasonality allowed into the model, the adjustment is too coarse.

The FB model appears to perform relatively the best in terms of capturing small magnitude month-to-month variations, a performance advantage that holds across low-seasonality counties. The shaded regions in each plot panel represent the 95% credible intervals and we note that DDP-FB is relatively more efficient in uncertainty quantification than are the other two models. The FB uses a fixed set of oscillating predictors and the coefficients are global across months. It represents the best balance between the coarse PCAR and the overly flexible ICAR.

Comparing the same models for a typical high-seasonality county in Figure 4 reveals that the DDP-PCAR dramatically oversmooths seasonality, which is, by contrast, very well predicted by both the DDP-ICAR and DDP-FB, though the DDP-FB expresses a slightly higher degree of smoothness, which likely accounts for its prediction stability on low-seasonality counties. The DDP-FB model dominates the other two in prediction accuracy over *both* low- and high-seasonality counties such that it achieves both the lowest Mean Squared Prediction Error (MSPE) and MAPE-C, as shown in Table 1. Figure 4 also reveals that prediction uncertainty is more efficient for the DDP-FB model. Accurate uncertainty quantification is important because it provides the states with valuable inputs on the relative confidence in the predictions that helps focus dialogue between the LAUS program and states regarding state judgments about their economic conditions. Even the DDP-PCAR achieves lower prediction fit statistics than does the DDP-ICAR because the spikes induced in low-seasonality counties by DDP-ICAR induce large errors and because there are relatively more counties that expressed relatively weak seasonality patterns. The  $\rho_k^*$  for the length scale adjustment location parameter for clusters containing high-seasonality counties are estimated at  $\approx 0.9$ , which produces dramatically attenuated seasonal patterns, even though the value is close to 1. Mixing is very good for clusters with high values of  $\rho_k^*$ , which suggests a limitation of the model construction, in practical application. As mentioned, the parameterization is overly coarse.

The above comparison between the three models encourages our selection of a model with FB to perform the by-county prediction of employment totals. Figures 5, 6, 7 compare use our predictor-assisted dependent Dirichlet process (DDP) clustering prior to use of a simple Dirichlet process prior (that excludes predictor information in the prior probabilities of co-clustering), both under the Fourier basis construction for seasonality. We label the latter model as “DP-FB”. We used all of the predictors, including the by-industry location quotients, the latitude and longitude (computed on population centroids) and level of unemployment insurance claims to influence the prior probabilities

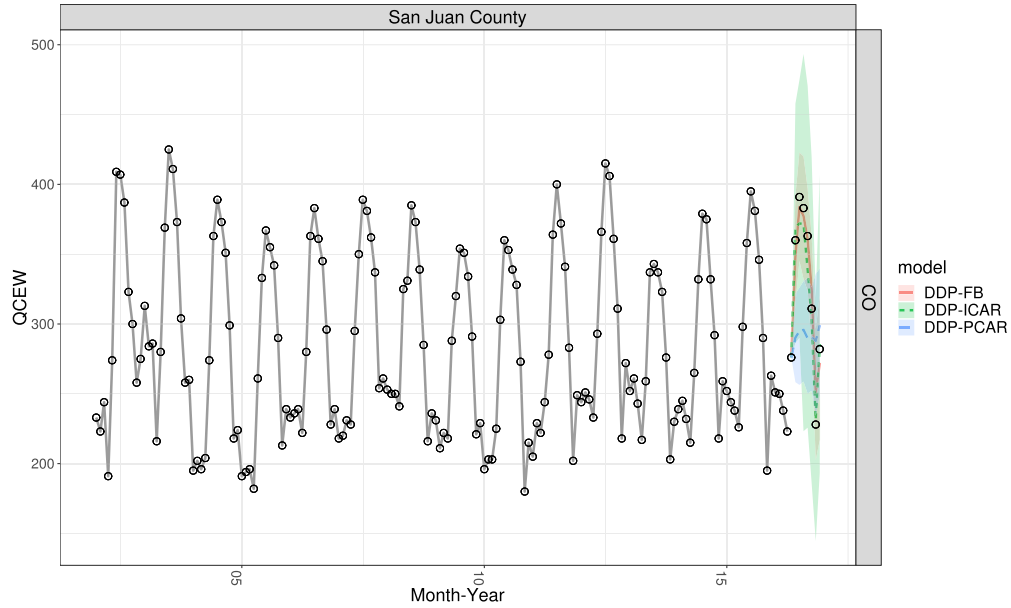


Figure 4: Comparing the prediction performance of the ICAR prior (green, short-dashed line) to a PCAR prior (blue, long-dashed line) and to the Fourier basis prior (pink, solid line) for a county where seasonality is *highly* expressed for the 7-month prediction period. The hollow circles represent the data points. The shading represents 95% credible intervals.

of co-clustering for the counties. The figures readily reveal that the DDP-FB uniformly outperforms the DP-FB, as confirmed by prediction error statistics in Table 1. It is interesting to note, however, that DP-FB outperforms DDP-ICAR, which highlights the importance of the Fourier basis construction for the prior on the seasonality term. We additionally tested all subsets of predictors for the DDP and found that including them all, together, performed best. It also bears mention that including only the spatial predictors in the DDP also reduced the prediction performance of DDP-FB, indicating that the non-spatial predictors are important to improve the clustering performance.

All of our models enumerated in Section 2 configure our data as a collection of time series that are spatially linked through our dependent Dirichlet process prior construction that is indexed by latitude and longitude (as well as other predictors). In other words, our models are parameterized as spatially-indexed time series. By contrast, we may also model these data as a time evolving spatial process by proposing a spatial model over counties that is allowed to evolve over time. We accomplish the latter modeling framework with,

$$y_{ij} | \mathbf{x}_{ij}^{p \times 1}, \beta_i, (\gamma_{mij})_{m=1, \dots, M} \stackrel{\text{ind}}{\sim} \mathcal{N} \left( \mathbf{x}_{ij}' \beta_i + \sum_{\ell=1}^{n_H} H_{ij\ell} \eta_{\ell j}, \tau_y^{-1} \right), \quad (26)$$

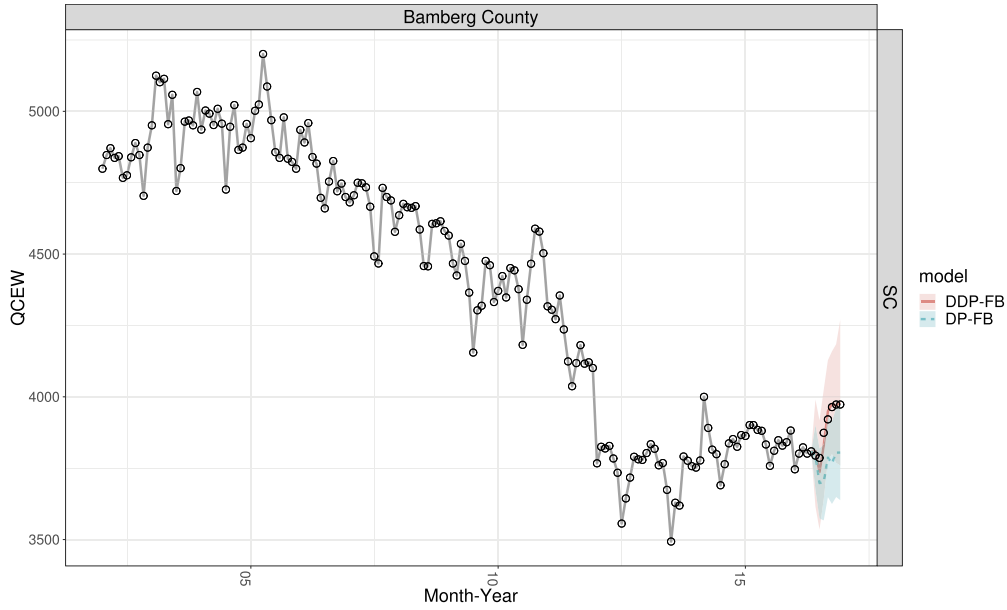


Figure 5: Comparing the prediction performance of predictor-assisted DDP clustering (pink, solid line) to a simple DP prior (turquoise, dashed line) plan for a county where seasonality is less expressed for the 7-month prediction period. The hollow circles represent the data points. The shading represents 95% credible intervals.

$$\boldsymbol{\eta}_j^{n_H \times 1} | \tau_\eta, \mathbf{H}_j, \Omega_\eta \stackrel{\text{ind}}{\sim} \mathcal{N}_{n_H} \left( \mathbf{0}, [\mathbf{H}_j' Q_\eta \mathbf{H}_j]^{-1} \right), \quad (27)$$

where we have replaced the time series random effects terms,  $\sum_{m=1}^M \gamma_{mij}$ , used in (1) with a spatial basis construction, where the  $n \times n_H$  basis matrix,  $\mathbf{H}_j = (H_{ij\ell})_{i,\ell}$ , encodes the spatial dependence among the counties. It is multiplied by  $n_H \times 1$  coefficients,  $\boldsymbol{\eta}_j$ . The  $\boldsymbol{\eta}_j$  are indexed by month,  $j \in 1, \dots, T$ , so they are allowed to evolve the spatial process over time. We next briefly introduce the construction for the basis matrix,  $\mathbf{H}_j$ , below, where each column represents a spatial contrast that we label as a “harmonic”; for example, the first two low resolution contrasts (from left-to-right) include north-south and east-west. Please see Hughes and Haran (2013) for a more detailed explanation. We construct and implement this model (also available in our R Core Team (2014) implementation) as an alternative to our approach.

We utilize the generalized Moran’s  $I$  operator for  $n \times p$  design matrix,  $\mathbf{X}_j$ , that projects an adjacency indexed spatial basis onto the space orthogonal to that defined by  $\mathbf{X}_j$  in order to avoid confounding and inflating estimates of posterior uncertainty (Hughes and Haran, 2013). The Moran’s  $I$  operator is defined on a graph,  $G = (V, E)$ , where  $V$  denotes the  $n$  vertices and  $E$  the edges between pairs,  $(i, i') \in (1, \dots, n)$  counties. Define the  $n \times n$  adjacency matrix,  $\Omega_\eta = (\omega_{ii'})$ , where  $\omega_{ii'} = \mathbf{1}(i, i' \in E, i \neq i')$ , with  $\text{diag}(\Omega_\eta) = \mathbf{0}$ . The associated spatial conditional autoregressive formulation for the

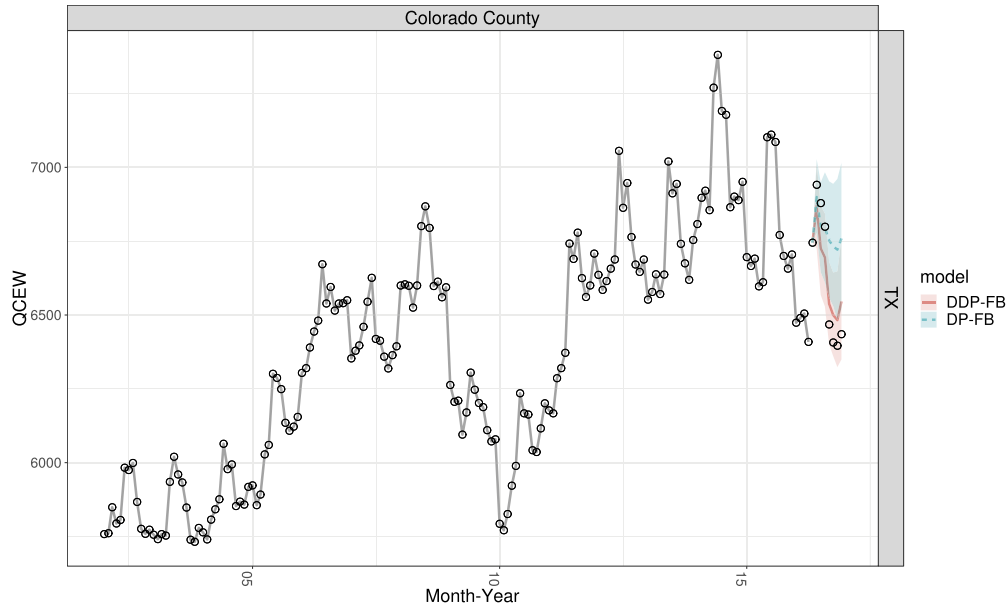


Figure 6: Comparing the prediction performance of predictor-assisted DDP clustering (pink, solid line) to a simple DP prior (turquoise, dashed line) plan for a county where seasonality is more expressed for the 7-month prediction period. The hollow circles represent the data points. The shading represents 95% credible intervals.

$n \times n$  precision matrix is accomplished with,  $Q_\eta = \Omega_\eta \mathbf{1} - \Omega_\eta$ , where each diagonal entry of  $Q_\eta$  counts the number of neighboring counties for that county. Construct  $\mathbf{P}_j = \mathbf{X}_j (\mathbf{X}_j' \mathbf{X}_j)^{-1} \mathbf{X}_j'$  to be the projection onto the space defined by  $\mathbf{X}_j$  and its orthogonal complement,  $P_j^\perp = \mathbb{I}_n - \mathbf{P}_j$ . Hughes and Haran (2013) formulate the generalized Moran's  $I$  operator with,

$$I_{X_j}(\Omega_\eta) = P_j^\perp \Omega_\eta P_j^\perp = \mathbf{H}_j' \Lambda \mathbf{H}_j \quad (28)$$

where the  $n \times n$  eigenmatrix,  $\mathbf{H}_j$ , is employed as a spatial basis matrix. The columns of  $\mathbf{H}_j$  are harmonics that express the frequencies of spatial association, with higher frequencies corresponding to rougher dependence structures. The frequency of the contrasts and resulting resolution of spatial smoothing increase from left-to-right in  $\mathbf{H}_j$ , so it is typical (Hughes and Haran, 2013) practice to select 10% of the columns to achieve a high quality estimation with substantial dimension reduction to facilitate computation. We follow this practice, so  $n_H = 0.1 \times n$  denotes the number of columns or harmonics used for estimation. We label this model as “MI-T” to denote that it is a Moran's  $I$  spatial model for the counties that is allowed to evolve over the months.

Figure 8 demonstrates that MI-T is both biased and less efficient in prediction relative to DDP-FB, even on this relatively low seasonality county where, perhaps, we would expect a model constructed as a time-evolving spatial process (like MI-T)

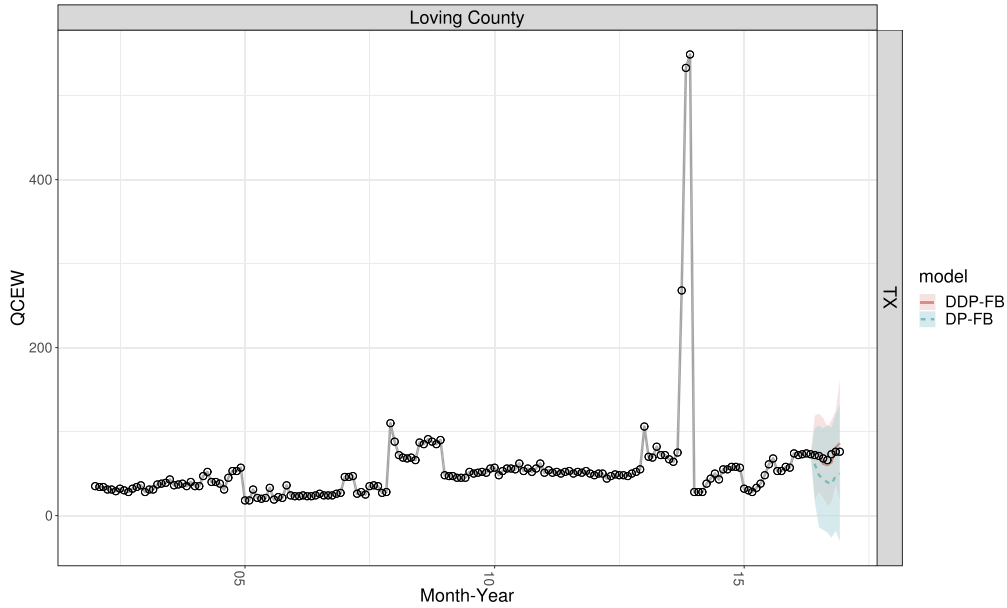


Figure 7: Comparing the prediction performance of predictor-assisted DDP clustering (pink, solid line) to a simple DP prior ((turquoise, dashed line)) plan for a tiny, idiosyncratic county for the 7-month prediction period. The hollow circles represent the data points.

to perform relatively well. We, additionally, configured models that employed both a Moran-I's basis,  $\sum_{\ell=1}^{n_H} H_{ij\ell} \eta_{\ell j}$  and the time series random effects terms,  $\sum_{m=1}^M \gamma_{mij}$  (including both seasonality and trend as one alternative and solely seasonality as another alternative). The fit of the expanded model was very similar to the models employing only time series terms and the prediction accuracy was worse.

Table 1 compares the root mean squared predictions errors for our suite of models. We see that DDP-FB performs the best on the predicted months, both for RMSPE computed on all 3108 counties and  $MAPE - C$ , computed on the 1751 CES counties. We recall that  $MAPE - C$  computes a percentage absolute error on each county and the score averages over the counties, so it treats relatively smaller counties as equivalently important to larger ones. Table 2 presents the distribution of  $MAPE - C$  by size of county for each of the DDP-FB and DDP-ICAR models. Firstly, we see how DDP-FB performs uniformly better across counties of different sizes. Secondly, we see how the two models converge as the county size category increases. So DDP-ICAR performs particularly badly on small counties, due to the instability in estimation of seasonality on low-seasonality counties, which are predominant among smaller-sized counties, as discussed above.

We propose to provide state analysts with the posterior predictive mean value for each county and month under the DDP-FB model as the predicted point estimate for

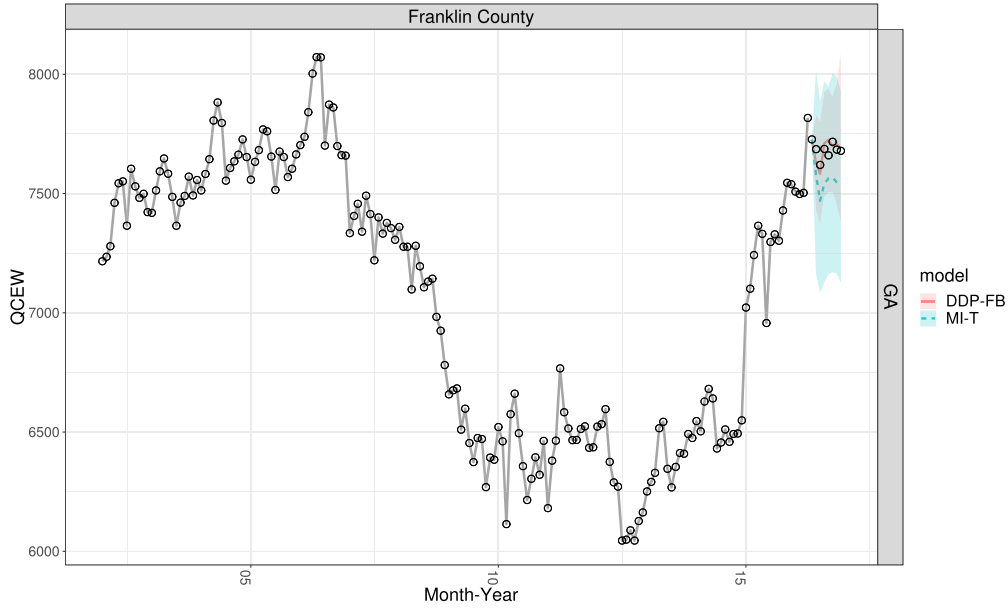


Figure 8: Comparing the prediction performance of predictor-assisted DDP clustering (pink, solid line) to a time varying Moran’s I spatial process (turquoise, dashed line) for the 7-month prediction period. The hollow circles represent the data points.

Model	RMSPE	MAPE-C
DDP-FB	919	1.29%
DDP-PCAR	1688	2.45%
DDP-ICAR	2103	2.71%
DP-FB	1570	2.11%
MI-t	2987	3.37%

Table 1: Root mean squared prediction errors (RMSPE) and mean absolute percentage error (MAPE-C) over the 7 predicted models across the  $n = 3108$  counties for comparison models.

each non-CES county. We additionally provide the associated 95% credible interval as an indicator of our confidence in each predicted value. The LAUS program work with the states to produce an unemployment rate estimate for each county and month using the employment prediction (in the denominator of the unemployment rate statistic). The estimate for the unemployment level (used in the numerator of the unemployment rate statistic) is derived by applying a county-to-state ratio to a state-level estimator (obtained from the Current Population Survey) in each month to obtain the county-level estimate of unemployment. An overall measure of uncertainty is not currently published for the resulting unemployment rate statistic due to this ratio estimation of the unemployment level for each county.

County Size		Models				
# Employees	# Counties	DDP-FB	DDP-ICAR	DDP-PCAR	DP-FB	MI-t
0–250	23	3.57%	11.95%	10.20%	8.41%	12.23%
250–500	48	2.90%	6.40%	5.32%	3.26%	6.51%
500–1000	116	2.30%	4.50%	3.95%	4.27%	6.31%
1000–2000	342	1.59%	3.12%	2.69%	2.39%	4.09%
2500–5000	392	1.24%	2.63%	2.45%	2.21%	3.01%
5000–10000	373	1.03%	2.17%	2.23%	1.68%	2.67%
10000+	457	0.8%	1.69%	1.70%	1.19%	2.21%

Table 2: Mean absolute percentage error distribution by county employment size category for comparator models.

## 5 Discussion

We constructed a suite of functional mixed effects models for an area-indexed collection of U.S. county time series that are designed to perform prediction for employment totals over multiple months. The last month of the prediction period is reported by the U.S. Bureau Labor Statistics (BLS) after consultation with the states. Our modeling framework is characterized by terms for trend, seasonality and the incorporation of predictors; in particular, measures of concentration across industry and spatial location centroids. These terms are specialized to county, but tied together with a predictor-dependent Dirichlet process partition model that allows the borrowing of strength among counties that express similar patterns in their time-indexed functions. We demonstrate that the incorporation of predictors into the estimation of partitions/clusters sharpens the prediction accuracy.

While there are many models in use in the U.S. Federal Statistical System that incorporate seasonality for a single time series, ours is among-the-first applications of a modeling approach for a collection of dependent time series where the degree of seasonality expressed varies among them. We have shown that use of a relatively simple Fourier basis under a flexible (Huang-Wand) prior for the covariance matrix of the county-indexed vector of coefficients performs robustly over both low- and high-seasonality counties. These strong results for our DDP-FB model suggest a modeling template for general use on collections of area-indexed time series. The semi-parametric framework allows for quick adaptation of model predictions during shifts in underlying economic conditions.

Future research will further explore opportunities to combine additional information during the prediction period, such as current month employment estimates for CES counties (since the CES survey estimates are available), into a model to predict employment for non-CES counties.

This case study application is particularly relevant because actual current month employment for U.S. counties is released to the public based on look-ahead predictions, rather than current month data (which are unavailable) and because this class of time- and spatially-indexed data is broadly representative of that published by government statistical agencies. It has been particularly challenging in this study to model a collection of time series under heterogenous patterns for seasonality and it is hoped that this

paper may catalyze the further development of Bayesian methods because they offer robust uncertainty quantification for smaller counties, which is important to data users.

## References

- Banerjee, S., Wall, M. M., and Carlin, B. P. (2003). “Frailty Modeling for Spatially Correlated Survival Data, with Application to Infant Mortality in Minnesota.” *Biostatistics (Oxford)*, 4(1): 123–142. 797
- Gelfand, A. E., Kottas, A., and Mac Eachern, S. N. (2005). “Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing.” *Journal of the American Statistical Association*, 100(471): 1021–1035. MR2201028. doi: <https://doi.org/10.1198/016214504000002078>. 800
- Hossain, M. M., Lawson, A. B., Cai, B., Choi, J., Liu, J., and Kirby, R. S. (2013). “Space-time stick-breaking processes for small area disease cluster estimation.” *Environmental and ecological statistics*, 20(1): 91–107. MR3039588. doi: <https://doi.org/10.1007/s10651-012-0209-0>. 802
- Huang, A. and Wand, M. P. (2013). “Simple Marginally Noninformative Prior Distributions for Covariance Matrices.” *Bayesian Analysis*, 8(2): 439–452. MR3066948. doi: <https://doi.org/10.1214/13-BA815>. 800
- Hughes, J. and Haran, M. (2013). “Dimension reduction and alleviation of confounding for spatial generalized linear mixed models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1): 139–159. MR3008275. doi: <https://doi.org/10.1111/j.1467-9868.2012.01041.x>. 799, 809, 810
- Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). “Fixed-width Output Analysis for Markov Chain Monte Carlo.” *Journal of the American Statistical Association*, 101(476): 1537–1547. MR2279478. doi: <https://doi.org/10.1198/016214506000000492>. 804
- Montesinos-López, A., Montesinos-López, O. A., de Los Campos, G., Crossa, J., Burgueño, J., and Luna-Vazquez, F. J. (2018). “Bayesian functional regression as an alternative statistical analysis of high-throughput phenotyping data of modern agriculture.” *Plant methods*, 14(1): 46. 798
- Mukherjee, C., Kasibhatla, P. S., and West, M. (2011). “Bayesian statistical modeling of spatially correlated error structure in atmospheric tracer inverse analysis.” *Atmospheric Chemistry and Physics*, 11(11): 5365–5382. URL <https://www.atmos-chem-phys.net/11/5365/2011/>. 797
- Müller, P., Quintana, F., and Rosner, G. L. (2011). “A Product Partition Model With Regression on Covariates.” *Journal of Computational and Graphical Statistics*, 20(1): 260–278. URL <http://www.tandfonline.com/loi/ucgs20>. MR2816548. doi: <https://doi.org/10.1198/jcgs.2011.09066>. 792, 801
- Neal, R. (2000a). “Slice Sampling.” *Annals of Statistics*, 31: 705–767. MR1994729. doi: <https://doi.org/10.1214/aos/1056562461>. 804



- Neal, R. M. (2000b). “Markov Chain Sampling Methods for Dirichlet Process Mixture Models.” *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: <https://doi.org/10.2307/1390653>. 802, 804
- Nobre, A. A., Sansó, B., and Schmidt, A. M. (2011). “Spatially Varying Autoregressive Processes.” *Technometrics*, 53(3): 310–321. URL <http://www.jstor.org/stable/23210406>. MR2857706. doi: <https://doi.org/10.1198/TECH.2011.10008>. 802
- Prado, R. and West, M. (2010). *Time series: modeling, computation, and inference*. CRC Press. MR2655202. 797
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>. 804, 809
- Rodríguez, A. and Dunson, D. B. (2011). “Nonparametric Bayesian Models through Probit Stick-breaking Processes.” *Bayesian Analysis*, 6(1): 145–178. MR2781811. doi: <https://doi.org/10.1214/11-BA605>. 802
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall Ltd/CRC. MR2130347. doi: <https://doi.org/10.1201/9780203492024>. 792, 796
- Sethuraman, J. (1994). “A Constructive Definition Of Dirichlet Priors.” *Statistica Sinica*, 4(2): 639–650. MR1309433. 800
- Wei, W. (2006). *Time Series Analysis Univariate and Multivariate Methods Second Edition*. Pearson Addison-Wesley, The United States of America. MR2517831. 798
- West, M. (2013). “Bayesian Dynamic Modelling.” In Damien, P., Dellaportas, P., Polson, N. G., and Stephens, D. A. (eds.), *Bayesian theory and applications*, chapter 8. Oxford, UK: Oxford University Press. MR3221162. 797