The Annals of Applied Statistics
2016, Vol. 10, No. 4, 2157–2181
DOI: 10.1214/16-AOAS968

BAYESIAN NONPARAMETRIC MULTIRESOLUTION ESTIMATION FOR THE AMERICAN COMMUNITY SURVEY

BY TERRANCE D. SAVITSKY

U.S. Bureau of Labor Statistics

Bayesian hierarchical methods implemented for small area estimation focus on reducing the noise variation in published government official statistics by borrowing information among dependent response values. Even the most flexible models confine parameters defined at the finest scale to link to each data observation in a one-to-one construction. We propose a Bayesian multiresolution formulation that utilizes an ensemble of observations at a variety of coarse scales in space and time to additively nest parameters we define at a finer scale, which serve as our focus for estimation. Our construction is motivated by and applied to the estimation of 1-year period employment totals, indexed by county, from statistics published at coarser areal domains and multi-year periods in the American Community Survey (ACS). We construct a nonparametric mixture of Gaussian processes as the prior on a set of regression coefficients of county-indexed latent functions over multiple survey years. We evaluate a modified Dirichlet process prior that incorporates county-year predictors as the mixing measure. Each county-year parameter of a latent function is estimated from multiple coarse-scale observations in space and time to which it links. The multiresolution formulation is evaluated on synthetic data and applied to the ACS.

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 across all states in the U.S. The local areas include counties and municipal civil divisions (MCDs), the latter of which are sub-county areas located in New England. The LAUS program utilizes estimated total employment for all counties and MCDs, published annually in the American Community survey, to construct a percentage of the total state employment for each local area. These ratios are applied to state-level employment totals, published monthly in the Current Population Survey (CPS), to render a synthetic (allocated) estimate of the employment totals for the local areas. This two-step process of estimating percentages-of-state total employment from the ACS for the local areas and their multiplication into CPS state-level estimates reflects the relatively small sample sizes and sparse coverage for local geographies in the CPS sample, which prevents stable (below some threshold for the coefficient of variation) employment total estimates for most local areas (directly from the CPS in one step). The ACS

Received June 2015; revised July 2016.
Key words and phrases. Survey sampling, small area estimation, latent models, Gaussian process, Dirichlet process, Bayesian hierarchical models, Markov chain Monte Carlo.

is a national survey conducted by the U.S. Census Bureau (Census) with dense geographic coverage and sufficiently large sample sizes to permit inference about local areas on many variables. The ACS provides annually updated estimates for many variables formerly only published in the decennial census long form.

Each ACS sample of households in a local area, such as a county, is realized under a sampling design that assigns probabilities of inclusion to all households in the sampling frame under a known sampling design distribution (which describes the distribution over the space of all possible samples). The ACS publishes sampling-weighted "direct estimates" of population-level quantities from many variables, such as the total employment, for each local area by summing over sampling-weighted household responses in the observed sample taken from that population to construct a (model-free) unbiased estimate for the entire local area. The ACS publishes a variance statistic (estimated with respect to the sampling design distribution in a nonmodel-based fashion) for each direct estimate of an area that may be interpreted as a measure of quality [see Särndal, Swensson and Wretman (1992)], and so the ACS publishes *both* a direct estimate and an associated variance for each area that are computed from the sampled household-level observations in that area. Särndal, Swensson and Wretman (1992) provide general approaches for computing the variance of the (population-level) direct estimate that is composed from the set of sampling-weighted household observations in a fashion that accounts for dependencies (e.g., under sampling without replacement) among household observations induced by the sampling design, and so the direct estimate is not an observation, but an estimate taken under the sampling design distribution from a collection of sampled household responses; therefore, we refer to the direct area estimates of total employment as "estimates" to distinguish them from (household) observations.

Employment (total) estimates are published at 1-, 3- and 5-year time periods (which we denote as "periods") for each of a wide variety of geographic domains. The longer time periods enable the collection and pooling of more household samples to improve the estimation precision expressed as a coefficient of variation (CV); hence, each period estimate corresponds to a single time interval computed from the total sample of households collected during that period. For example, if we fix an area and some time period, there is an unknown total employment for the *population* in that area in that time period we wish to estimate from a partial *sample* of households taken from that area. We don't observe the whole population, but only a sample of it. The population total estimate is constructed with the above-noted sampling-weighted summation using only those observations included in the sample. As the sample size grows, the information in the sample becomes progressively more representative of the population (and the variance of the sampling weights decrease). Under repeated sampling from the sampling design distribution, we would expect the variance of total employment estimates for the population in the focus area and chosen time period, estimated from the collection of repeated samples, to then decrease as the sample size (used for all of the repeated samples) increases. The Census determines which periods and area domains to publish estimates in the ACS based on the supporting population size in each area domain in order to ensure an acceptable CV; for example, 1-year period estimates are published for all area domains with populations $> 65{,}000$, while 3-year period estimates are provided for populations $> 20{,}000$ and 5-year period estimates are otherwise provided. A domain for which 1-year period estimates are published will also have published 3- and 5-year period estimates, while a domain for which 3-year period estimates are published will also have published 5-year period estimates. Most counties and MCDs in the U.S. are relatively small, such that only 26% of all *counties* have published ACS 1-year period estimates. In addition to pooling household observations across years into multi-year periods, the ACS also aggregates counties into larger geographic domains, such as metropolitan or micropolitan areas, to achieve a larger sample size that allows publication of 1-year period estimates.

The LAUS program is forced to use only the 5-year period estimates to compute the allocation proportions in order that employment total estimates be available for *all* counties and MCDs under the same (pooled) set of years. The multi-period ACS estimates are published each year; for example, the 5-year period estimate published in 2012 includes a pooled estimate for the period of 2008–2012, while the 5-year period estimate published in 2013 represents a pooled estimate for 2009–2013. Although new sample observations are added to the 5-year published estimates with each year, the resulting pooled, multi-year interval estimate is, however, lagged and possibly overly smoothed, which may result in a failure of the allocation proportion scheme to capture near-term changes in economic conditions, such as the recent Great Recession, that may dramatically alter the estimated proportions from one year to the next. The modeling approach we devise for this paper utilizes the ACS estimates published at multiple combined geographic and time period scales that are coarser than county-by-year to estimate latent, 1-year period employment totals for all counties and MCDs in the U.S. The LAUS program may then employ these model-based county-year employment total estimates to construct their local allocation proportions for all counties and MCDs in lieu of lagged, 5-year period ACS estimates.

Bayesian hierarchical modeling is extensively used in small area estimation applied to survey direct estimates published as official statistics by government agencies with the goal to reduce estimation uncertainty by borrowing information among parameters indexed by spatial area and often time period [Ghosh et al. (1998)]. The use of hierarchical modeling facilitates the borrowing of estimation strength by shrinking all or some subset of domain-period parameters to a common mean.

Even the most sophisticated small area modeling approaches, however, parameterize each regression mean to be linked one-to-one with an observed data point [Hawala and Lahiri (2012)]. These models may not be used to extract denoised, *single* year employment estimates for over 74% of those counties and MCDs that don't have available 1-year period ACS estimates. While the recent works of Bradley, Wikle and Holan (2014, 2015) appear to develop estimates for small domains from larger ones, they allocate or apportion larger domain estimates. [See the definition of $h(A, i) \equiv |A \cap B_i|/|A|$ just below Equation (5) on page 7 of Bradley, Wikle and Holan (2015), where B_i is on a finer grid than A for which there exists an estimate that is allocated to B_i.] They don't attempt to estimate latent values for finer areas nested within coarser ones that are viewed to generate the observed coarse estimates. Their prior is constructed to focus on inference and prediction under a spatial (adjacency-based) prior construction, while our approach intends to induce a more general data-adaptive dependence structure.

In the sequel, we construct parameters to be indexed on a fine scale and nest within one or more coarse-scale observations in space and time. A parameter is defined for each county-by-year to represent the total employment, and the set of by-year parameters for each county are, together, viewed as a function. There are typically multiple 1- and/or 3-year period (employment total) estimates published for coarser spatial domains. This collection of published employment total estimates from the ACS may be used to provide information about each county-by-year parameter based on the spatial nesting relationship of each county in a domain (such as a metropolitan area) and the inclusion (nesting) of each year in a multi-year period.

We devise a flexible nonparametric mixture formulation for estimation of regression coefficients used to construct a latent continuous function of years for county. The nonparametric mixture allows the data to shrink together the county-indexed functions with similar (though not exactly equal) by-year trends. This data-induced dimension reduction permits identification of the functions estimated from the coarser set of estimates that nest them. We refer to our approach as a "multiresolution" formulation because it utilizes observations defined at varied areal or time period resolutions for estimation of the fine-scale, by-county functions.

We specify the parameterization for our multiresolution likelihood in Section 2, followed by a set of priors that, along with the likelihood, construct a nonparametric clustering (mixture) model in Section 3. This model is extended into a predictor-assisted clustering formulation in Section 4. We present estimated results for the collection of county/MCD-year parameters from the ACS in Section 5. We perform a simulation study to assess the accuracy of the ACS estimates in Section 6 and offer a concluding discussion in Section 7. A brief overview of our algorithm to sample the set of full conditional posterior distributions defined by our model is discussed in a Technical Supplement [Savitsky (2016)].

2. Multiresolution likelihood. We begin by introducing the parameterization of the 1-year employment totals for counties and municipal civil division (MCD) domains and how they sum into multi-period and aggregate area employment total estimates to formulate our likelihood. We will subsequently introduce the nonparametric prior distributions that specify two alternative probability models, both of which employ this likelihood construction.

2.1. *Multiresolution parameterization.* In the sequel, we will use "county" as a generic label to denote county and MCD, the latter of which is a New England township designation where MCDs are nested within counties. Let $f_{\ell j}$ denote the (latent) employment total for $\ell = 1, \ldots, (N = 4751)$ counties over years, $j = 1, \ldots, (T = 5)$ (for the ACS published survey years of 2008, ..., 2012 that we use). The counties are geographically nested in larger core-based statistical areas (CBSAs), such as metropolitan (metro) and micropolitan (micro) areas, combinations of those larger areas (called core statistical areas or CSAs), including balance of states that subtract out all larger CBSAs and CSAs from each state. Larger states generally have both metro and micro areas, as well as larger combinations of these. (The Census defines all CBSAs and CSAs to fully nest within a state). Smaller states may have only one-to-a-few micro areas and no larger CSAs, other than the balance of the state estimate that subtracts away the micro areas. We denote all areas that geographically nest counties (which includes the counties themselves) by the term "group," $b = 1, \ldots, B$, and all counties geographically nest in one or more groups. We use published employment total estimates for $B = 6074$ ACS groups (that include the $N = 4751$ counties). We write that a county "links" to a group if it geographically nests in that group. The definitions of counties and groups remain fixed throughout the 2008–2012 time period so that the geographic county-to-group links are fixed. A single county may link to multiple groups because it may (geographically) nest within a group which is, in turn, nested within other groups. Figure 1 presents an example for Amesbury Town, Massachusetts, which links to 4 other groups through successive nestings. Approximately 75% of the N counties link to 3–5 groups (including themselves), and the remaining 25% link to 6–7 groups. We have so far defined how a county, ℓ, links to or nests within larger geographic groups, b. By extension, a 1-year period (e.g., 2008) is indexed by j and is linked to or nests within multi-year period, q. We index the multi-year periods by $q = 1, \ldots, Q$, where each index value links a particular set of years. Table 1 presents the set of years, $j = 1, \ldots, (T = 5)$ (indexing the columns), that link with each period (row), q, where 1 denotes a link and 0 not.

We construct a likelihood statement for each group-period estimate published in the ACS data, y_{bq}, by summing over parameters indexed by counties, (ℓ), that nest in group, b, and years, (j), that nest (are included) in associated period, q, with

$$y_{bq} \mid (f_{\ell j})_{\ell \in b, j \in q} \overset{\text{ind}}{\sim} \mathcal{N}\left(\sum_{\ell \in b} \sum_{j \in q} f_{\ell j}, \sigma_{bq}^2\right), \tag{1}$$

$$f_{\ell j} = \mathbf{x}_{\ell j}' \boldsymbol{\beta}_{\ell j}, \tag{2}$$

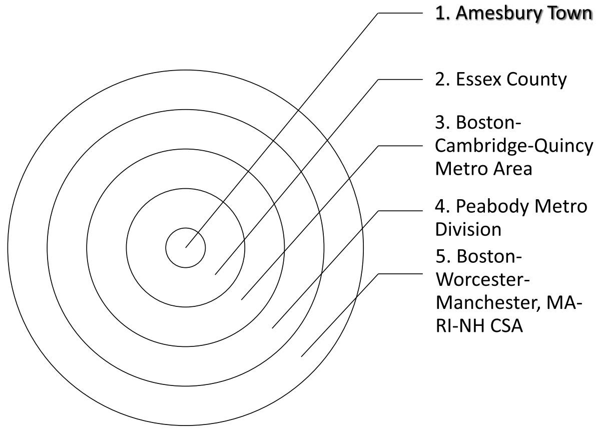

FIG. 1. *Example of group nesting structure for Amesbury Town, Massachusetts.*

where the associated group-period sampling variances (of the direct estimate), $\{\sigma_{bq}^2\}$, are *known* because they are published by the Census in the ACS, as discussed in Section 1. Since y_{bq} is the employment total estimate from the ACS for group, b, and period, q, the $(f_{\ell j})$ in Equation (1) are defined to represent latent employment totals for those counties, ℓ, and years, j, nested in (group, period), (b, q). Equation (1) constructs the (latent) mean of y_{bq} by summing the latent county-year employment totals, $(f_{\ell j})$, that jointly nest within group, b, and period, q. A $P \times 1$ set of predictors, $\mathbf{x}_{\ell j}$, defined at the *county-year* (fine) scale, is incorporated into the model for the function, $f_{\ell j}$, with associated $P \times 1$ coefficients, $\boldsymbol{\beta}_{\ell j}$. We construct the county-year predictors, $\mathbf{x}_{\ell j}$, with an intercept and a set of predictors available for all counties ($\ell = 1, \dots, N$) and 1-year periods

TABLE 1
Period, $q = 1, \dots, (Q = 9)$ *links to years*, $j = 1, \dots, (T = 5)$ (*representing* 2008–2012 *calendar years*)

	Year				
Period, *q*	**2008**	**2009**	**2010**	**2011**	**2012**
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	1	1	0	0
7	0	1	1	1	0
8	0	0	1	1	1
9	1	1	1	1	1

($j = 2008, \dots, 2012$) from administrative data. The Quarterly Census of Employment and Wages (QCEW) is a census instrument targeted to business establishments (rather than households targeted by the ACS) that collects employment totals (on a monthly basis), which we aggregate to county and year. Our QCEW county-year predictors are employment totals for 12 "super sectors" defined in the North American Industry Classification System (NAICS): 1. Agricultural; 2. Natural resources and mining; 3. Construction; 4. Manufacturing; 5. Trade, transportation, utilities; 6. Information; 7. Financial activities; 8. Professional and business services; 9. Leisure and hospitality; 10. Other services; 11. Public administration; 12. Unclassified. We intend these 12 predictors, together, to describe the composition of the economic activity for each county by year. We also include state records of individual unemployment claims aggregated to counties in our predictor set as a measure of economic health. Our predictors will be critical to identify the regression coefficients and to regulate the borrowing of information for their estimation (through shrinkage). QCEW predictors are published at the county-year level, though disclosure limitations cause some values to be suppressed for small counties. Those values were available to us, however, in our analysis (though they may be imputed).

We believe that similarities in the trends of employment (and unemployment) totals among counties are generally not adjacency induced, but driven by underlying similarities in the economies between counties; for example, a rural county directly abutting or near to an urban area would be expected to express very different employment total trends, while yet being similar to a rural county in another state. Perhaps we may see a higher adjacency-based localized dependence as the area resolution decreases (e.g., for census tracts), though our focus is on counties. Our likelihood construction that estimates a latent by-county total using coarser areas in which it nests is an alternative approach to capturing spatial association, but one that may be more flexible than adjacency. Two counties that nest in the same group (e.g., a metropolitan area) are not required to have a similar size for total employment, only that their sum be coherent with the response value for each larger group in which they together nest.

The likelihood of Equation (1) sums the latent county-year employment totals, $(f_{\ell j})$, nested in each group-period estimate, y_{bq}. Conversely, there are multiple employment total estimates from the American Community Survey (indexed by group-period) that link to each county-year parameter and provide some information to support the estimation of that parameter. Fix a county, "ℓ," linked to a group, "b," where group b, in turn, includes published observations for 3- and 5-year periods, but is not large enough (in population) to include 1-year period estimates. Suppose we are interested to recover the published employment total estimates for group b that are linked to the ℓ-2010 county-year. The rows labeled 6–9 in Table 1 show that there are 4 available ACS employment total estimates for group b, which provide information for 2010 to estimate $f_{\ell 2010}$. There will potentially be many published estimates used to estimate $f_{\ell 2010}$ in the case that it nests in multiple groups. The implication is that each $f_{\ell j}$ may appear in the likelihood statements for multiple (y_{bq}). The (y_{bq}) are drawn as conditionally independent in Equation (1), given the collection of county-year latent function parameters, $(f_{\ell j})$, because the dependence among the (y_{bq}) is encoded by shared $(f_{\ell j})$ under our model formulation so that unconditionally [e.g., by marginalizing over the $(f_{\ell j})$] the (y_{bq}) are dependent.

We next define the prior distributions that permit flexibility in the borrowing of information for shrinkage in the estimation of the county-year regression coefficients.

3. Clustering model, $(Y|X)$. We next specify the probability model as the likelihood and set of prior distributions for our conditional regression and follow by discussing the selected prior formulations:

Likelihood

$$y_{bq}|(f_{\ell j})_{\ell \in b, j \in q} \overset{\text{ind}}{\sim} \mathcal{N}\left(\sum_{\ell \in b}\sum_{j \in q} f_{\ell j}, \sigma_{bq}^2\right). \tag{3}$$

Regression formulation

$$f_{\ell j} = \mathbf{x}'_{\ell j}\boldsymbol{\beta}_{\ell j}. \tag{4}$$

Conditional prior on coefficients

$$\overset{P\times T}{\mathbf{B}_\ell} = (\overset{P\times 1}{\boldsymbol{\beta}_{\ell 1}}, \dots, \boldsymbol{\beta}_{\ell T}) \overset{\text{ind}}{\sim} \mathbf{0} + \mathcal{N}_{P\times T}(\overset{P\times P}{\boldsymbol{\Lambda}_{y,\ell}^{-1}}, \mathbf{C}(\boldsymbol{\kappa}_\ell)). \tag{5}$$

Gaussian process covariance function

$$\overset{T\times T}{\mathbf{C}(\boldsymbol{\kappa}_\ell)} \equiv \mathbf{C}_\ell = (C_{\beta_{\ell pj},\beta_{\ell pk}})_{j,k\in(2008,\dots,2012)}, \qquad \ell = 1,\dots,N, \tag{6a}$$

$$C_{\beta_{\ell pj},\beta_{\ell pk}} = \frac{1}{\kappa_{\ell,1}}\left(1 + \frac{(t_{ij}-t_{ik})^2}{\kappa_{\ell,2}\kappa_{\ell,3}}\right)^{-\kappa_{\ell,3}}. \tag{6b}$$

Dirichlet process (DP) prior for covariance parameters

$$\boldsymbol{\Theta}_\ell = \{\boldsymbol{\Lambda}_{y,\ell}, \boldsymbol{\kappa}_\ell = (\kappa_{\ell,1},\kappa_{\ell,2},\kappa_{\ell,3})\}, \tag{7a}$$

$$\boldsymbol{\Theta}_1,\dots,\boldsymbol{\Theta}_N|G \sim G, \tag{7b}$$

$$G|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0), \tag{7c}$$

$$G_0 = \mathcal{W}(\boldsymbol{\Lambda}_{y,\ell}|P+1, \mathbb{I}_P) \times \prod_{d=1}^{D=3} \mathcal{G}a(\kappa_{\ell,d}|a,b). \tag{7d}$$

3.1. *Prior on functions.* We collect the $P \times T$ matrix of coefficients, $\mathbf{B}_\ell = (\boldsymbol{\beta}_{\ell 1}, \ldots, \boldsymbol{\beta}_{\ell T})$, indexed by county, $\ell = 1, \ldots, N$, on which we impose a conditional matrix variate Gaussian prior in Equation (5), using the notation of Dawid (1981), where the $P \times P$, $\boldsymbol{\Lambda}_{y,\ell}$, represents the precision matrix for the set of $P \times 1$ columns of $\mathbf{B}_\ell$ and the $T \times T$, $\mathbf{C}(\boldsymbol{\kappa}_\ell)$, denotes the covariance matrix for the rows of $\mathbf{B}_\ell$. The $P \times T$ mean matrix of zeros for $\mathbf{B}_\ell$ is denoted by $\mathbf{0}$. The county-indexed covariance matrix, $\mathbf{C}_\ell$, is parameterized by $\boldsymbol{\kappa}_\ell$. This specification is equivalent to the $TP \times TP$ covariance matrix constructed as $\boldsymbol{\Lambda}_{y,\ell}^{-1} \otimes \mathbf{C}(\boldsymbol{\kappa}_\ell)$ under a multivariate Gaussian prior on the vector obtained by stacking the rows of $\mathbf{B}_\ell$. The separable or tensor form we use for the covariance matrix reflects parsimony relative to a general $TP \times TP$ covariance matrix.

We fix a particular county, ℓ, and introduce a Gaussian process covariance formulation in Equation (6) that we construct for each of the $P, T \times 1$ rows of $\mathbf{B}_\ell = (\boldsymbol{\beta}_{\ell 1}, \ldots, \boldsymbol{\beta}_{\ell T})'$ in Equation (5). The parameters, $\boldsymbol{\kappa}_\ell = (\kappa_{\ell,1}, \kappa_{\ell,2}, \kappa_{\ell,3})$, are used to specify a rational quadratic covariance formula in Equation (6b) for each cell of $\mathbf{C}(\boldsymbol{\kappa}_\ell)$. The vertical magnitude of surfaces rendered from a GP with the rational quadratic covariance formula is directly controlled by $\kappa_{\ell,1}$, while $\kappa_{\ell,2}$ controls the mean length scale or period, and $\kappa_{\ell,3}$ controls smooth deviations from the mean length scale, which allows estimation of local, as well as global, features in the time-indexed functions, $(\mathbf{f}_\ell)$ [Rasmussen and Williams (2006)]; see Savitsky, Vannucci and Sha (2011) for more background on the Gaussian process covariance formulations. The $P \times P$ precision matrix, $\boldsymbol{\Lambda}_{y,\ell}$, allows the data to estimate a dependence among the P sets of $T \times 1$ functions, each drawn from the Gaussian process.

3.2. *Clustering the distributions of the coefficients*, $(\mathbf{B}_\ell)$. We specify a Dirichlet process (DP) prior for $(\boldsymbol{\Theta}_\ell)$ in Equations (7), where $(\boldsymbol{\Theta}_\ell)_{\ell=1,\ldots,N}$ receive a random distribution prior, G, drawn from a Dirichlet process (DP), parameterized with a concentration parameter, α, a precision parameter that controls the amount of variation in G around prior mean G_0. The base or mean distribution $G_0 = \mathcal{W}(P+1, \mathbb{I}_P) \times \prod_{d=1}^{D=3} \mathcal{G}a(a,b)$, a P-dimensional Wishart distribution for the $P \times P$, $\boldsymbol{\Lambda}_{y,\ell}$, and a product of Gamma priors for the $D = 3$ parameters in the rational quadratic specification for the parameter vector, $\boldsymbol{\kappa}_\ell$, that parameterize the $T \times T$ covariance matrix, $\mathbf{C}$, respectively.

Equations (5) and (7) together define a marginal mixture prior of the form $\mathbf{B}|G \overset{\text{i.i.d.}}{\sim} \int \mathbf{0} + \mathcal{N}_{P \times T}(\boldsymbol{\Lambda}_y, \mathbf{C}(\boldsymbol{\kappa}))G(d(\boldsymbol{\Lambda}_y, \boldsymbol{\kappa}))$, where G is the mixing measure. The DP prior imposed on $\boldsymbol{\Theta}_\ell$ allows the data to estimate probabilistic clusters such that those counties, (ℓ), whose $(\boldsymbol{\Theta}_\ell)$ are assigned to the same cluster will draw their coefficients, $(\mathbf{B}_\ell)$, from the same Gaussian mixture component. Drawing coefficients under the same mixture component offers more flexility than directly clustering the $(\mathbf{B}_\ell)$, which we do because we don't expect any of the coefficients [and associated $T \times 1$ functions, $(\mathbf{f}_\ell)$] to be exactly equal. Rather, we expect subsets of functions to be "similar," which we define as drawing their coefficients (assigned to the same cluster) from the same Gaussian distribution.

We write the unknown measure G constructively in the (stick-breaking) form as a set of weighted locations [Sethuraman (1994)],

$$G = \sum_{h=1}^{\infty} p_h \delta_{\boldsymbol{\Theta}_h^*}, \tag{8}$$

where G is a countably infinite mixture of weighted point masses (or "spikes") with "locations," $\boldsymbol{\Theta}_1^*, \ldots, \boldsymbol{\Theta}_M^*$, in the support of G indexing the unique values for the $(\boldsymbol{\Theta}_\ell)$, where $M \leq N$ (the total number of counties from the finite population). We record cluster memberships of counties with $\mathbf{s} = (s_1, \ldots, s_N)$, where $s_\ell = \ell$ denotes $\boldsymbol{\Theta}_\ell = \boldsymbol{\Theta}_\ell^*$ so that $\{\mathbf{s}, (\boldsymbol{\Theta}_m^*)\}$ provides an equivalent parameterization to $(\boldsymbol{\Theta}_\ell)$ and we recover $\boldsymbol{\Theta}_\ell = \boldsymbol{\Theta}_{s_\ell}^*$. The weight $p_h \in (0, 1)$ is composed as $p_h = v_h \prod_{k=1}^{h-1}(1 - v_k)$, where v_h 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 α will produce more clusters (unique locations). Since each location is drawn from G_0, as the number of unique locations increases, the estimated G approaches the base distribution G_0. We place a further gamma prior on α to allow posterior updating in recognition of the relatively strong influence it conveys on the number of clusters formed [Escobar and West (1995)].

Although we believe county-indexed dependence for total employment is not generally adjacency-related, we, nevertheless, tested inclusion of a spatial (adjacency) random effects term in our models, but found it induced over-smoothing and deteriorated the quality of the fitted, $(f_{\ell j})$. Our likelihood in Equation (1) *does*, however, encode spatial information by nesting the county-year parameters in groups and periods. Additionally, the existence of a true local dependence between any two counties, (ℓ, ℓ'), that is not explicitly modeled may, nevertheless, produce a high posterior probability for their sharing of a common mixture component (cluster) under the DP mixture prior on $(\mathbf{B}_\ell)$.

4. Predictor-assisted clustering model, (Y, X). Our estimation task is challenging because the latent county-year employment totals, $(f_{\ell j})$, are at a finer resolution that than the group-period indexed ACS estimates, (y_{bq}), and so we would like to borrow the maximum amount of information provided in our data by incorporating the predictor values into the prior probabilities for the co-clustering of the county covariance parameters of $(\mathbf{B}_\ell)$. If the $P \times T$ matrix of predictors, $\mathbf{X}_\ell$, for county, ℓ, is very similar to $\mathbf{X}_{\ell'}$ for county, ℓ', then we would like to define a higher *prior* probability for $\boldsymbol{\Theta}_\ell = \boldsymbol{\Theta}'_\ell = \boldsymbol{\Theta}_m^*$.

We modify an approach of Müller, Quintana and Rosner (2011) to allow the definition of a DP prior construction that incorporates the predictors, $(\mathbf{X}_\ell)_{\ell=1,\ldots,N}$, into the clustering prior distribution. We will treat the $P \times T$ predictor matrices,

$(\mathbf{X}_1, \dots, \mathbf{X}_N)$, as though they were random (though we believe they are *not* random) as a computational device to induce the utilization of the predictors in the construction of the prior probability for the $N \times 1$ cluster assignments, $\mathbf{s}$. We next construct the details of how to incorporate the $(\mathbf{X}_\ell)$ into the prior for cluster assignments by constructing a model for the $(\mathbf{X}_\ell)$ that treats them as random (even though they are fixed) as a computation device.

We specify a probability model for the joint likelihood $(\mathbf{Y}, (\mathbf{X}_\ell))$ to include values for the predictors in the determination of cluster assignments:

Conditional likelihood for $(\mathbf{Y}|(\mathbf{X}_\ell))$

$$y_{bq}|(\mathbf{x}_{\ell j}, \boldsymbol{\beta}_{\ell j})_{\ell \in b, j \in q} \overset{\text{ind}}{\sim} \mathcal{N}\biggl(\sum_{\ell \in b} \sum_{j \in q} \mathbf{x}'_{\ell j} \boldsymbol{\beta}_{\ell j}, \sigma^2_{bq}\biggr). \tag{9}$$

Mixture formulation for $\mathbf{X}_\ell$ *to induce predictor-dependent clustering prior*

$$\overset{P \times 1}{\mathbf{x}_{\ell j}} \overset{\text{ind}}{\sim} \mathbf{N}_P(\boldsymbol{\delta}_{\ell j}, \mathbf{H}_x^{-1}), \tag{10a}$$

$$\overset{P \times T}{\boldsymbol{\Delta}_\ell} = (\boldsymbol{\delta}_{\ell 1}, \dots, \boldsymbol{\delta}_{\ell T}), \tag{10b}$$

$$\boldsymbol{\Delta}_\ell \overset{\text{ind}}{\sim} \mathbf{0} + \mathcal{N}_{P \times T}\bigl(\overset{P \times P}{\boldsymbol{\Lambda}^{-1}_{x,\ell}}, \overset{T \times T}{\mathbf{Q}}(x, \ell)^{-1}\bigr), \tag{10c}$$

$$\mathbf{H}_x \sim \mathcal{W}(P + 1, \mathbb{I}_P), \tag{10d}$$

$$\overset{T \times T}{\mathbf{Q}}(x, \ell) = \tau_{x,\ell}(\mathbf{D}_x - \rho_{x,\ell} \boldsymbol{\Omega}_{x,\ell}). \tag{10e}$$

Mixture formulation for regression coefficients, $\mathbf{B}_\ell$*, in model for* $(\mathbf{Y}|(\mathbf{X}_\ell))$

$$\overset{P \times T}{\mathbf{B}_\ell} \overset{\text{ind}}{\sim} \mathbf{0} + \mathcal{N}_{P \times T}\bigl(\overset{P \times P}{\boldsymbol{\Lambda}^{-1}_{y,\ell}}, \mathbf{C}(\boldsymbol{\kappa}_\ell)\bigr). \tag{11}$$

Dirichlet process (DP) prior for covariance parameters of $\mathbf{Y}$ *that now includes* $(\mathbf{X}_\ell)$

$$\boldsymbol{\Theta}_\ell = \{\boldsymbol{\Lambda}_{y,\ell}, \boldsymbol{\kappa}_\ell, \boldsymbol{\Lambda}_{x,\ell}, \tau_{x,\ell}, \rho_{x,\ell}\}, \tag{12a}$$

$$\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_N | G \sim G, \tag{12b}$$

$$G|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0), \tag{12c}$$

$$\begin{aligned} G_0 = {} & \mathcal{W}(\boldsymbol{\Lambda}_{y,\ell}|P + 1, \mathbb{I}_P) \\ & \times \mathcal{W}(\boldsymbol{\Lambda}_{x,\ell}|P + 1, \mathbb{I}_P) \\ & \times \prod_{d=1}^{D=3} \mathcal{G}a(\kappa_{\ell,d}|a, b) \times \mathcal{G}a(\tau_{x,\ell}|a, b) \times \mathcal{U}(\rho_{x,\ell}|-1, 1). \end{aligned} \tag{12d}$$

We have expanded our probability model of Section 3 to now include a model for the predictors. The precision matrix, $\mathbf{Q}(x, \ell)$, for the prior on the $T \times 1$ rows of the coefficient matrices, $(\boldsymbol{\Delta}_\ell)$, of the model for $(\mathbf{x}_{\ell j})$ in Equation (10e) is constructed as conditional autoregressive (CAR) [Rue and Held (2005)] that is similar in idea to the GP prior on $\mathbf{B}_\ell$, but tends to render rough, nondifferentiable surfaces, rather than the smooth surfaces generated by a GP prior. We use the CAR prior because it is computationally faster to draw posterior samples than the GP, and we are not concerned with generating denoised functions from $\mathbf{X}_\ell$, but rather to model $\mathbf{X}_\ell$ as a computational device for inserting information about predictors into the prior probability of co-clustering. The CAR precision matrix is composed with a $T \times T$, $\mathbf{D}_x$, a diagonal matrix that sums the rows of the $T \times T$, $\boldsymbol{\Omega}_x$, a similarity or adjacency matrix between pairs of time points (with zeros for the diagonal values), and so each entry in $\mathbf{D}_x$ expresses the relative influence or precision for each time point. The parameter $\tau_{x,\ell} \sim \mathcal{G}a(a = 1, b = 1)$, which we set to be *a priori* weakly informative, controls the scale and $\rho_{x,\ell} \sim \mathcal{U}(-1, 1)$ controls the degree of autocorrelation. The CAR prior may be heuristically thought of as a local, random walk smoother with a fixed length scale (unlike the GP, where the data estimate the length scale); see Savitsky and Paddock (2013) for more details about the CAR prior.

The construction of the parameter vector, $\boldsymbol{\Theta}_\ell$, on which the Dirichlet process prior is imposed in Equation (12) to induce a prior over partitions or clusterings is expanded to now *include* parameters from the model for predictors $\mathbf{X}_\ell$. We next show that treating $(\mathbf{X}_\ell)$ as random inserts them into the prior for cluster assignments, $\mathbf{s}$. The covariance parameters for the predictors, $\{\boldsymbol{\Lambda}_{x,\ell}, \tau_{x,\ell}, \rho_{x,\ell}\}$, are now included in an augmented $\boldsymbol{\Theta}_\ell = \{\boldsymbol{\Lambda}_{y,\ell}, \boldsymbol{\kappa}_\ell, \boldsymbol{\Lambda}_{x,\ell}, \tau_{x,\ell}, \rho_{x,\ell}\}$ under the DP prior of Equations (7) to incorporate information about $\mathbf{X}_\ell$ into the prior over clusterings. We can see how the predictors under Equations (9)–(12) influence cluster assignments by examining the conditional prior distributions for cluster assignments, $\mathbf{s} = (s_1, \dots, s_N)$, after marginalizing over the random measure G:

$$f\big(s_\ell = s|\mathbf{s}_{-i}, \boldsymbol{\Theta}^*_s, \alpha, \mathbf{B}_\ell, \boldsymbol{\Delta}_\ell\big) \propto \begin{cases} \dfrac{n_{-\ell,s}}{n-1+\alpha} L(\boldsymbol{\Delta}_\ell) & \text{if } 1 \le s \le M^-, \\ \dfrac{\alpha/c^*}{n-1+\alpha} L(\boldsymbol{\Delta}_\ell) & \text{if } s = M^- + h, \end{cases} \tag{13}$$

where the $(\boldsymbol{\Delta}_\ell)$ convey the information about the $(\mathbf{X}_\ell)$ in

$$L(\boldsymbol{\Delta}_\ell) = \prod_{j=1}^{T} \mathcal{N}_P\big(\mathbf{x}_{\ell j}|\boldsymbol{\delta}_{\ell j}, \mathbf{H}_x^{-1}\big) \times \mathcal{N}_{P\times T}\big(\boldsymbol{\Delta}_\ell|\boldsymbol{\Lambda}^*_{x,s}, \mathbf{Q}\big(\tau^*_{x,s}, \rho^*_{x,s}\big)\big),$$

where we recall from Section 3 that the $*$ superscript indexing the covariance/precision parameters indicates they are the unique locations where we sample $\{\mathbf{s}, (\boldsymbol{\Theta}^*_m)_{m=1,\dots,M}\}$. The terms involving $\boldsymbol{\Delta}_\ell$ reflect the insertion of $\mathbf{X}_\ell$ into the DP *prior* because the $\mathbf{X}_\ell$ are *fixed*, not random. Information about $\mathbf{X}_\ell$ is solely used in the prior for cluster assignments, $\mathbf{s}$, and not in the model for regression coefficients, $(\mathbf{B}_\ell)$, and so we otherwise discard $\mathbf{X}_\ell$ and focus our inference on the conditional distribution for $\mathbf{Y}|(\mathbf{X}_\ell)$, but with the prior distribution for cluster assignments now updated to include predictor information.

We have incorporated information about the predictors in determination of cluster assignments by treating them as random under an augmented version of our DP scale mixture formulation. We do *not*, however, believe the predictors, $(\mathbf{X}_\ell)$, are random, and we employ the probability model for $(\mathbf{X}_\ell)$ solely as a computational device to create a new prior for cluster assignments that incorporates the predictor values in a predictor-dependent Dirichlet process mixture formulation. We zoom out from the detailed modeling, discussed above, for incorporating predictors $(\mathbf{X}_\ell)$ into the prior for assignment to clusters and now show an equivalent summary comparison for the joint prior for cluster assignments, $\mathbf{s}$, excluding and including the predictors.

The joint prior cluster assignments under the clustering prior of Equation (7) is stated with

$$f(s_1, \dots, s_N) \propto \alpha^{M-1} \prod_{m=1}^{M} (n_m - 1)!, \tag{14}$$

after marginalizing out the random measure G, where $n_m = \sum_{\ell=1}^{N} \mathbb{I}(s_\ell = m)$ denotes the number of counties assigned to cluster m. The predictor-assisted approach, which parameterizes a joint distribution for $\mathbf{Y}$, $(\mathbf{X}_\ell)_{\ell=1,\dots,N}$, adjusts Equation (14) to *add* information about the predictors with

$$f(s_1, \dots, s_N | \overset{P\times T}{\mathbf{X}_1}, \dots, \mathbf{X}_N) \propto \alpha^{M-1} \prod_{m=1}^{M} g(\mathbf{X}_m^*)(n_m - 1)!, \tag{15}$$

where our notation conditions on the $(\mathbf{X}_\ell)$ for emphasis, though this prior *doesn't* treat them as random. They are fixed. We define

$$g(\mathbf{X}_m^*) = \int \prod_{\ell : s_\ell = m} f(\mathbf{X}_\ell | \boldsymbol{\Delta}_\ell, \mathbf{H}_x) f(\boldsymbol{\Delta}_\ell | \boldsymbol{\Theta}_{x,m}^*) f(\boldsymbol{\Theta}_{x,m}^*) \, d\boldsymbol{\Delta}_\ell \, d\boldsymbol{\Theta}_{x,m}^* \, d\mathbf{H}_x, \tag{16}$$

where $\boldsymbol{\Theta}_{x,m}^* = \{\boldsymbol{\Lambda}_{x,m}^*, \tau_{x,m}^*, \rho_{x,m}^*\}$ are the unique locations. The integration is performed numerically in our MCMC procedure. Equation (15), with $(\mathbf{X}_m^*)$ defined in Equation (16), is directly derived from the conditional formulation in Equation (13) for any vector of cluster assignments $\mathbf{s}$. We have only modeled the $\mathbf{X}_\ell$ as a computational device for inserting the predictors into the prior where two establishments, ℓ and ℓ', with similar values of $\mathbf{X}_\ell$ and $\mathbf{X}_{\ell'}$ are more likely to co-cluster *a priori* because the $(\mathbf{X}_\ell)$ are fixed. What is especially satisfying is that we have estimated a complicated predictor-dependent Dirichlet process mixture model with a much simpler Dirichlet process model through imposing a probability model on the 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, and so care is warranted in the choice of which and how many predictors to use for indexing the prior for cluster assignments.

We ran our sampling chains for 30,000 iterations, discarding half as burn-in, and thinning each chain by saving every 20th draw. Convergence of the sampler was 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)]. We further computed the scale reduction factor, $\hat{R}$, of Gelman and Rubin (1992), which is composed as the square root of the ratio of a weighted average of between-chain and within-chain variances divided by the within-chain variance. The statistic of Gelman and Rubin (1992) is computed from multiple chains for each parameter of interest. We achieve a value of $\hat{R}$ very near to 1, which indicates convergence. Our general experience is that the fixed width estimator produces stopped chains about twice as long as the Gelman and Rubin (1992) statistic; please see the Technical Supplement in Savitsky (2016) for details of the posterior sampling schemes used for both models and the resulting mixing performance.

5. Results for the ACS. We next illustrate estimation results for our model by comparing the fitted function for a selected county with the collection of ACS estimates to which it is linked at each time point. To make the comparison meaningful, we only want to include the portion of each published estimate that provides information about that county; for example, if a county is nested, along with other counties, in a metropolitan area for which we have an ACS estimate, y_{bq}, we'd like to extract from the estimate only the portion of the observed employment total that provides information about that county. We compute a "pseudo" statistic, $\tilde{y}_{bq,\ell j}$, in Equation (17) for each group, b, and period, q, linked to a latent, county-year function parameter, $f_{\ell j} = \mathbf{x}'_{\ell j}\boldsymbol{\beta}_{\ell j}$, by subtracting away from estimate y_{bq} (to which county-year, ℓ-j, is linked) all other model-estimated county-year function values (besides that for ℓ-j) for which y_{bq} also provides information [including years (j^*) other than j for county ℓ]. The quantity $\hat{\boldsymbol{\beta}}_{\ell^* j^*}$ in Equation (17) represents the posterior mean of the sampled values from our MCMC,

$$\tilde{y}_{bq,\ell j} = y_{bq} - \sum_{\ell^* \neq \ell \in b} \sum_{j \in q} \mathbf{x}'_{\ell^* j} \hat{\boldsymbol{\beta}}_{\ell^* j} - \sum_{j^* \neq j \in q} \mathbf{x}'_{\ell j^*} \hat{\boldsymbol{\beta}}_{\ell j^*}. \tag{17}$$

We will plot the pseudo-statistics to demonstrate that the posterior for each matrix of $P \times T$ coefficients, $\mathbf{B}_\ell$, weights the contribution of each estimate, y_{bq}, in proportion to its precision (inverse variance), such that (ACS) estimates associated to group-periods closer in geography (that nests relatively fewer counties) and time exert more influence on the model-estimated result. Our presentation of results, to follow, will illustrate the fit mechanism by plotting each pseudo-statistic, $\tilde{y}_{bq,\ell j}$ for county-year, ℓ-j, with the size of the displayed point in proportion to its precision.

The figures that follow display fitted estimates from the conditional model of Section 3. Figure 2 displays the fitted function (in the pink line), along with the collections of pseudo-statistics in each year for a randomly selected county with 1-year period ACS observations. The size of each pseudo-statistic is in proportion to its precision, with 1-year period estimates colored in red, 3-year period estimates in green and 5-year period estimates in blue. Fixing a year, one may observe how densely this county is nested in areas of similar size by examining whether there are many pseudo-statistics of similarly large precision values; for example, examining the pseudo-statistics for $j = 2012$, we note the largest pseudo-statistic is the 1-year period estimate for DuPage County. The 1-year period statistic for DuPage County is embedded in a set of somewhat lower precision statistics, indicating that DuPage County nests in a group of larger areas (e.g., such as metropolitan areas), which we know because the pseudo-statistics for these areas are colored in red so that they are 1-year period estimates and because their precisions are smaller, indicating that they are larger areas that nest DuPage County. The green pseudo-statistic above this group of red, 1-year period pseudo-statistics indicates there is a relatively high precision 3-year period estimate available for DuPage County that provides information about $f_{\ell j}$ for $\ell =$ DuPage County and $j = 2012$. Sim-

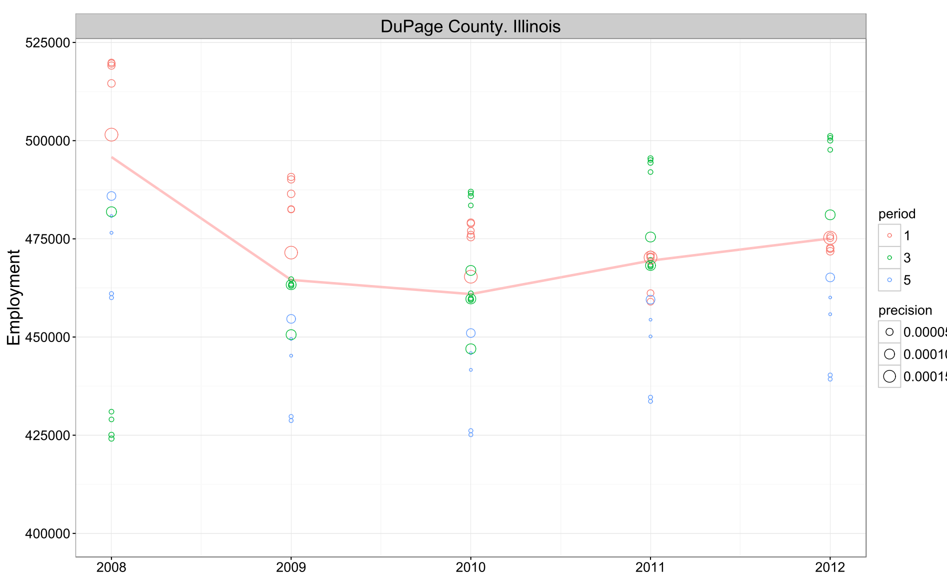

FIG. 2. *Estimated function vs. pseudo-data for* 1*-year county*: *Fitted function* (*pink line*) *compared to the collection of pseudo-data points in each year*, 2008–2012, *for a large-sized* (*by population*) *county*, *DuPage County*, *IL*, *with published* 1*-year period estimates*. *Each hollow circle represents a pseudo-statistic*, *and its size is proportional to its estimated precision*. *Each hollow circle is colored based on the period of the data point*; *red denotes a* 1*-year period*, *green denotes a* 3*-year period and blue denotes a* 5*-year period*.

ilarly, the blue-colored pseudo-statistic below the group of 1-year period pseudo-statistics is a 5-year period pseudo-statistic for DuPage County of relatively large precision that influences the modeled estimate. All to say, the collection of the relatively high-precision pseudo-statistics together influence the estimation of $f_{\ell j}$, for $\ell =$ DuPage County and $j = 2012$, with their estimation influence in direct proportion to their precisions. The remaining, relatively farther away pseudo-statistics of small precision provide relatively little information in the estimation of DuPage County because their population sizes are much bigger than DuPage County and likely nest many other counties. Since DuPage County has observed 1-year period estimates, those will be the most precise (and hence largest in size) for estimating this county. Nevertheless, we see that while the fitted trend is similar to that expressed by the 1-year estimates, it differs because the fitted values are influenced by pseudo-statistics representing other groups and periods in which the county-years nest. These values lie below the 1-year values and pull down the fitted function away from the 1-year period estimate.

We may *not* use these pseudo-data plots to assess the fit quality, however, precisely because the pseudo-statistics are convolved with the estimation procedure. We may, nevertheless, comment on the coherence or closeness among estimated pseudo-statistics with relatively larger precision values, which offers comment on the strength of estimation.

Figure 3 displays the estimated function compared to pseudo-statistics for a county with 3- and 5-year period observations, but not 1-year period observations. We see a good coherence between estimated by year 1- and 3-year pseudo-statistics constructed from near in size groups in which this county nests and also among the pseudo-statistics constructed from 3- and 5-year period estimates that nest each year in the figure.

Figure 4 presents an MCD for which only a single 5-year period estimate is available. These results also express a good coherence between the pseudo-statistic in each year constructed from the 1-year period estimate of the county in which the MCD nests and the pseudo-statistic in each year constructed from the 5-year period estimate for the focus county.

We observe in these figures that some of the pseudo-statistics are very large in magnitude—highly positive or negative—though their small precisions result in their exerting little to no influence in the estimation of the (parameters for the) functions, $(f_{\ell j})$. The overly high magnitude values occur in those cases where a county is nested in an area far different in size than itself, for example, nested in a balance of metropolitan areas, which will potentially include hundreds of counties. While a state-level estimate may be relatively precise for estimating a large, state-level quantity, it is highly imprecise for estimating a small, constituent piece. Thus, there is almost no information borrowed from a group that is far larger in size than a constituent county, reflecting a limitation in the ability of the model to borrow information.

In general, we find that the super sector employment total predictors from the Quarterly Census of Employment and Wages (QCEW), available at the county-

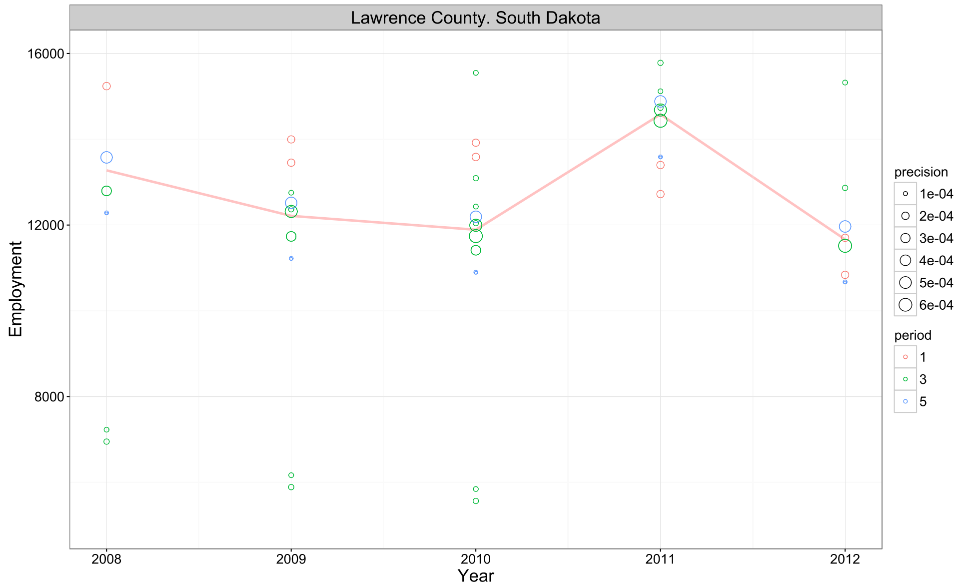

FIG. 3. *Estimated function vs. pseudo-data for* 3*-year county*: *Fitted function* (*pink line*) *compared to the collection of pseudo-data points in each year,* 2008–2012, *for a medium-sized* (*by population*) *county*, *Lawrence County*, *SD*, *with published* 3*-year* (*but not* 1*-year*) *period estimates*. *Each hollow circle represents a pseudo-statistic and its size is proportional to its estimated precision*. *Each hollow circle is colored based on the period of the data point*.

year resolution, helps to identify the county-year functions by providing magnitude information that stabilizes the estimation of by-county regression coefficients since the county employment totals span vast differences in the size of their labor markets. Yet, the resulting modeled estimate is typically quite different in level and trend (not shown) than the total of the QCEW super sector employment values. We are not surprised because the QCEW provides place-of-work employment obtained from business establishments, while the ACS is a household survey providing place-of-residence employment.

Our estimation model entirely focuses on estimating fine-level, county-year parameters, using groups and periods that nest them. A question arises about the quality of estimation at the *state* level composed by summing over the county-year parameters nested in each state-year. The roll-up of estimated functions to the states produces estimates for all states that are within 1–2% of 1-year period state-level estimates in the ACS. Figure 5 shows the estimated summed functions compared to the observed data points for three randomly selected states, which illustrates the estimation of latent functions at the county-year level provides a good estimation for state-level, 1-year period observations.

We may not directly assess the fit performance of the estimated county-year functions for 3- and 5-year counties due to the absence of observed 1-year data

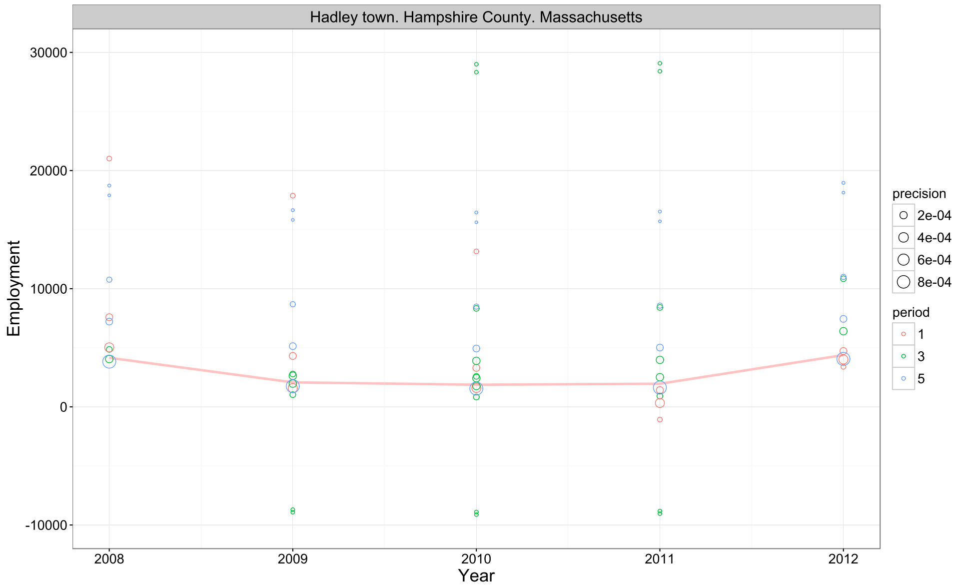

FIG. 4. *Estimated function vs. pseudo-data for* 5*-year county*: *Fitted function* (*pink line*) *compared to the collection of pseudo-statistics in each year*, 2008–2012, *for a small-sized* (*by population*) *township* (*MCD*), *Hadley*, *Hampshire County*, *MA*, *with published* 5*-year* (*but not* 1*- or* 3*-year*) *period estimates*. *Each hollow circle represents a pseudo-statistic*, *and its size is proportional to its estimated precision*. *Each hollow circle is colored based on the period of the data point*.

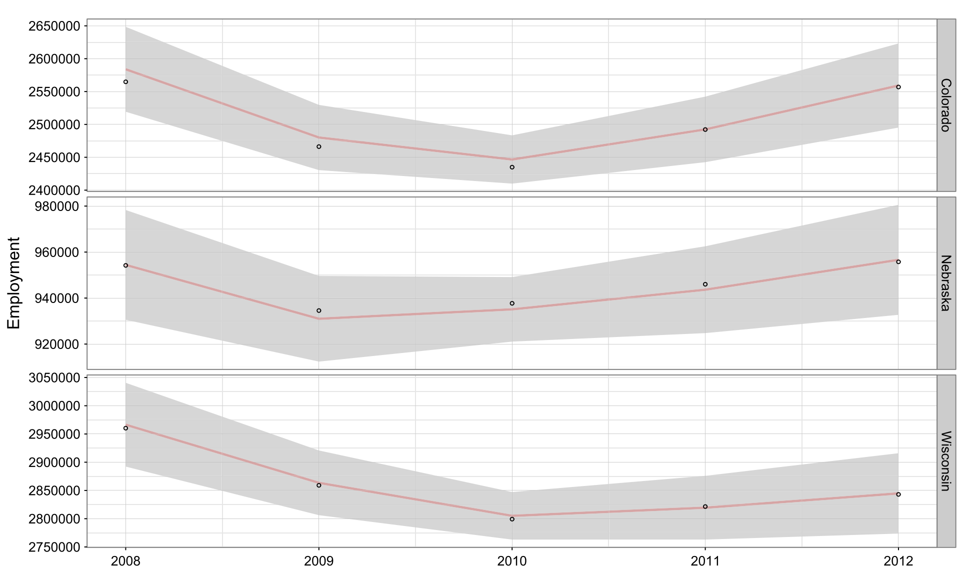

FIG. 5. *County-year fitted values summed to state level* (*pink line*) *versus data values* (*hollow circles*) *for randomly selected states*. *The gray shading represents the* 95% *credible intervals*.

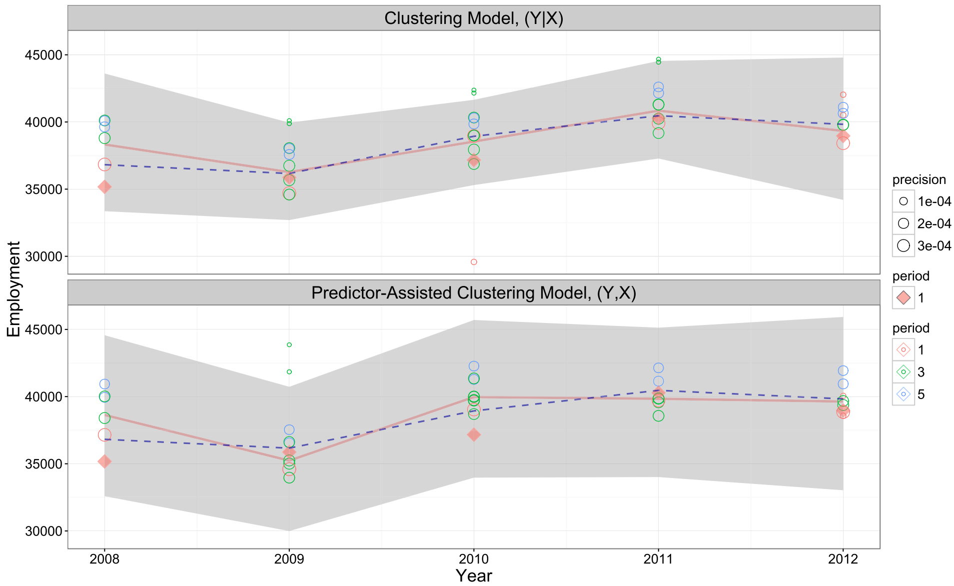

FIG. 6. *Comparison of model-estimated values for a* 1*-year county* (*Craven County*, *NC*) *when excluding* 1*-year data values*. *The bottom plot panel provides results for the predictor-assisted clustering model* [*which we label* (Y, X)], *while the plot in the top panel excludes predictors in the prior for cluster assignments* [*which we label* $(Y|X)$]. *The solid*, *pink line in each plot panel presents the posterior mean fitted function when excluding the* 1*-year data points*, *while the dashed*, *blue line presents the posterior mean when including the* 1*-year data points*. *The gray shading represents the* 95% *credible intervals as estimated on the models excluding* 1*-year data points*. *The associated pseudo-statistics are also estimated from the models excluding* 1*-year data points*. *The solid pink diamonds plot the* 1*-year data points*.

values. An indication of fit quality may, however, be provided by *holding out or excluding* the (five) 1-year data values for a county *with* available 1-year data values and comparing how the models—that exclude or include predictors in the prior distributions for cluster assignments—estimate the county-year function to when the 1-year values are included. Figure 6 presents estimated county-year function parameters for Craven County, North Carolina. The bottom panel displays estimated results under the predictor-assisted, dependent clustering model of Section 4, while the top panel displays the same under the clustering model that excludes predictors (in the prior for assignment to clusters) of Section 3. The solid, pink line in each plot panel presents the posterior mean fitted function when excluding the 1-year data values, while the dashed, blue line presents the same when including the 1-year values. The gray shading displays the associated 95% credible intervals under exclusion of the 1-year data values, and the associated pseudo-statistics are also constructed using the fitted functions under exclusion of these values. Finally, the pink, diamond points display the 1-year data values.

We explored a number of 1-year counties at random and found a high degree of similarity between the estimated county-year functions with and without inclusion of the 1-year data values under both models. Craven County is something of a worst-case result (due to its relatively small size) that provides clearer differentiation between the performances of the two models. We see that both models amplify the estimated employment decline from 2008–2009 when the 1-year data values are *excluded*, which increases the influence of the other groups containing Craven County. Yet, the model excluding predictors in assigning clusters well captures both the increasing trend from 2010–2011 and the decreasing trend from 2011–2012. The predictor-assisted clustering model expresses a slightly steeper decline, followed by a more rapid recovery. The predictors appear to have induced co-clustering among counties with this pattern during the Great Recession, causing an overemphasis on this period. The fitted results under both models may be sensitive to the choice of the county-year predictors because they are *below* the resolution of the observed data; for example, perhaps if we include additional predictors that provide information about poverty concentration or education achievement, then the predictor-assisted model may or may not outperform, and so predictors should be carefully chosen based on their ability to comment on the economic conditions of each county. The larger credible intervals for the predictor-assisted clustering model reflects the large space of partitions or clusterings induced when including the predictors in the prior for the mixing measure. These results generally suggest that the spatial and temporal nesting construction that underpin our models may provide reasonable estimates because we have shown that the models well reproduce the denoised, by-year functions that would be estimated under inclusion of the 1-year ACS estimates when the 1-year ACS estimates are *excluded* from the models. Under exclusion of the 1-year estimates, the models must rely on both coarser period estimates for Craven County, as well as the information provided by other areas in which Craven County nests.

Our primary modeling goal is not merely to fit the observed data, (y_{bq}), but to discover true employment trends at resolutions lower than the observed data. We, nevertheless, conduct posterior predictive checks for both models fitted to the observations, (y_{bq}), by computing the Bayesian posterior predictive p-value, $p_B = P(T(y^{\text{rep}}, (f_{\ell j})) \geq T(y, (f_{\ell j}))|\mathbf{Y})$, using $T(y, (f_{\ell j})) = \sum_{b=1}^{B} \sum_{q=1}^{Q} (y_{bq} - \sum_{\ell \in b} \sum_{j \in q} f_{\ell j})^2/\sigma_{bq}^2$, the Pearson chi-squared statistic [Gelman et al. (2015)]. The p-value is rendered by generating replicate data, y^{rep}, from our models on each posterior sampling iterations, computing the statistic, and counting the number of times it is greater under the replicated than the real data. The intent is to provide a baseline check on whether the estimated model generates data of a similar structure to the real data, where values closer to 0.5 may indicate a good fit of the model to the data. We are not surprised by posterior predictive p-values of $(0.48, 0.47)$ for the conditional and joint models, respectively. The harder challenge is to assess the goodness of fit for estimation of county-level functions for which we don't have estimates. Section 6, which immediately follows, constructs

TABLE 2

Fit performance comparison between model including predictors in prior for cluster assignments, (Y, X), and model excluding predictors in clustering, $(Y|X)$. Lower values indicate better fit performance for all included fit statistics

	$(\boldsymbol{Y}, \boldsymbol{X})$	$(\boldsymbol{Y}\vert\boldsymbol{X})$
−LPML	233,517	228,181
DIC_3	449,663	450,199
$\bar{D}$	444,634	446,928

synthetic data under a known truth to allow us to conduct an assessment of fit for our model-estimated county-level functions.

We conclude this section by comparing the relative fit performances between our two model alternatives in Table 2, where we see that the predictor-assisted clustering model doesn't produce a notably lower mean deviance, $\bar{D}$, than the simpler model to justify the added complexity. The table additionally displays the DIC_3 criterion [Celeux et al. (2006)] that focuses on the marginal (predictive) density $\widehat{f(\mathbf{y})}$ in lieu of $\widehat{f(\mathbf{y}|\text{parameters})}$, which is more appropriate for mixture models. Also shown is the log-pseudo-marginal likelihood that employs "leave-one-out" cross-validation [Gelfand and Dey (1994)]. The similar fit statistics, combined with the lower perturbation in the estimated functions illustrated in Figure 6, incline us to prefer the simpler model of Section 3.2.

6. Simulation study. Our examination of results for the ACS helped provide insight on the fit performance, but perhaps does not fully address the quality of fit for counties with only 3- and 5-year data values, and so we generate known synthetic values for coefficients, $(\mathbf{B}_\ell)$ from Equation (5), employing the posterior means of covariance parameters $(\hat{\mathbf{\Lambda}}_{y,\ell}, \hat{\boldsymbol{\kappa}}_\ell)$ from the model of Section 3. We compute $f_{\ell j} = \mathbf{x}'_{\ell j}\beta_{\ell j}$, where $\mathbf{X}_\ell$ is observed (known). We next generate $y_{bq} \stackrel{\text{ind}}{\sim} \mathcal{N}(\sum_{\ell \in b}\sum_{j \in q} f_{\ell j}, \sigma^2_{bq})$. The same nesting relationships of (county, year) to (group, period) from the ACS are duplicated for the simulation study so that we are generating a synthetic version of ACS employment counts. Of course, this simulation assumes that our spatial and temporal nesting construction is the correct generating model, which we do not know to be the case, though the fit performances on 1-year counties when excluding the 1-year data values suggests that this assumption may be broadly reasonable.

Figure 7 presents the pseudo-statistics, fitted function (denoted by a pink line) and associated 95% credible interval (denoted by gray shading) along with the true function (denoted by the dashed, blue line) for a 3-year county. It reveals that our model also does well on a county for which we have 3-year period estimates, but not 1-year period estimates.

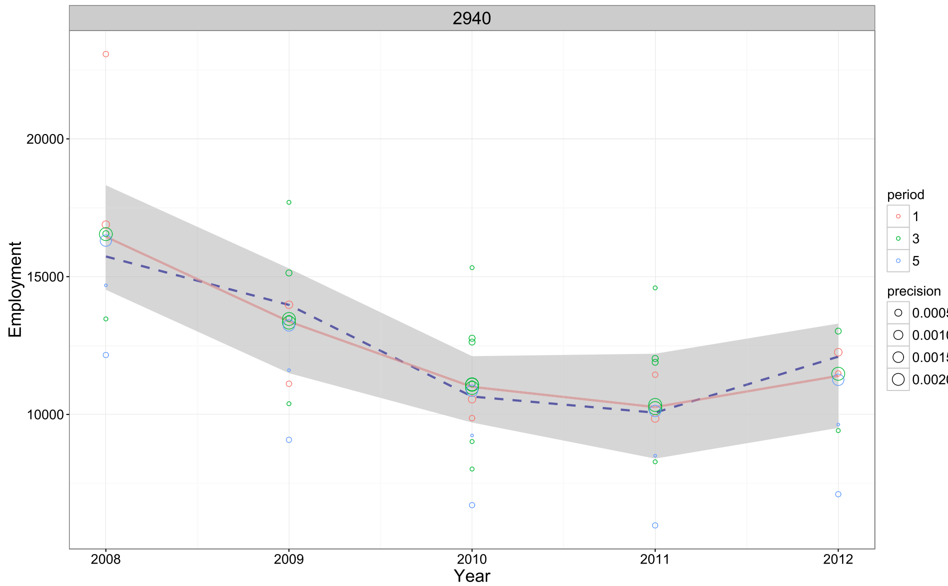

FIG. 7. *Fitted versus data values for simulated* 3*-year county.*

We find similar results for counties where each has only a single 5-year period ACS estimate [see the Technical Supplement in Savitsky (2016) for a plot] as for those with only 3-year period estimates in that the true trend is captured, though there is some over-smoothing for 5-year counties, particularly those that are nested far in space into much larger areas (along with many other 5-year counties). Adding data for upcoming years will bring in additional 5-year period estimates, which are expected to improve the quality of estimation for these far-nested counties by borrowing strength over (overlapping) multi-year periods.

7. Discussion. Motivated by the use of ACS employment data at the BLS to allocate statewide CPS employment estimates to sub-state, local areas, we have developed a general approach to estimate fine-scale time and areal-indexed parameters using an ensemble of coarse-scale observations that spatially and temporally nest the parameters. We specify the likelihood to link subsets of the parameters that exhaustively nest each group-period observation. Our best-performing Bayesian multiscale model of Section 3 formulates a relatively simple nonparametric mixture model for estimating the latent county functions in a fashion that facilitates the shrinking together of similar functions by the data. The flexible shrinking under the Bayesian nonparametric approach, which penalizes complexity, combined with leveraging nesting relationships to identify an ensemble of observations that provide information about each latent parameter, provides a broadly useful approach.

Many ACS users, such as the LAUS program in BLS, would prefer to employ 1-year period estimates for counties, but are relegated to using 5-year period published estimates in the case where analyses are conducted across all counties in the U.S. Results from our simulation study demonstrate that our approach performs well to uncover the latent true county-year parameters for 3-year counties and 5-year counties, where the 5-year counties nest within similarly sized groups (along with few other counties). There was some notable over-smoothing of the estimated county function (though the magnitude and global trend are captured) for 5-year counties exclusively nested in much larger sized groups, which occurs because we only have a single, 5-year period estimate for these counties. We expect improvements in the fit accuracy for these counties as we add upcoming years to the five years of data that we considered for our analysis because our mixtures of Gaussian process formulations borrow strength across years. Employing an ensemble of estimates published at varied resolutions even adds value for the modeling of counties with 1-year period estimates by incorporating the additional estimates associated to groups nesting each 1-year county. Our approach may be applied to any variable from the ACS, as well as to other data sets that express a multiresolution structure where a collection of latent parameters are linked to multiple observations that are coarser in scale. It is difficult, however, to think of an example of such a data structure because the usual case is to make inference about coarser level parameters from data observed at finer resolutions [e.g., inferring about the performance of a teacher from the individual performances of students in their classrooms [Savitsky and McCaffrey (2013)]. More generally, then, our multiresolution modeling formulation serves as an example of what is possible for performing inference on structured data using Bayesian hierarchical models.

Our estimates were composed using 3-year period estimates for which the Census will discontinue publication going forward (from 2016). Leveraging historical 3-year period estimates in our nonparametric multiresolution modeling remains, however, very useful because they continue to offer information for borrowing of estimation strength among the $(f_{\ell j})$ because we retain previous years of data when adding new years. Our modeling approach and computation are sufficiently flexible to input data at any level of coarseness in space and time.

Acknowledgments. The authors wish to thank our colleagues at the Bureau of Labor Statistics whose focus on continuous improvement led them to sponsor this project. We thank the following important contributors:

1. Sean B. Wilson, Senior Economist, who led the effort to gain approval for the project and evaluate the quality of results.
2. Garrett T. Schmitt, Senior Economist, who helped us structure the data and associated county-group links.

SUPPLEMENTARY MATERIAL

Technical Appendices (DOI: 10.1214/16-AOAS968SUPP; .pdf). The online supplement contains three technical appendices with detailed material on the following topics:

1. posterior computation;
2. posterior mixing;
3. simulation study 5-year county results.

REFERENCES

BRADLEY, J. R., WIKLE, C. K. and HOLAN, S. H. (2014). Bayesian spatial change of support for count-valued survey data. Available at http://adsabs.harvard.edu/abs/2014arXiv1405.7227B.

BRADLEY, J. R., WIKLE, C. K. and HOLAN, S. H. (2015). Spatio-temporal change of support with application to American Community Survey multi-year period estimates. *Stat* **4** 255–270. MR3414659

CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Rejoinder to "Deviance information criteria for missing data models." *Bayesian Anal*. **1** 701–706 (electronic). MR2282197

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc*. **90** 577–588. MR1340510

GELFAND, A. E. and DEY, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514. MR1278223

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2015). *Bayesian Data Analysis*, 3rd ed. Chapman & Hall/CRC, Boca Raton, FL.

GELMAN, A. and RUBIN, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci*. **7** 457–511. Available at http://www.stat.columbia.edu/~gelman/research/published/itsim.pdf.

GHOSH, M., NATARAJAN, K., STROUD, T. W. F. and CARLIN, B. P. (1998). Generalized linear models for small-area estimation. *J. Amer. Statist. Assoc*. **93** 273–282. MR1614644

HAWALA, S. and LAHIRI, P. (2012). Hierarchical Bayes estimation of poverty rates. Technical report, U.S. Census Bureau—Small Area Income and Poverty Estimates. Available at https://www.census.gov/did/www/saipe/publications/files/hawalalahirishpl2012.pdf.

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc*. **101** 1537–1547. MR2279478

MÜLLER, P., QUINTANA, F. and ROSNER, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist*. **20** 260–278. MR2816548

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. MR2514435

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. Chapman & Hall/CRC, Boca Raton, FL. MR2130347

SÄRNDAL, C., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York. MR1140409

SAVITSKY, T. D. (2016). Supplement to "Bayesian nonparametric multiresolution estimation for the American Community Survey." DOI:10.1214/16-AOAS968SUPP.

SAVITSKY, T. D. and MCCAFFREY, D. F. (2013). Bayesisan hierarchical multivariate formulation with factor analysis for nested ordinal data. *Psychometrika* **79** 275–302.

SAVITSKY, T. D. and PADDOCK, S. M. (2013). Bayesian nonparametric hierarchical modeling for multiple membership data in grouped attendance interventions. *Ann*. *Appl*. *Stat*. **7** 1074–1094. MR3113501

SAVITSKY, T., VANNUCCI, M. and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist*. *Sci*. **26** 130–149. MR2849913

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist*. *Sinica* **4** 639–650. MR1309433

OFFICE OF SURVEY METHODS RESEARCH
U.S. BUREAU OF LABOR STATISTICS
2 MASSACHUSETTS AVE. N.E
WASHINGTON, DC 20212
USA
E-MAIL: Savitsky.Terrance@bls.gov