Inflection points in community-level homeless rates*

Chris Glynn[†], Thomas H. Byrne,[‡] and Dennis P. Culhane[§]

Abstract

Statistical models of community-level homeless rates typically assume a linear relationship to covariates. This linear model assumption precludes the possibility of inflection points in homeless rates – thresholds in quantifiable metrics of a community that, once breached, are associated with large increases in homelessness. In this paper, we identify points of structural change in the relationship between homeless rates and community-level measures of housing affordability and extreme poverty. We utilize the Ewens-Pitman attraction distribution to develop a Bayesian nonparametric mixture model in which clusters of communities with similar covariates share common patterns of variation in homeless rates. A main finding of the study is that the expected homeless rate in a community increases sharply once median rental costs exceed 32% of median income, providing statistical evidence for the widely used definition of a housing cost burden at 30% of income. Our analysis also identifies clusters of communities that exhibit distinct geographic patterns and yield insight into the homelessness and housing affordability crisis unfolding on both coasts of the United States.

1 Introduction

Homeless rates in the United States vary significantly from one community to another. According to the U.S. Department of Housing and Urban Development (HUD), roughly 1 in 1,250 people were counted as homeless in Glendale, CA in January 2017, while 1 in 70 people were counted as homeless in Mendocino County, CA that same month (HUD, 2017). This more than seventeen-fold increase in the rate of homelessness within the state of California suggests that homelessness is critically influenced by features of individual communities. In this study, we investigate complex and potentially nonlinear relationships between homeless rates and community-level predictors.

Quantifying the association between homeless rates and covariates of a community is practically useful along two dimensions. First, it sharpens public focus on the social forces related to homelessness – leading to improved monitoring and intervention opportunities to help the most vulnerable citizens. Second, it provides a set of measurable objectives to guide public policy.

A significant number of studies have investigated statistical associations between covariates of a community and homelessness[1] (Corinth, 2015; Byrne et al., 2013; Lee et al., 2003; Quigley

[*] This work was supported by Zillow. The authors thank Kaoru Irie, Surya Tokdar, Cory Hopkins, Alexander Casey, and Svenja Gudell for helpful discussions.

[†] Peter T. Paul College of Business and Economics, University of New Hampshire, christopher.glynn@unh.edu

[‡] School of Social Work, Boston University, tbyrne@bu.edu

[§] School of Social Policy & Practice, University of Pennsylvania, culhane@upenn.edu

[1] In this paper, we examine inter-community variation in homeless rates based on point-in-time counts across HUD-defined continuums of care. An alternative approach to assessing the relationship between community factors and homeless rates is to look at neighborhoods within a city as "communities" and measure rates of shelter admission from those communities based on last address. See, for example, Culhane et al. (1996) and Rukmana (2008)

et al., 2001); however, existing statistical models of homeless rates alternate between two extreme assumptions. At one extreme, analyses assume a single global parameter so that the relationship between homelessness and housing costs, for example, is the same nationwide (see, e.g., Byrne et al. (2013)). Assuming a single global parameter is rigid, and it precludes the possibility that local social structures mitigate (or exacerbate) the role that housing costs play in housing vulnerability. At the other extreme, Glynn and Fox (2019) endow each community with a local parameter in a hierarchical statistical model. Assuming local effects for each community is problematically flexible, as there is scarce data on the size of the homeless population in each community – leading to imprecise estimates of model parameters. In the presence of scarce data, there is a trade-off between model flexibility and the precision of parameter estimates.

Between these extremes of model rigidity and flexibility exists a middle ground where clusters of similar communities share model parameters. This modeling strategy has both statistical and applied advantages. From a statistical perspective, pooling information across similar communities provides sharper estimation of the association between community-level covariates and homelessness. From an applied perspective, identifying clusters of communities is a way to define highly-relevant peer groups for development and evaluation of policy interventions.

We have two primary objectives in this paper:

(O_1) Flexibly estimate the relationship between community covariates and homeless rates to identify points where structural changes in the relationship occur; and

(O_2) Identify clusters of communities where homeless rates exhibit common patterns of variation.

The statistical challenge is to estimate the complex functional relationship between homeless rates and community-level covariates from scarce data. Because there is limited variation in the features of a community from one year to the next, data from a single community is concentrated in a limited region of predictor space. Estimating the complete response surface in predictor space requires pooling data across related communities and fusing together local estimates. To estimate the response surface locally, we pool data from communities with similar covariates utilizing a Bayesian nonparametric mixture model where prior probabilities of cluster assignments depend on covariates. A consequence of this covariate-dependent clustering strategy is that communities in our analysis are no longer treated as exchangeable, and standard nonparametric mixture models, such as the Dirichlet process mixture model (Antoniak, 1974; Escobar and West, 1995), are not suitable for our analysis. To include community-level covariates in the prior probability of a partition over communities, we utilize the Ewens-Pitman attraction (EPA) distribution of Dahl et al. (2017). The EPA distribution is one in a broader class of nonexchangeable prior distributions for random partition models (Müller et al., 2011; Park and Dunson, 2010; Shahbaba and Neal, 2009), which have been successfully used in applied analyses when the number of covariates is small (Page and Quintana, 2016, 2015; Dahl, 2008).

The EPA distribution is a prior distribution over the space of partitions indexed by pairwise similarity between observational units (communities in our case). The applied intuition is that communities with similar covariates have a higher prior probability of membership in the same cluster than communities with covariates that are dissimilar (Page and Quintana, 2018). We utilize the EPA distribution rather than dependent Dirichlet processes (MacEachern, 2000) or distance-dependent Chinese restaurant processes (Blei and Frazier, 2011) so that we directly model the partition of communities with covariate information. Three important aspects of our model are (i) the number of clusters; (ii) cluster membership; and (iii) the relationship between community covariates and homelessness within clusters are all jointly estimated as part of the inference procedure. We compute fully Bayesian posterior distributions with a custom Markov chain Monte Carlo algorithm that seamlessly combines the Polya-Gamma data augmentation strategy of Polson et al. (2013) with the Gibbs sampling algorithm of Dahl et al. (2017) and a forward filtering backward sampling (FFBS) algorithm to account for community-specific temporal trends.

Our analysis focuses on three measures of a community: rental costs, measured by Zillow's Rent Index (ZRI), median household income, and the percent of residents living in extreme poverty. While the cost of housing is consistently identified as a predictor of homelessness both across (Byrne et al., 2013) and within (Glynn and Fox, 2019) communities, housing costs in absolute dollar amounts are an incomplete measure of housing affordability. The combination of housing costs and household income – specifically, the percent of income spent on housing costs – more completely reflects the relative affordability of housing across communities. By focusing on median housing costs as a share of median income, we more directly compare housing affordability in communities with different housing markets and economies. While median housing affordability measures account for varying housing markets and income levels, they do not reflect the size of the population in a community whose income is inadequate to meet the cost of housing. To control for the size of the population in each community that is most vulnerable to homelessness, we also include in our model the percent of a community living in extreme poverty.

Our analysis identifies a structural change in homeless rates when housing costs in a community reach 32% of median income. After housing costs exceed 32% of median income, the expected homeless rate in a community increases sharply. We also find three dominant modes of variation in homeless rates, with 377 of 386 total communities in our analysis falling into one of three clusters: communities in the first cluster – primarily located in the midwest, mid-Atlantic, and southeast – tend to have very low homeless rates and modest housing costs; communities in the second cluster – including most of New England, Florida, the mountain west and central United States – have intermediate homeless rates and housing costs on par with the national average; communities in cluster three, which span much of the west coast and include large metropolitan areas on the east coast, have very high homeless rates and high costs of housing.

The paper proceeds as follows: in Section 2, we describe the data used in our analysis; in Section 3, we present our EPA-based mixture model of homeless populations and describe choices for prior distributions; in Section 4, we detail our Markov chain Monte Carlo inference procedure; in Section 5, we present posterior predictive distributions for homeless rates over a range of housing affordability and extreme poverty levels, and we also identify clusters of communities sharing similar associations; in Section 6, we conclude with a discussion of our findings and how the clusters of communities can be effectively utilized for policy prescriptions.

2 Data

The data used in our analysis spans the years 2011 to 2017 and comes from three sources: HUD, the American Community Survey (ACS), and the real estate analytics firm Zillow.

Each year, HUD produces a nationwide estimate of the number of people experiencing homelessness on a single night. The national estimate is based on local enumeration efforts called point-in-time (PIT) counts. While the PIT counts are conducted in January, the data is typically released the following November. At the local level, counts are conducted in roughly 400[2] continuums of care (CoCs), geographic units that coordinate support services for homeless and whose boundaries are typically coterminous with a single city, a single county, or a group of counties. In 2017, PIT estimates were produced for 399 CoCs across all 50 states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, and Guam.

To estimate homeless rates, it is essential to know the relative size of CoCs; however, the total population of a CoC is not reported by HUD. Discrepancies between geographic boundaries of CoCs and boundaries of geographic units for which total population estimates are made available by the U.S. Census Bureau mean that total population estimates for some CoCs are not readily available. To overcome this mismatch, we develop a crosswalk between HUD CoCs – the most granular geographic unit for which homeless data is available nationally – and census tracts. To match census tracts with CoCs, we utilize a process conceptually similar to that described by Byrne et al. (2013). Specifically, we use geospatial data from HUD on the boundaries of each CoC and compute the geographic centroid of each census tract. If the tract centroid falls within the boundaries of a CoC, we match the whole tract to the CoC. Based on this assignment of tracts to CoCs and tract-level ACS 5-year population estimates, we construct approximate total population measures for each CoC. For example, to construct the CoC total populations in 2011, we use the 2007-2011 ACS 5-year estimates. These CoC total population estimates and PIT counts facilitate comparisons of homeless rates across communities of various sizes. We have made the code used to conduct the geospatial matching and construct the CoC total population estimates publicly available on the GitHub page of one of the authors (Byrne, 2018).

We focus our analysis on three particular covariates of a community: rental costs, measured by Zillow's rent index (ZRI), median household income, and the percent of residents living in extreme poverty. Median household income data and the percent of residents living in extreme poverty are also reported in ACS. We weight tract-level measures of median income and extreme poverty by the tract-level populations and aggregate to construct CoC-level measures of median household income and rates of extreme poverty. To measure rental costs, we follow Glynn and Fox (2019) and utilize a custom-computed ZRI. The critical difference in the rental data for this analysis and that used by Glynn and Fox (2019) is that in the present study, Zillow directly computed a rent index for each CoC based on geospatial data provided by HUD. The rent index methodology is identical to Zillow's existing ZRI methodology (Bun, 2012), but it is brought to the non-standard CoC geographies – providing a measure of rent not previously available to researchers utilizing PIT count data. Table 1 presents a snapshot of the data for the New York City CoC (NY-600). While countless measures of a community are potentially associated with homelessness – including apartment vacancy rates, unemployment rates, demographics, etc. – most are highly correlated with the covariates that we include in our analysis.

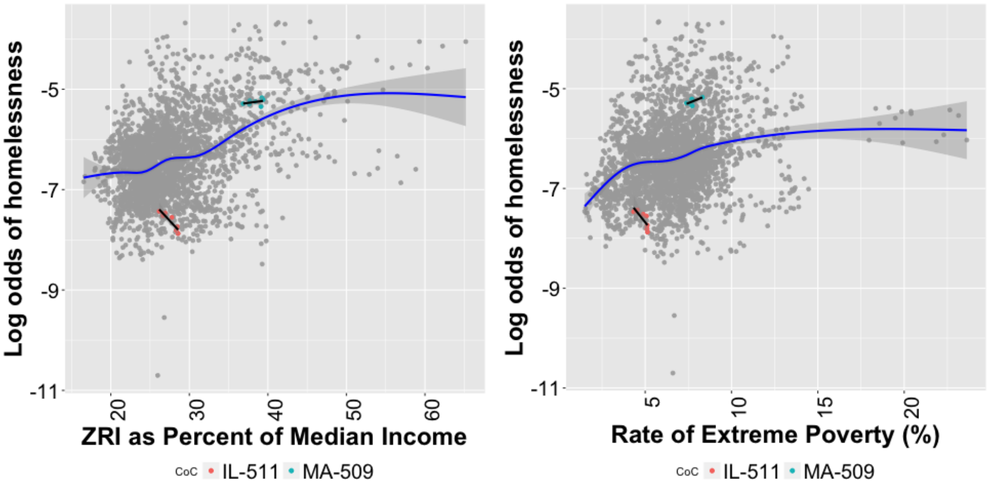
Observe in Figure 1 that as both ZRI (as a percentage of median income) and the rate of extreme poverty increase, the estimated log odds of homelessness generally increases as well; however, this is not universally true. In Figure 1a, observe that the data strands for the Cook County (IL) CoC and the Cambridge (MA) CoC exhibit very different associations with ZRI / Median Income. A single linear model is too rigid to realistically explain the disparate associations; however, the CoC-level data sequences are only 7 years long, and inference on local model parameters characterizing the individual relationships visualized in Figure 1a may not be robust. To over-

[2] The exact number of CoCs varies from year to year due to the creation or dissolution of CoCs or the merger of two or more existing CoCs. In 2007, there were 461 CoCs; in 2017 there were 399.

	Count	Population	ZRI ($)	Income ($)	Poverty (%)
2011	51,123	7,944,958	1,738.62	54,974.00	8.60
2012	56,672	8,009,322	1,768.21	55,510.05	8.82
2013	64,060	8,074,863	1,843.62	56,036.71	9.03
2014	67,810	8,159,782	2,010.27	57,029.83	9.08
2015	75,323	8,231,358	2,175.81	57,758.77	8.95
2016	73,523	8,268,601	2,322.79	59,552.74	8.79
2017	76,501	8,305,844	2,469.76	61,346.72	8.63

Table 1: Homeless count and community covariates of New York City CoC (NY-600), including all five burroughs of New York City.

come this data scarcity at the CoC-level and facilitate robust inference, we pool observations in a cluster of CoCs sharing a similar relationship. The GAM-smoothings of the log odds ratios in Figures 1a and 1b illustrate nonlinear increases in homeless rates associated with increases in ZRI/median income and rates of extreme poverty.



(a) Affordability

(b) Poverty

Figure 1: Imputed log odds of homelessness plotted against ZRI as a percentage of income (left) and rates of extreme poverty (right). The highlighted data are from the Cambridge (MA) CoC and the Cook County (IL) CoC, and the line segments through the MA-509 and IL-511 highlighted data correspond to ordinary least squares model fits. The solid lines spanning the full range of the x-axes in both figures present Generalized Additive Model (GAM)-smoothings of the CoC-level log odds.

3 A Bayesian nonparametric model for homeless counts

The novel modeling contribution of the study is a mixture model for latent homeless rates based on the Ewens-Pitman attraction (EPA) distribution (Dahl et al., 2017). The EPA distribution is a prior over the space of CoC-partitions, and it is indexed by pairwise similarity of CoCs themselves. Unlike the partition distribution implied by the Dirichlet Process prior (Ferguson, 1973) where data is modeled as exchangeable, the EPA distribution models dependence in cluster assignments based on covariates of CoCs. It assigns higher probability to partitions where CoCs with similar levels of housing affordability and extreme poverty belong to the same cluster. Before introducing the modeling innovation in Section 3.2, we first discuss our strategy for modeling the unobserved homeless rate in a community given the HUD-reported PIT counts and our noisy estimates of CoC-level total populations.

3.1 Modeling homeless rates as latent variables

Modeling homeless rates requires some care, as several data quality challenges prevent simply dividing PIT counts in a given year by the total CoC population. Hopper et al. (2008) provide evidence that street counts do not fully reflect the size of the homeless population in a community. This systematic undercount of homeless populations artificially lowers homeless rates and necessitates modeling the mechanism by which individuals are excluded from PIT counts. Uncertainty in the size of the homeless population is one aspect of the data quality challenge. Uncertainty in the total population of each CoC is a second aspect. While we observe the ACS 5-year estimates of total population at the tract level, tract populations are aggregated to form a noisy estimate at the CoC level. At both the tract and CoC level, the total population is not exactly known. Modeling noise in the numerator and denominator of a rate calculation allows for a more complete accounting of uncertainty in homeless rates.

To address these data quality challenges, we adopt the modeling framework proposed by Glynn and Fox (2019) and treat unobserved homeless rates as parameters in a hierarchical Bayesian statistical model. The hierarchical model has three levels: (i) a component model for the total population of CoC i in year t, denoted $N_{i,t}$; (ii) a component model for the unobserved total homeless population, denoted $H_{i,t}$; and (iii) a component model for the counted number of homeless, denoted $C_{i,t}$. In this hierarchical model, uncertainty in $N_{i,t}$ and $H_{i,t}$ propagate to estimates of the latent homeless rate, denoted $p_{i,t}$. We summarize critical components of the framework here.

Total Population. The total population of CoC i in year t is modeled with a Poisson random variable,

$$N_{i,t}|\lambda_{i,t} \sim Poisson(\lambda_{i,t}). \tag{1}$$

The expected total population in year t, $\lambda_{i,t}$, is further modeled over time in a way that admits a forward filtering backward sampling algorithm to infer $\lambda_{i,t}$ from the ACS 5-year estimates from 2011-2017. We refer the reader to Glynn and Fox (2019) for a discussion of prior distributions for $\lambda_{i,t}$, which are not the core focus of the current study.

Total homeless population. The total number of homeless $H_{i,t}$ is a small subpopulation of the CoC's total population. To model the size of the homeless subpopulation conditional on the total population of the CoC, a binomial thinning step is employed,

$$H_{i,t}|N_{i,t}, p_{i,t} \sim Binomial(N_{i,t}, p_{i,t}). \tag{2}$$

While $H_{i,t}$ is modeled as a latent variable given $N_{i,t}$, it is important to note that $H_{i,t}$ itself is not directly observed. We treat $H_{i,t}$ as missing data and impute it as part of our model fitting procedure. The homeless rate, $p_{i,t}$, is the focus of Section 3.2.

Homeless count. The counted number of homeless, a quantity less than or equal to $H_{i,t}$, is modeled as a conditionally binomial random variable

$$C_{i,t}|H_{i,t}, \pi_{i,t} \sim Binomial(H_{i,t}, \pi_{i,t}). \tag{3}$$

The parameter $\pi_{i,t} \in [0, 1]$ is the probability that a person who is homeless will be counted as homeless. We adopt priors for $\pi_{i,t}$ utilized by Glynn and Fox (2019) to carry out our analysis. As $H_{i,t}$ is not observed, it is not possible to learn $\pi_{i,t}$. We view $\pi_{i,t}$ as a nuisance parameter and integrate over it so that the marginal model $C_{i,t}|H_{i,t}$ is beta-binomial distributed.

3.2 A nonparametric mixture model for $p_{i,t}$

The primary modeling innovation of the study is a mixture model for $p_{i,t}$ based on the EPA distribution. As outlined in 2, homeless rate $p_{i,t}$ is the unobserved probability of homelessness in a Bayesian logistic regression. We transform $p_{i,t}$ to the real line with a logit transformation

$$\psi_{i,t} = log\left(\frac{p_{i,t}}{1 - p_{i,t}}\right) = F_i'\beta_{i,t} + X_{i,t}'\phi_i + \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim N(0, \sigma^2_{\psi_i}). \tag{4}$$

The log odds of homelessness in CoC i in year t, denoted $\psi_{i,t}$, is modeled as the composition of a dynamic latent factor $F_i'\beta_{i,t}$ and the regression $X_{i,t}'\phi_i$. We discuss each in turn.

Regression $X_{i,t}'\phi_i$. The $p \times 1$ vector $X_{i,t}$ is a set of community-level predictors and ϕ_i is a $p \times 1$ vector of regression coefficients. Our modeling objective is to induce a shared parameter vector across all CoCs in the same cluster. To achieve this objective, we reparameterize the collection $\phi_1, \ldots, \phi_n$ by the partition $\boldsymbol{\pi}_n = \{S_1, \ldots, S_{q_n}\}$ and shared cluster-level coefficients $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_{q_n})$. The partition $\boldsymbol{\pi}_n$ splits the CoC index set $\{1, \ldots, n\}$ into q_n mutually exclusive and non-empty subsets $S_1, \ldots, S_{q_n}$. When index $i \in S_k$, we say that CoC i belongs to cluster k and define cluster membership variable $Z_i = k$. The regression vector ϕ_i is then constructed from the set of unique $p \times 1$ vectors $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_{q_n})$ so that

$$\phi_i = \sum_{k=1}^{q_n} \tilde{\phi}_k \mathbb{1}_{\{Z_i = k\}}, \tag{5}$$

where each $\tilde{\phi}_k$ is independently drawn from a p-dimensional Normal distribution, $\tilde{\phi}_k \sim N(\mu_0, \Sigma_0)$. Hyperparameter choices for μ_0 and Σ_0 are discussed in Section 3.3.

In this study, we include a leading one in covariate vector

$$X_{i,t} = \begin{bmatrix} 1 & ZRI_{i,t}/MedianIncome_{i,t} & ExtPoverty_{i,t} \end{bmatrix}'.$$

The leading one results in a shared cluster-level intercept or expected rate of homelessness. One way of interpreting the cluster-level intercept is as the baseline homeless rate in a particular group of communities, an important metric for policymakers.

The model for inducing shared parameters in clusters of CoCs is completed by an EPA prior distribution over all possible partitions of CoCs. The EPA prior distribution for the partition of CoCs, $p(\boldsymbol{\pi}_n|\alpha, \delta, f, \boldsymbol{\omega})$, is indexed by a concentration parameter α (similar to the Dirichlet process), a discount parameter $\delta \in [0, 1)$, and similarity function f. The EPA distribution, which depends on the sequence in which CoCs are assigned to clusters and thus not exchangable, is also indexed by a permutation of CoC indices denoted $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$.

Cluster assignment probabilities depend on CoC covariates through similarity function f. The similarity function $f : \mathbb{R}^{3\times 3} \to (0, 1]$ maps distance between CoCs in covariate space to the unit interval, quantifying the pairwise similarity between two CoCs,

$$f(X_{\omega_j,T}, X_{\omega_i,T}) = \exp\{-\tau||X_{\omega_j,T} - X_{\omega_i,T}||_2\}. \tag{6}$$

CoCs ω_i and ω_j with identical covariates will have a similarity of one. If their covariates are far apart in $\mathbb{R}^3$, the similarity will be closer to zero. Decay in similarity is governed by temperature τ, a hyperparameter chosen by the modeler. In this analysis, we let $\tau = 0.35$ so that two CoCs ω_j and ω_i with $||X_{\omega_j,T} - X_{\omega_i,T}||_2 = 10$ are quite different, with similarity of $f(X_{\omega_j,T}, X_{\omega_i,T}) = 0.03$. For example, two CoCs that have the same level of extreme poverty but housing affordability measures that differ by 10% have very little similarity between them and a higher prior probability of being in different clusters. As τ increases, the probability that all members of a cluster are located near each other in predictor space increases as well.

The probability mass function $p(\boldsymbol{\pi}_n|\alpha, \delta, f, \boldsymbol{\omega})$ is constructed from the sequential product of conditional probabilities

$$p(\boldsymbol{\pi}_n|\alpha, \delta, f, \boldsymbol{\omega}) = \prod_{\ell=1}^{n} p_\ell(\alpha, \delta, f, \boldsymbol{\pi}(\omega_1, \ldots, \omega_{\ell-1}), \tag{7}$$

where $p_1(\alpha, \delta, f, \boldsymbol{\pi}_0) = 1$. For $\ell > 1$, $p_\ell(\alpha, \delta, f, \boldsymbol{\pi}(\omega_1, \ldots, \omega_{\ell-1}))$ is the probability that CoC ω_ℓ is assigned to cluster k given the previous assignments of CoCs $\omega_1, \ldots, \omega_{\ell-1}$, parameters α and δ, and similarity function f.

$$p_\ell(\alpha, \delta, f, \boldsymbol{\pi}(\omega_1, \ldots, \omega_{\ell-1})) = Pr(Z_{\omega_\ell} = k|\alpha, \delta, f, \boldsymbol{\pi}(\omega_1, \ldots, \omega_{\ell-1})) \tag{8}$$

$$= \begin{cases} \left(\frac{\ell-1-\delta q_{\ell-1}}{\alpha+\ell-1}\right) \frac{\sum\limits_{\{\omega_s:Z_{\omega_s}=k\}} f(X_{\omega_\ell,T},X_{\omega_s,T})}{\sum\limits_{s=1}^{\ell-1} f(X_{\omega_\ell,T},X_{\omega_s,T})}, \text{ for } k = 1, \ldots, q_{\ell-1} \\ \\ \frac{\alpha+\delta q_{\ell-1}}{\alpha+\ell-1} \text{ for } k = 0 \text{ (e.g., a new cluster)} \end{cases} \tag{9}$$

where $q_{\ell-1}$ is the number of clusters (subsets) in the partition of the first $\ell-1$ CoCs, $\boldsymbol{\pi}(\omega_1, \ldots, \omega_{\ell-1})$. Note that the probability of assignment depends on the order in which the CoCs are assigned. We address this non-exchangeability issue by utilizing a prior distribution for permutations and numerically integrating all over all possible permutations in our MCMC algorithm (see Section 4), resulting in a joint posterior distribution that is invariant to the ordering of the CoCs. Following Dahl et al. (2017), we use a uniform prior distribution so that $p(\boldsymbol{\omega}) = \frac{1}{n!}$ for all permutations.

The EPA distribution depends on the ratio of similarities

$$\frac{\sum\limits_{\{\omega_s:Z_{\omega_s}=k\}} f(X_{\omega_\ell,T}, X_{\omega_s,T})}{\sum\limits_{s=1}^{\ell-1} f(X_{\omega_\ell,T}, X_{\omega_s,T})}.$$

The numerator is the sum of similarity between CoC ω_ℓ and all other CoCs assigned to cluster k. The denominator is the total sum of similarity across all previously assigned $\ell - 1$ CoCs. Taken together, the ratio is the proportional attraction of CoC ω_ℓ to cluster k. By fixing $\delta = 0$, the cluster assignment process is a modified Chinese Restaurant Process. In fact, if the similarity function is constant (e.g., $f(X_{\omega_\ell,T}, X_{\omega_s,T}) = 1$) and $\delta = 0$, then the EPA distribution simplifies to the partition distribution implied by the Dirichlet Process. See Section 4.1 of Dahl et al. (2017). For this reason, we fix $\delta = 0$ and interpret the induced prior distribution for the collection $(\phi_1, \ldots, \phi_n)$ as a stochastic process prior that is similar to the Dirichlet process but – due to the EPA distribution over $\boldsymbol{\pi}_n$ – tilts a CoC's random cluster assignment towards a cluster where other members share similar covariates.

Innovation variance $\sigma^2_{\psi_i}$. The number of clusters q_n is significantly impacted by the choice of innovation variance $\sigma^2_{\psi_i}$ in 4. If the innovation variance is small, the variation of log odds around a particular regression line is tight, and many clusters are needed to explain variation in the $n = 386$ CoCs. As the innovation variance $\sigma^2_{\psi_i}$ increases, larger deviations in homeless rates from the regression fit are expected, and fewer clusters are needed. We model each $\sigma^2_{\psi_i}$ with an inverse gamma (IG) distribution, allowing the data to appropriately inform the innovation variance and number of clusters.

$$\sigma^2_{\psi_i} \sim IG(a_\psi, b_\psi) \tag{10}$$

A consequence of this model choice for $\sigma^2_{\psi_i}$ is that conditional on the latent factor $\beta_{i,t}$ and ϕ_i, the log odds of homelessness $p(\psi_{i,t}|\beta_{i,t}, \phi_i) = \int_0^\infty p(\psi_{i,t}|\beta_{i,t}, \phi_i, \sigma^2_{\psi_i})p(\sigma^2_{\psi_i})d\sigma^2_{\psi_i}$ is t-distributed. The heavy tails of $\psi_{i,t}|\beta_{i,t}, \phi_i$ allow for CoC-specific variation in homeless rates and a regression model that is robust to outlier homeless counts driven by idiosyncratic local events.

Dynamic latent factor $\beta_{i,t}$. The cluster-level regression coefficient ϕ_i models variation in $\psi_{i,t}$ associated with predictors $X_{i,t}$; however, there are many covariates of a community that are either excluded from $X_{i,t}$ or not directly observed. To account for these unobserved local covariates, we include a CoC-level dynamic latent factor $F_i'\beta_{i,t}$, allowing for small departures from the cluster-level regression that may be due to local policies, cultural attitudes toward homelessness, affordable housing initiatives, and many other difficult to observe local factors. The $F_i'\beta_{i,t}$ term reflects whether the environment in CoC i contributes to or reduces homelessness beyond the level associated with predictors $X_{i,t}$ in a specific cluster. To account for temporal trends in these latent factors at the CoC-level, we model $\beta_{i,t}$ with a two-dimensional state-space model

$$\beta_{i,t} = A\beta_{i,t-1} + w_{i,t}, \qquad w_{i,t} \sim N(0, W). \tag{11}$$

The dynamic latent factor model in 11 makes two important contributions: first, the 2×1 $\beta_{i,t}$ vector provides a mechanism to include (in aggregate) the unobserved community features that are excluded from $X_{i,t}$; second, it allows for temporal trends in homeless rates that are not well explained by predictors $X_{i,t}$. The locally linear trend model for $\beta_{i,t}$ is achieved by choosing $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $F_i' = \begin{bmatrix} 1 & 0 \end{bmatrix}$. See West and Harrison (1997) for more detail on dynamic model structures.

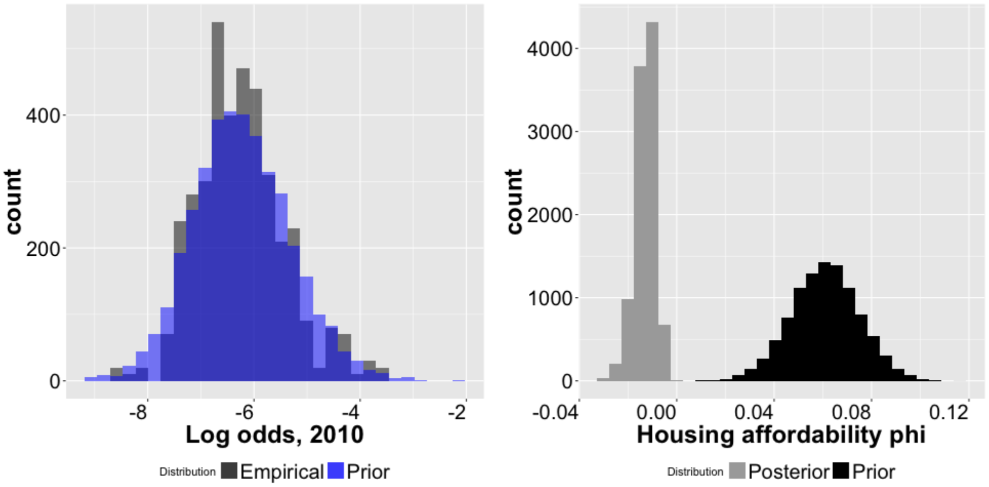
3.3 Prior choices

Prior distributions for $(\beta_{i,0}, \alpha, \sigma^2_{\psi_i})$ and hyperparameters μ_0 and Σ_0 are chosen by matching the first two moments of the implied prior distribution of $\psi_{i,0}$ to the empirical distribution for the log odds of homelessness computed from 2010 data. Since the data used in our analysis begins in 2011, we use data from 2010 to inform priors. The empirical distribution of log odds of homelessness in 2010 is unimodal and symmetric with sample mean -6.24 and sample variance 0.69 (see Figure 2a). The expectation of $\psi_{i,0}$, computed by taking the expectation of 4, is $E[\psi_{i,0}] = F'_{i,0}E[\beta_{i,0}] + X'_{i,0}E[\phi_i]$. We choose $E[\beta_{i,0}] = 0$ to encode our prior belief that the expected homeless rate in a community is the cluster-level contribution from CoC-predictors, $E[\psi_{i,0}] = X'_{i,0}E[\phi_i]$. The choice of $E[\phi_i]$ is akin to choosing mean μ_0. In Section 3.2, we noted that the cluster-level intercept may be interpreted as the baseline rate of homelessness in a community. Thus, we utilize PIT counts from 2010 on chronic homelessness to inform the first element $\mu_0^{(1)} = -8.28$. Remaining elements of μ_0 are chosen so that the difference between the sample mean in 2010 and $\mu_0^{(1)}$ is divided evenly across coefficients for housing affordability and extreme poverty, and $\mu_0^{(2)} = \mu_0^{(3)} = \frac{-6.24 - \mu_0^{(1)}}{\frac{1}{n}\sum_{i=1}^{n}\left(X_{i,0}^{(2)} + X_{i,0}^{(3)}\right)}$. When we include CoC data on housing affordability, $X^{(2)}$, and the rate of extreme poverty, $X^{(3)}$, we compute $\mu'_0 = \begin{bmatrix} -8.28 & 0.061 & 0.061 \end{bmatrix}$.

With the means of prior distributions chosen so that $E[\psi_{i,0}]$ matches the sample mean in the 2010 data, we follow a similar strategy in choosing prior variances. The objective is to compose $Var(\psi_{i,0})$ from contributions that are consistent with the modeler's uncertainty in each parameter. The variance $Var(\psi_{i,0})$ may be decomposed with an application of the law of total variance,

$$\begin{aligned} Var(\psi_{i,0}) &= E[Var(\psi_{i,0}|\beta_{i,0}, \phi_i, \sigma^2_{\psi_i})] + Var(E[\psi_{i,0}|\beta_{i,0}, \phi_i, \sigma^2_{\psi_i}]) && (12) \\ &= E[\sigma^2_{\psi_i}] + F'_{i,0}Var(\beta_{i,0})F_{i,0} + X'_{i,0}Var(\phi_i)X_{i,0}. && (13) \end{aligned}$$

We begin by fixing the latent factor covariance matrix $Var(\beta_{i,0}) = diag(0.1, 1 \times 10^{-6})$, which allows for meaningful systematic (as opposed to idiosyncratic) deviations in a community's homeless rate from the homeless rate of the cluster. The variance of ϕ_i, denoted by Σ_0, is chosen to encode the belief that our most uncertain component is the intercept, the baseline rate of homelessness. We fix $\Sigma_0 = diag(0.4, 0.0002, 0.0002)$. The choice of 0.0002 for the variance of coefficients associated with housing affordability and poverty encodes a strong prior belief that these parameters are positive, but it does not rule out a negative association, as illustrated in Figure 2b, where the posterior for the housing affordability coefficient concentrates on negative values in one of the clusters. The remaining variance component is $\sigma^2_{\psi_i} \sim IG(3, 0.1)$, which puts a diffuse prior on observational noise in homeless rates – encoding a belief that in some CoCs, the homeless rate is close to the regression fit, while in other CoCs, the rate fluctuates significantly due to random local factors. Dahl et al. (2017) note the relationship between α and the concentration parameter in the Dirichlet process, and we follow Escobar and West (1995) in utilizing the conventional $\alpha \sim Ga(1, 1)$ prior distribution. We note that prior choices for $Var(\beta_{i,0})$, α and $\sigma^2_{\psi_i}$ impact the inferred number of clusters. By choosing relatively diffuse priors for each, we give the data a significant role in informing the number of clusters. The marginal prior for $\psi_{i,0}$ is illustrated in Figure 2a. Observe that the induced prior for $\psi_{i,0}$ is slightly more diffuse than the empirical distribution of log odds in 2010, providing for the possibility that homeless rates in CoCs nationwide are actually more variable than was observed in 2010 alone.



(a) Moment Matching

(b) Posterior update

Figure 2: Left: The empirical distribution of log odds of homelessness in 2010 and the implied prior distribution for $\psi_{i,0}$. Right: the prior and posterior distributions for $\tilde{\phi}_k^{(2)}$, the parameter associated with housing affordability.

4 Markov Chain Monte Carlo

Our objective is to sample from the posterior distribution

$$p(\tilde{\boldsymbol{\phi}}, Z_{1:n}, \beta_{1:n,1:T} | N_{1:n,1:T}, C_{1:n,1:T}). \tag{14}$$

Our computational strategy is to condition on observations $N_{i,t}$ and $C_{i,t}$ while numerically integrating each $\sigma^2_{\psi_i}$, latent variables $H_{i,t}$ and $\psi_{i,t}$, and concentration parameter α from the joint posterior. Importantly, we also integrate over the permutation $\boldsymbol{\omega}$ so that the posterior distribution is invariant to the order in which we assign CoCs in the EPA partitioning.

$$\begin{aligned} & p(\tilde{\boldsymbol{\phi}}, Z_{1:n}, \beta_{1:n,1:T} | N_{1:n,1:T}, C_{1:n,1:T}) \\ & = \int p(\psi_{1:n,1:T}, H_{1:n,1:T}, \sigma^2_{\psi,1:n}, \alpha, \boldsymbol{\omega}, \ldots \\ & \ldots \tilde{\boldsymbol{\phi}}, Z_{1:n}, \beta_{1:n,1:T}, | N_{1:n,1:T}, C_{1:n,1:T}) dH_{1:n,1:T} d\psi_{1:n,1:T} d\sigma^2_{\psi,1:n} d\alpha d\boldsymbol{\omega}. \end{aligned}$$

The computational scheme is a parameter expanded Gibbs sampler: to integrate over $\psi_{i,t}$ in the logistic model, we utilize Pólya-Gamma data augmentation (Polson et al., 2013); to draw latent factor sequence $\beta_{i,1:T}$, we rely on forward filtering and backward sampling (FFBS); to sample $\boldsymbol{\omega}$, ϕ and Z, we use the Gibbs steps of Dahl et al. (2017). The MCMC algorithm is initialized by sampling from the posterior when $(\phi_1, \ldots, \phi_n)$ is modeled with a Dirichlet Process mixture model (e.g., when $f(X_i, X_j) = 1$) using the standard MCMC algorithms of Neal (2000). We run our MCMC algorithm for 20,000 iterations and discard the first 10,000 as a burn-in.

4.1 Sampling steps

There are nine different sampling steps required in the MCMC algorithm. Step 1 is for latent variable $H_{i,t}$. Step 2 samples a Pólya-Gamma auxiliary variable ζ (Polson et al., 2013). Conditional on $\zeta_{i,t}$, we sample the log odds of homelessness $\psi_{i,t}$ in Step 3. Given the sequence of log odds draws $\psi_{i,1:T}$ and draws $\tilde{\phi}_k$ and $Z_i = k$, we sample the latent factor sequence $\beta_{i,1:T}$ utilizing FFBS. Step 5 updates the innovation variance $\sigma^2_{\psi_i}$ by sampling from an inverse gamma full conditional distribution. Step 6 updates the concentration parameter α by sampling from a mixture of Gamma distributions (Escobar and West, 1995). Step 7, Step 8, and Step 9 are from the EPA sampling algorithm of Dahl et al. (2017).

1. For each i, t, sample the total number of people experiencing homelessness in metro i and year t, $H_{i,t}$, from a discrete distribution with support $[C_{i,t}, N_{i,t}]$. The probability mass for each possible value is

$$\begin{aligned}
&p(H_{i,t}|N_{i,t}, C_{i,t}, p_{i,t}, a_{i,t}, b_{i,t}) \propto \\
&\frac{\Gamma(H_{i,t}+1)}{\Gamma(C_{i,t}+1)\Gamma(H_{i,t}-C_{i,t}+1)} \frac{\Gamma(C_{i,t}+a_{i,t})\Gamma(H_{i,t}-C_{i,t}+b_{i,t})}{\Gamma(H_{i,t}+a_{i,t}+b_{i,t})} \times \dots \\
&\times \frac{\Gamma(a_{i,t}+b_{i,t})}{\Gamma(a_{i,t})\Gamma(b_{i,t})} \binom{N_{i,t}}{H_{i,t}} p_{i,t}^{H_{i,t}} (1-p_{i,t})^{(N_{i,t}-H_{i,t})}.
\end{aligned}$$

Sampling $H_{i,t}$ depends on prior beliefs about count accuracy $\pi_{i,t} \sim Beta(a_{i,t}, b_{i,t})$ in 3. We follow Glynn and Fox (2019) and specify

$$\begin{aligned}
E[\pi_{i,t}] &= \frac{0.95 \times C_{i,0}^{\text{Sheltered}} + 0.6 \times C_{i,0}^{\text{Unsheltered}}}{C_{i,0}}, \\
Var(\pi_{i,t}) &= 0.0015,
\end{aligned}$$

and compute

$$a_{i,t} = E[\pi_{i,t}] \left(\frac{(1-E[\pi_{i,t}])E[\pi_{i,t}]}{Var(\pi_{i,t})} - 1 \right), \tag{15}$$

$$b_{i,t} = \frac{Var(\pi_{i,t})}{E[\pi_{i,t}]^2} \left(\frac{a_{i,t}^2}{E[\pi_{i,t}]} + a_{i,t} \right). \tag{16}$$

2. For each i, t, sample the auxiliary Pólya-Gamma random variates to augment the total homeless variable, $\zeta_{i,t}|N_{i,t}, \psi_{i,t} \sim PG(N_{i,t}, \psi_{i,t})$.

3. For each i, t, sample the normally distributed

$$\begin{aligned}
&\psi_{i,t}|\zeta_{i,t}, N_{i,t}, H_{i,t}, Z_i = k, \tilde{\phi}_k, \sigma^2_{\psi_i} \\
\sim &N\left(\left(\zeta_{i,t} + \frac{1}{\sigma^2_{\psi_i}} \right)^{-1} \left(H_{i,t} - \frac{N_{i,t}}{2} + \frac{1}{\sigma^2_{\psi_i}} \left(\beta_{i,t} + X'_{i,t}\tilde{\phi}_k \right) \right), \left(\zeta_{i,t} + \frac{1}{\sigma^2_{\psi_i}} \right)^{-1} \right)
\end{aligned}$$

.

4. For each i, conditional on $\psi_{i,t}$, $Z_i = k$, and $\tilde{\phi}_k$, construct Dynamic Linear Model

$$y^*_{i,t} = F'_i \beta_{i,t} + \epsilon_{i,t} \tag{17}$$
$$\beta_{i,t} = A\beta_{i,t-1} + w_{i,t}, \tag{18}$$

   where $y^*_{i,t} = \psi_{i,t} - X'_{i,t}\tilde{\phi}_k$. Then jointly sample sample $\beta_{i,1:T}|\psi_{i,1:T}, Z_i = k, \tilde{\phi}_k, \sigma^2_{\psi_i}$ from a multivariate normal distribution using standard FFBS computations for Dynamic Linear Models. See West and Harrison (1997); Carter and Kohn (1994); Fruhwirth-Schnatter (1994).

5. For each i, sample the inverse gamma distributed $\sigma^2_{\psi_i}|Z_i = k, \tilde{\phi}_k, \beta_{i,1:T}, \psi_{i,1:T} \sim IG(a_\psi + \frac{T}{2}, b_\psi + \frac{1}{2}\sum_{t=1}^T \left(\psi_{i,t} - \beta_{i,t} - X'_{i,t}\tilde{\phi}_k\right)^2$.

6. Sample $\alpha|\tilde{\boldsymbol{\phi}}$ from a mixture of Gamma distributions as in Escobar and West (1995).

   - First, sample $g|\alpha \sim Be(1, \alpha + 1)$.
   - Compute ρ_g, where $\frac{\rho_g}{1-\rho_g} = \frac{q_n}{n(1-\log(g))}$.
   - Sample $\alpha|g, \tilde{\boldsymbol{\phi}} \sim \rho_g Ga(1 + q_n, 1 - \log(g)) + (1 - \rho_g)Ga(q_n, 1 - \log(g))$.

7. Update permutation $\boldsymbol{\omega}$ following the Metropolis-Hastings step of Dahl et al. (2017). First, propose a new permutation $\boldsymbol{\omega}^*$ by updataing r randomly chosen elements of the current permutation $\boldsymbol{\omega}$. The remaining $n - r$ elements of $\boldsymbol{\omega}^*$ are identical to $\boldsymbol{\omega}$. The updated elements of $\boldsymbol{\omega}^*$ are randomly shuffled, and $\boldsymbol{\omega}^*$ is accepted with probability $\min\left\{1, \frac{p(\boldsymbol{\pi}_n|\alpha,\delta=0,\boldsymbol{\omega}^*,f)}{p(\boldsymbol{\pi}_n|\alpha,\delta=0,\boldsymbol{\omega},f)}\right\}$. The integer r is chosen so that the acceptance probability is $\approx 40\%$, which in our simulations is achieved when $r = 90$.

8. For each ω_i, sample Z_{ω_i} from the the discrete full conditional distribution over $k = 0, 1, \ldots, q_n$,

$$\begin{aligned} &p(Z_{\omega_i} = k|Z_{(-\omega_i)}, \alpha, \boldsymbol{\omega}, \psi_{\omega_i,1:T}, \tilde{\boldsymbol{\phi}}, \sigma^2_{\psi_{\omega_i}}, \beta_{\omega_i,1:T}) \propto \\ &p(\boldsymbol{\pi}_n^{Z_{\omega_i}\to k}|\alpha, \delta = 0, \boldsymbol{\omega}, f)p(\psi_{\omega_i,1:T}|Z_{\omega_i} = k, \tilde{\phi}_k, \beta_{\omega_i,1:K}, \sigma^2_{\psi_{\omega_i}}). \end{aligned}$$

   Note that $p(\boldsymbol{\pi}_n^{Z_{\omega_i}\to k}|\alpha, \delta = 0, \boldsymbol{\omega}, f)$ is computed using equation 8 for the partition $\boldsymbol{\pi}_n^{Z_{\omega_i}\to k}$, which is constructed by assigning CoC ω_i to cluster k in the current partition $\boldsymbol{\pi}_n$. If $k = 0$ (corresponding to a new cluster), sample $\phi_0 \sim N(\mu_0, \Sigma_0)$.

9. For each $k = 1, \ldots, q_n$, sample $\tilde{\phi}_k|Z_{1:n}, \psi_{1:n,1:T}, \beta_{1:n,1:T}, \{\sigma^2_{\psi_i}\}_{i=1}^n \sim N(a^*, A^*)$, where

   - $A^* = \left(\Sigma_0^{-1} + \sum_{\{i:Z_i=k\}} \frac{1}{\sigma^2_{\psi_i}} \sum_{t=1}^T X_{i,t}X'_{i,t}\right)^{-1}$
   - $a^* = (A^*)^{-1}\left(\Sigma_0^{-1}\mu_0 + \sum_{\{i:Z_i=k\}} \frac{1}{\sigma^2_{\psi_i}} \sum_{t=1}^T X_{i,t}(\psi_{i,t} - \beta_{i,t})\right)$.

4.2 Posterior predictive distributions of homeless rates

Inferred relationships between homeless rates and CoC-predictors are best summarized by the posterior predictive distribution of the log odds of homelessness in a new community with predictor-vector $X_{n+1,T}$ when there are no latent factors, $p(\psi_{n+1,T}|\beta_{n+1,T}=0, X_{n+1,T}, C_{1:n,1:T}, N_{1:n,1:T})$. The posterior predictive is computed by integrating over $Z_{n+1}, \tilde{\boldsymbol{\phi}}$, and $\sigma^2_{\psi_{n+1}}$ in the joint distribution

$$p(\psi_{n+1,T}, Z_{n+1}, \tilde{\boldsymbol{\phi}}, \sigma^2_{\psi_{n+1}}|\beta_{n+1,T}=0, X_{n+1,T}, C_{1:n,1:T}, N_{1:n,1:T}). \tag{19}$$

We perform this integration numerically and construct samples of the posterior predictive homeless rate in a new CoC according to the following four step procedure.

1. For the m^{th} MCMC iteration, sample cluster assignment $Z^{(m)}_{n+1}$ for a new CoC according to the model implied probability mass function

$$Pr(Z^{(m)}_{n+1}=k|\alpha^{(m)}, \delta=0, f, \boldsymbol{\pi}(\omega^{(m)}_1, \ldots, \omega^{(m)}_n)) \tag{20}$$

$$= \begin{cases} \left(\frac{n}{\alpha^{(m)}+n}\right) \frac{\sum\limits_{\{\omega_s : Z_{\omega_s}=k\}} f(X_{n+1,T}, X_{\omega_s,T})}{\sum\limits_{s=1}^{n} f(X_{n+1,T}, X_{\omega_s,T})}, \text{ for } k=1,\ldots,q^{(m)}_n \\ \\ \frac{\alpha^{(m)}}{\alpha^{(m)}+n} \text{ for } k=0 \text{ (e.g., a new cluster) }. \end{cases} \tag{21}$$

2. Sample $\left(\sigma^2_{\psi_{n+1}}\right)^{(m)} \sim IG(a_\psi, b_\psi)$.

3. If $Z^{(m)}_{n+1}=k$ for $k=1,\ldots,q^{(m)}_n$, sample $\psi^{(m)}_{n+1} \sim N\left(X'_{n+1,T}\tilde{\phi}^{(m)}_k, \left(\sigma^2_{\psi_{n+1}}\right)^{(m)}\right)$. If $Z^{(m)}_{n+1}=0$, first draw $\tilde{\phi}^{(m)}_0 \sim N(\mu_0, \Sigma_0)$ and then sample $\psi^{(m)}_{n+1} \sim N\left(X'_{n+1,T}\tilde{\phi}^{(m)}_0, \left(\sigma^2_{\psi_{n+1}}\right)^{(m)}\right)$.

4. Transform to the homeless rate with the logistic transformation, $p^{(m)}_{n+1} = \frac{1}{1+e^{-\psi^{(m)}_{n+1}}}$.

While the functional form $\psi_{n+1,T} = X'_{n+1,T}\tilde{\phi}_k + \epsilon_{n+1,T}$ is locally linear in predictor space conditional on $Z_{n+1}=k$, $X_{n+1,T}$, and $\tilde{\phi}_k$, the marginal predictive distribution of $\psi_{n+1,T}$ is not necessarily linear as $X_{n+1,T}$ changes. Since cluster assignment prior probabilities depend on covariates, Z_{n+1} may also change as $X_{n+1,T}$ changes, and $\psi_{n+1,T}$ may exhibit nonlinear form as a function of $X_{n+1,T}$. This flexible functional form allows us to to identify structural changes in the relationship between homeless rates and CoC covariates, a main objective of the analysis.

The temperature τ in 6 controls the degree to which information is borrowed across predictor space when computing the posterior predictive. When predicting the homeless rate in a new metropolitan area, it makes sense to rely heavily on data from other big metropolitan areas rather than giving equal weight to data from rural CoCs. This common sense objective requires that pairwise similarity between CoCs rapidly decay in X, necessitating larger values of τ. A consequence of choosing $\tau = 0.35$ is that the posterior predictive distribution of the homeless rate in a new community

$$p_{n+1,T} | \beta_{n+1,T} = 0, X_{n+1,T}, C_{1:n,1:T}, N_{1:n,1:T}$$

is heavily informed by data from CoCs that are close neighbors in predictor space. We believe this an important feature of our model.

5 Results

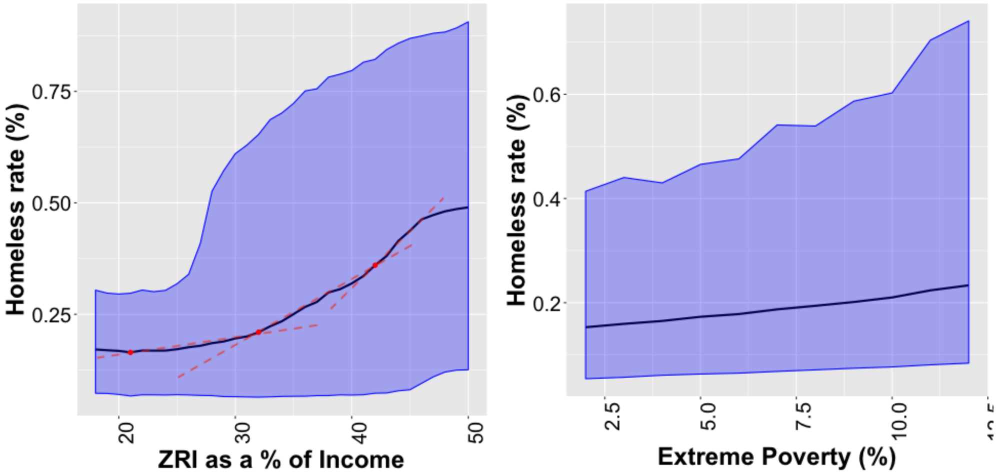
There are three main findings of our study: (i) there is an inflection point when ZRI reaches 32% of median income – after which the expected homeless rate in a community sharply increases; (ii) we identify six different clusters of CoC's that exhibit distinct geographic patterns; and (iii) unobserved factors in a CoC beyond poverty and housing affordability contribute meaningfully to increases (decreases) in homeless rates over time. In Section 5.1, we illustrate the complex nonlinear associations between homeless rates, housing affordability, and extreme poverty. In Section 5.2, we present findings from our cluster analysis. In Section 5.3, we examine the net contribution of additional unobserved factors to the overall homeless rate – allowing us to identify temporal trends in homeless rates that are not explained by housing affordability or poverty.

5.1 Inflection points in CoC-predictors

A primary objective of this analysis is to identify levels of housing affordability and extreme poverty which, once exceeded, predict significant increases in homeless rates. Identifying these inflection points can help communities prepare for rapid growth in homeless populations. In Figure 3, we summarize the relationship between homeless rates, housing affordability, and extreme poverty with the posterior predictive distribution computed in Section 4.2. The posterior predictive is a two-dimensional surface for the homeless rate over a grid of housing affordability and extreme poverty values. In Figure 3, we present cross sections of the surface at the sample average of both predictors. In Figure 3a, we predict the homeless rate as a function of housing affordability when extreme poverty is the sample average (6.64%). For example, we expect a homeless rate of ≈ 0.32% (y-axis) in a community where rental costs consume 40% (x-axis) of median income and extreme poverty is on par with the national average. San Diego is an example of a community with these characteristics. In 2017, the extreme poverty rate in San Diego was 6.26% and ZRI consumed 40.16% of median income. The estimated homeless rate in San Diego in 2017 was 0.37% – right in the middle of the predicted range.

To identify inflection points in housing affordability, we numerically evaluate the second derivative of $\tilde{\psi}(x_*) := E[\psi_{n+1,T} | C_{1:n,1:T}, N_{1:n,1:T}]$ for covariate vector $X_*' = \begin{bmatrix} 1 & x_* & 6.64 \end{bmatrix}$. To estimate the location of inflection points, we compute the posterior probability that the second derivative $\tilde{\psi}''(x_*)$ exceeds threshold κ, corresponding to structural changes in the slope of $\tilde{\psi}(x_*)$. This probability is computed with posterior samples in equation 22.

$$P\left(\tilde{\psi}''(x_*) > \kappa\right) \approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\left\{\tilde{\psi}^{(m)}(x_*+1) - 2\tilde{\psi}^{(m)}(x_*) + \tilde{\psi}^{(m)}(x_*-1) > \kappa\right\}} \tag{22}$$



(a) Housing affordability

(b) Poverty rate

Figure 3: Left: The posterior predictive distribution for homeless rates as ZRI/median income increases. The red points mark estimated inflection points and are connected by dashed red lines to aid visual clarity of piecewise linear relationships. Right: the posterior predictive distribution for the homeless rate as a function of extreme poverty. The shaded intervals illustrate the 90% predictive uncertainty intervals.

We classify x_* as an inflection point if $P\left(\tilde{\psi}''(x_*) > \kappa\right) > 0.5$. When we apply this procedure with $\kappa \geq 0.0125$, three clear inflection points emerge: 21%, 32%, and 42%. In Figure 3a, we have marked these inflection points in red and connected them with dashed red lines to aid in visual clarity. Observe that when ZRI as a percent of median income is between 21-32%, the rate of increase in the expected homeless rate is not nearly as sharp as the rate of increase from 32 - 42%. A third inflection point occurs at 42%, and the slope of the homeless rate increases even further until it begins to flatten at 46%. Observe that the expected homeless rate is approximately piecewise linear over the 21-32%, 32-42%, and 42-46% intervals. We believe that the slope flattens at 46% because the people most vulnerable to homelessness have already been impacted by the time affordability reaches 46%. The estimated 32% threshold is particularly noteworthy because it closely corresponds to the 30% definition of affordable housing used by HUD and the Census Bureau: when housing costs exceed 30% of income, a family is defined as cost burdened (HUD, 2018). When families become acutely cost burdened, we find that the expected homeless rate sharply increases.

The uncertainty intervals in Figure 3a are quite wide. When ZRI is 40% of median income, the 90% predictive interval for the homeless rate spans 0.07% on the low end to 0.8% on the high end. There are two important reasons for this. First, uncertainty in the quality of homeless counts $C_{1:n,1:T}$ and CoC-level populations $N_{1:n,1:T}$ is propagated to the posterior predictive. The second reason is the underlying nature of the CoCs themselves. Many exclusive communities have high costs of housing but very low homeless rates by design. This accounts for the wide range of homeless rates across communities with the same level of housing affordability, which we observe in the raw data (see Figure 1). As a result, the uncertainty intervals in Figure 3a are wide to reflect significant variation in homeless rates across different communities with the same level of housing affordability.

In Figure 3b, we present the cross section of the predicted homeless rate as a function of extreme poverty for a community where ZRI is 28% of income, the sample average. The predictor vector is $X'_* = \begin{bmatrix} 1 & 28 & x_* \end{bmatrix}$. We interpret Figure 3b as following: the expected homeless rate is 0.20% (y-axis) in a community where 8% (x-axis) of the population lives in extreme poverty and relative housing costs are on par with the national average. The 90% predictive interval ranges from 0.07% to 0.59%. In Albuquerque, NM (7.75% in extreme poverty, 28.7% for ZRI/median income) we estimate that in 2017 the homeless rate was 0.32% – again within the predicted range. Observe in Figure 3b that the relationship between homeless rates and extreme poverty is characterized by a single line. There are no estimated inflection points in the rate of extreme poverty, as the slope of the line is uniform.

5.2 Clusters of CoCs

There is significant interest from a policy perspective in identifying a peer group of CoCs likely to benefit from the same type of intervention. To form these peer groups, we identify frequent co-occurences of CoCs i and j in the same cluster and compute a pairwise similarity matrix from MCMC samples of Z_i and Z_j. Based on the posterior probability of CoCs i and j sharing a cluster, we utilize the adjusted Rand index of Fritsch and Ickstadt (2009) to compute a point estimate of the partition $\hat{\boldsymbol{\pi}}_{386}$.

We find six different clusters; however, most CoCs (377 of 386) are assigned to clusters one, two, and three. Observe in Table 2 that, of the first three clusters, cluster one has (on average) the lowest homeless rate (0.09%), the most affordable housing (26.69%) and the lowest rate of extreme poverty (5.96%). Of clusters one through three, cluster three has (on average) the highest homeless rate (0.63%), the least affordable housing (39.44%), and the highest rate of extreme poverty (7.39%). The largest cluster – both by number of CoCs and by population – is cluster two, which is home to 50.24% of the U.S. population. While only 13.86% of the total U.S. population lives in cluster three, it contains 45.59% of the homeless included in the 2017 PIT counts.

Cluster	1	2	3	4	5	6
Size (# of CoCs)	136	192	49	7	1	1
Share of Total Pop (%)	34.24	50.24	13.86	1.20	0.32	0.14
Share of PIT Count (%)	13.42	40.39	45.59	0.21	0.16	0.23
Homeless Rate (%)	0.09	0.19	0.63	0.04	0.09	0.37
Affordability Rate	26.69	29.81	39.44	27.71	23.50	33.60
Poverty Rate (%)	5.96	6.83	7.39	7.32	4.21	5.47

Table 2: Cluster characteristics in EPA partitioning: The Share of Total Pop (%) and Share of PIT Count (%) are the percentage of the total US population and HUD counted number of homeless in each cluster in 2017. Homeless Rate (%) is the mean estimated homeless rate. Affordability is the cluster-level mean of ZRI as a percentage of median income, and poverty is the cluster-level mean of the extreme poverty rate.

Although the model contains no specific mechanism for spatial patterns in homeless rates, there is clear spatial structure in our cluster assignments. Observe that cluster one is common in the Midwest, Mid-Atlantic, and parts of the South. Most of New England, Georgia, Florida, the mountain west and southwest United States are assigned to cluster two. Cluster three occupies much of the west coast – including San Francisco, Portland (OR), and Seattle – as well as eastern metropolitan areas in Boston, New York City, Washington, D.C., and Atlanta. The communities in cluster three, with ZRI at 39% of median income on average, are well above the inflection point of 32% identified in Section 5.1. Figure 4 is a data-driven confirmation of observations made by homeless coordinators and policy makers around the country: while homeless counts are generally falling in most parts of the United States, there are pockets on both coast where states of emergency have been declared to combat homeless crises.

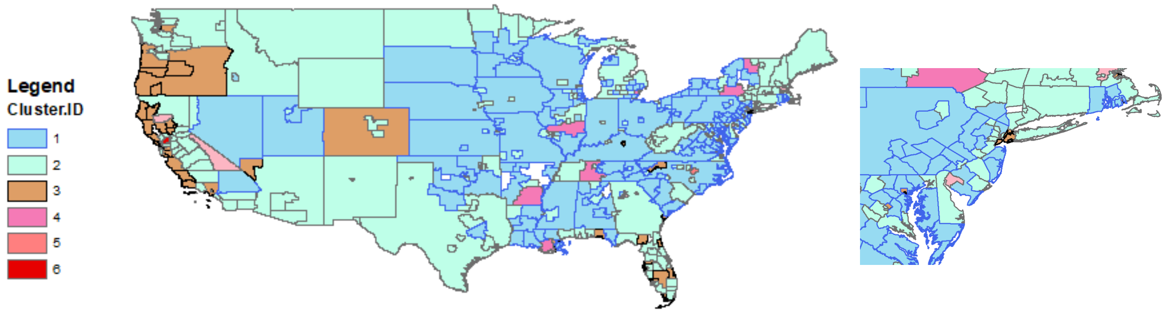


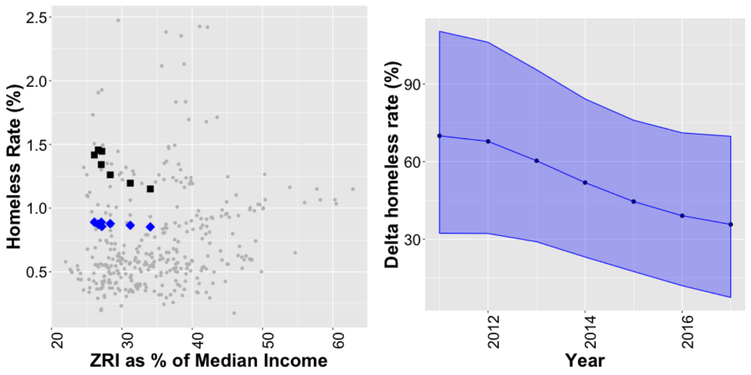
Figure 4: Map of clusters in the continental United States (left) and the northeast corridor (right) from Washington, D.C. to Boston, MA. Clusters exhibit strong spatial structure.

Clusters four through six correspond to CoCs that are relatively unique. Cluster four has seven members in the south, midwest, and upstate New York – the Southeast Arkansas, Houma-Terrebonne/Thibodaux (Louisiana), Central Tennessee, South Central Illinois, Dearborn/Wayne County (Michigan), Frankin County (New York), and Binghamtom (New York) CoCs (see Figure 4). In these communities, the average homeless rate is very low (0.04%) considering the high rate of extreme poverty (7.32%). The sole member of cluster five is Raleigh/Wake County, North Carolina, which has a low homeless rate, very affordable housing, and low poverty rates (see Table 2). The sole member of cluster six is the Vallejo/Solano County CoC in the San Francisco Bay area, which stands out for its relatively high homeless rate despite low poverty rates. The Vallejo / Solano County CoC has a particularly strong association between the homeless rate and worsening housing affordability.

5.3 CoC-level latent factors

There are many dimensions of a community. Poverty and housing affordability, while important covariates of a CoC, may not adequately explain variation in homeless rates – particularly in the presence of policy interventions aimed at reducing homelessness. To account for the many unobserved contributors to homelessness in a community, we include community-level dynamic latent factors $\beta_{i,1:T}$ in our statistical model. We interpret $F_i'\beta_{i,t}|C_{1:n,1:T}, N_{1:n,1:T}$, as the deviation of the homeless rate in CoC i from the rate expected of CoCs with similar covariates in the same cluster.

The Atlanta Continuum of Care provides an illustrative example of the role that latent factors play in our analysis. Atlanta, a member of cluster three in Section 5.2, has a particularly high homeless rate (0.93%) for a CoC with housing costs at 34% of median income in 2017. Relative to peer CoCs in cluster three with similar housing costs, the homeless rate in Atlanta is significantly higher than expected (see Figure 5a). While the high homeless rate in Atlanta is partly explained by the fact that 12% of the population lives in extreme poverty, poverty and housing costs are an incomplete accounting of the factors at play. Observe in Figure 5a that the estimated homeless rates in 2011-2017 (squares) are significantly higher than the homeless rates predicted by housing affordability and extreme poverty alone (diamonds). The underprediction indicates that other factors are contributing to homelessness, which we model with the latent factor $F_i'\beta_{i,t}$. Since latent factors in Atlanta are adding to the homeless rate beyond the rate expected of peers in cluster three with similar covariates, the posterior distribution $F_i'\beta_{i,T}|C_{1:n,1:T}, N_{1:n,1:T}$ concentrates on positive values (Figure 5b). We interpret Figure 5b as the percent increase in the predicted homeless rate from a model that includes $F_i'\beta_{i,t}$ compared to the predicted rate when $F_i'\beta_{i,t} = 0$, expressed mathematically as $100 \times \left(\frac{1+\exp\{-X_{i,t}'\phi_i\}}{1+\exp\{-F_i'\beta_{i,t}-X_{i,t}'\phi_i\}} - 1\right)$. The negative trend observed in Figure 5b also helps explain why the homeless rate in Atlanta has fallen over the years 2011 to 2017, despite the fact that housing affordability has deteriorated from 27% of income in 2011 to 34% in 2017. The important takeaway is that some combination of factors in Atlanta beyond housing affordability and poverty are contributing to this lowered homeless rate, and we estimate this net factor for each CoC with the the posterior $F_i'\beta_{i,t}|C_{1:n,1:T}, N_{1:n,1:T}$. The latent factor distribution over time provides a mechanism to evaluate the CoC's changing environment for homelessness – including policy interventions.



(a) Model Fit

(b) Latent factors

Figure 5: Atlanta Continuum of Care (GA-500). Left: Model fit for the homeless rate including latent factors (squares); the model fit for the homeless rate excluding latent factors (diamonds); and the homeless rates of other CoCs in cluster three (circles). Right: Posterior distribution for the percentage increase in the homeless rate associated with latent factors in Atlanta from 2011-2017.

6 Discussion

In this paper, we present a Bayesian nonparametric model of community-level homeless rates. The EPA-based mixture model shares information across CoCs where homeless rates are similarly related to covariates of a community, and we utilize posterior predictive distributions to identify structural changes in homeless rates as a function of housing affordability and extreme poverty. A main finding of the analysis is that the expected homeless rate in a community sharply increases once ZRI exceeds 32% of the median income – a finding that closely matches the federal definition of affordable housing (HUD, 2018). We identify three dominant clusters of CoCs that exhibit common relationships between homelessness and community features. Among the three main clusters, the lowest homeless rate, most affordable housing, and lowest extreme poverty rate are found in cluster one. Cluster three communities have, on average, the highest homeless rate, the least affordable housing, and the most poverty.

Our findings extend prior research that has examined the overall relationship between community-level factors and homelessness in an important way. We show that the relationship between homeless rates, housing affordability, and extreme poverty follow a nonlinear functional form. This stands in contrast to prior studies that have almost exclusively assumed the relationship between such factors and homelessness to be linear. Our relaxation of this assumption reveals important policy-relevant findings. For example, we find that maintaining a rent/income ratio less than 32% may be an important target for communities in order to avoid sharp increases in homelessness.

The study also provides new insight into geographic patterns of homelessness in the United States. A relatively small number of cities with large populations (cluster 3) are experiencing surges in homelessness related to unaffordable housing and extreme poverty. The average housing affordability metric is higher in cluster three (39.44%) than the 32% break point we identify, which partly explains rapid growth in the homeless populations of many of these CoCs. Communities in clusters one and two are not nearly as cost burdened, with average housing affordability measures of 26.7% and 29.8%, respectively. The majority of the United States is less sensitive to increases in housing costs than the 49 communities in cluster 3. This may explain why, despite increased homelessness in cluster 3 cities like Los Angeles, New York, and Seattle, homelessness in the United States as a whole has declined since the recession of 2008.

Prior research on community-level determinants of homelessness was motivated by policy considerations: Factors identified as key drivers of higher (or lower) rates of homelessness may subsequently be used by communities as policy levers to be pulled in their efforts to address homelessness. However, prior research in this vein operated under the implicit assumption that pulling the same levers with the same strength and in the same direction will have an identical effect regardless of the community in question. Our findings suggest that such an assumption is likely to be incorrect, and that communities would be wise to take a more nuanced approach in how they contend with structural factors in seeking to reduce homelessness. More concretely, our identification of six clusters of communities based on rental costs, household income, and the rate of extreme poverty points to the potential need for at least six distinct approaches for offsetting the respective impact of these factors on homelessness in a community. Our estimation of community-level latent factors adds even more nuance that might influence policy strategies. Comparing the relative contributions of latent factors, housing affordability, extreme poverty, and the cluster baseline to the overall rate of homelessness in a community can provide additional insight into which policy levers may be most impactful for individual communities.

A limitation of the current study is our use of the CoC as the primary observational unit. Many CoCs are geographically large, with Rhode Island, North Dakota, South Dakota, and Wyoming each representing statewide CoCs. Housing affordability and extreme poverty measures at the CoC-level may conceal dynamics of local markets, adding to the inference challenge in some larger CoCs. While we do not know of better nationwide data on homeless populations, we recognize the challenge of working with PIT counts to investigate the relationship between homelessness and community covariates. This research augments but is not a substitute for the invaluable local knowledge of CoC-coordinators and service organizations in addressing the needs of homeless populations in individual communities.

References

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. The Annals of Statistics, 2(6):1152–1174.

Blei, D. M. and Frazier, P. I. (2011). Distance dependent chinese restaurant processes. Journal of Machine Learning Research, 12(Aug):2461–2488.

Bun, Y. (2012). Zillow rent index: Methodology. https://www.zillow.com/research/zillow-rent-index-methodology-2393/. [Online; accessed 04/2/2017].

Byrne, T. (2018). HUD-CoC-Geography-Crosswalk. https://github.com/tomhbyrne/HUD-CoC-Geography-Crosswalk.

Byrne, T., Munley, E. A., Fargo, J. D., Montgomery, A. E., and Culhane, D. P. (2013). New perspectives on community-level determinants of homelessness. Journal of Urban Affairs, 35(5):607–625.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. Biometrika, 81(3):541–553.

Corinth, K. C. (2015). Ending homelessness: More housing or fewer shelters? AEI Economics Working Papers 863788, American Enterprise Institute.

Culhane, D. P., Lee, C., and Wachter, S. M. (1996). Where the homeless come from: A study of the prior address distribution of families admitted to public shelters in New York City and Philadelphia. Housing Policy Debate, 7(2):327–365.

Dahl, D. B. (2008). Distance-based probability distribution for set partitions with applications to bayesian nonparametrics. JSM Proceedings, Section on Bayesian Statistical Science, American Statistical Association, Alexandria, VA.

Dahl, D. B., Day, R., and Tsai, J. W. (2017). Random partition distribution indexed by pairwise information. Journal of the American Statistical Association, 112(518):721–732.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, 90(430):577–588.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. The Annals of Statistics, pages 209–230.

Fritsch, A. and Ickstadt, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. Bayesian Anal., 4(2):367–391.

Fruhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. Journal of Time Series Analysis, 15(2):183–202.

Glynn, C. and Fox, E. B. (2019). Dynamics of homelessness in urban America. Annals of Applied Statistics, 13(1):573–605.

Hopper, K., Shinn, M., Laska, E., Meisner, M., and Wanderling, J. (2008). Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods. American Journal of Public Health, 98(8):1438–1442.

HUD (2017). Pit and hic data since 2007. https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/. [Online; accessed 08/7/2018].

HUD (2018). Affordable housing. www.hud.gov/program_offices/comm_planning/affordablehousing. [Online; accessed 12/2/2018].

Lee, B. A., Price-Spratlen, T., and Kanan, J. W. (2003). Determinants of homelessness in metropolitan areas. Journal of Urban Affairs, 25(3):335–356.

MacEachern, S. N. (2000). Dependent dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University, pages 1–40.

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.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9(2):249–265.

Page, G. L. and Quintana, F. A. (2015). Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates. Bayesian Anal., 10(2):379–410.

Page, G. L. and Quintana, F. A. (2016). Spatial product partition models. Bayesian Anal., 11(1):265–298.

Page, G. L. and Quintana, F. A. (2018). Calibrating covariate informed product partition models. Statistics and Computing, 28(5):1009–1031.

Park, J.-H. and Dunson, D. B. (2010). Bayesian generalized product partition model. Statistica Sinica, pages 1203–1226.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using polya-gamma latent variables. Journal of the American Statistical Association, 108(504):1339–1349.

Quigley, J. M., Raphael, S., and Smolensky, E. (2001). Homeless in america, homeless in california. Review of Economics and Statistics, 83(1):37–51.

Rukmana, D. (2008). Where the homeless children and youth come from: A study of the residential origins of the homeless in Miami-Dade County, Florida. Children and Youth Services Review, 30(9):1009–1021.

Shahbaba, B. and Neal, R. (2009). Nonlinear models using dirichlet process mixtures. Journal of Machine Learning Research, 10(Aug):1829–1850.

West, M. and Harrison, J. (1997). Bayesian Forecasting and Dynamic Modeling. Springer-Verlag, New York, NY, second edition.