Inflection points in community-level homeless rates*

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

Abstract

Previous studies that quantify the relationship between homeless rates and features of a community typically assume a global linear relationship. This linear model assumption precludes the possibility of inflection points in homeless rates – thresholds in quantifiable metrics of a community which, 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 develop a Dirichlet process mixture model that allows clusters of communities with similar features to exhibit 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 empirical evidence for the widely used definition of a housing cost burden at 30% of income. The Dirichlet process model also generates clusters that share common characteristics and exhibit distinct geographic patterns – yielding insight into the homelessness and housing affordability crises in large metropolitan areas 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[1]. Quantifying the association between homeless rates and features 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.

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

[†] 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).

A large collection of literature has investigated statistical associations between features of a community and homelessness (Byrne et al., 2013; Lee et al., 2003; Quigley et al., 2001); however, existing statistical models for variation in 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 (2018) 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 estimates for model parameters.

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 features 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.

In this paper, we have to primary objectives:

(O_1) Flexibly estimate the relationship between community features 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.

To achieve these goals, we develop a Dirichlet process (Ferguson, 1973) mixture model of homeless rates that partitions communities into clusters where the relationship between rates of homelessness and features of communities is common. Homeless rates are modeled as the unobserved probability of homelessness in a Bayesian logistic regression. Building on Glynn and Fox (2018), a distinction is made between the counted and total number of homeless, and sampling variability in the homeless counts and uncertainty in the size of the total homeless population flow through to the model for the homeless rate. Three important aspects of our model are (i) the number of clusters; (ii) cluster membership; and (iii) the relationship between community features and homelessness within clusters are all jointly estimated as part of the inference procedure. A Markov chain Monte Carlo algorithm is developed that seamlessly combines the Polya-Gamma data augmentation strategy of Polson et al. (2013) with Neal's Algorithm 2 for Dirichlet process mixtures (Neal, 2000) and a forward filtering backward sampling (FFBS) algorithm to account for community-specific trends. An important consequence of our Bayesian nonparametric model for homeless counts is the ability to flexibly estimate increases in homeless rates with locally linear regressions.

In this study, we focus on three aspects 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, 2018) 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, taking into account that high rents in big cities are also typically supported by relatively higher salaries whereas lower rents in rural areas may still represent a significant portion of one's income. By focusing on median housing costs as a share of median income, we are able to 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 prone to homelessness, we also include the percent of a community living in extreme poverty in our statistical analysis.

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 381 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 Dirichlet process 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 localized posterior predictive distributions for the relationship between homeless rates and community features and identify clusters of CoCs 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 assess variation in homeless rates, it is essential to account for variation in the size of

[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.

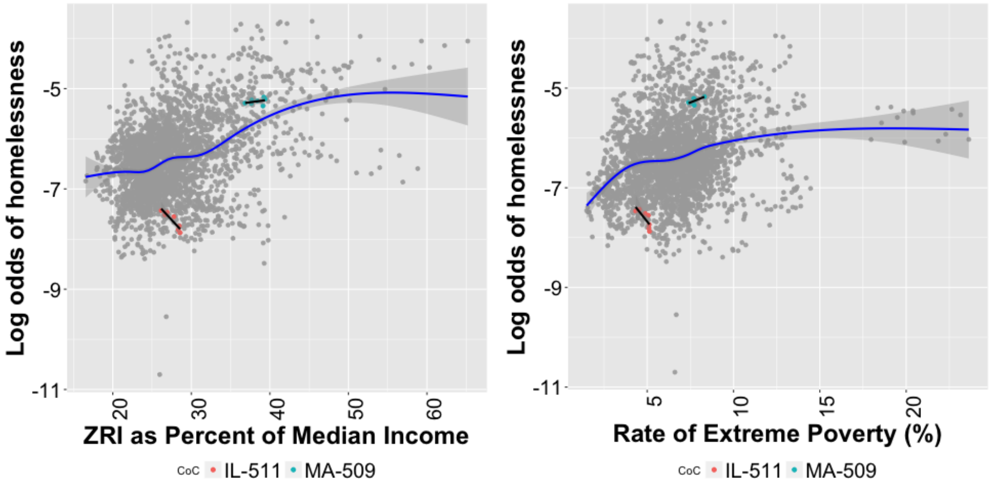
CoCs; however, the total population of a CoC is not reported by HUD. Discrepancies between geographic boundaries of CoCs and boundaries of geographic units for which total population estimates are made available by the U.S. Census Bureau mean that total population estimates for some CoCs are not readily 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 from 2011-2016. 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 features 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 (2018) and utilize a custom-computed variant of ZRI. The critical difference in the rental data for this analysis and that used by Glynn and Fox (2018) is that in the present study, Zillow 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, 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).

	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 features of New York City CoC (NY-600), including all five burroughs of New York City.

While countless features of a community are potentially associated with homelessness – including apartment vacancy rates, unemployment rates, demographics, etc. – most (if not all) are highly correlated with the features that we have included in our analysis: cost of rental housing, median income, and rates of extreme poverty. Including many highly correlated predictors in a statistical model presents estimation problems that are avoided by focusing on a few important predictive features. Figure 1 demonstrates that as both ZRI (as a percentage of median income) and the rate of extreme poverty increase, so too does the estimated log odds of homelessness. In Figure 1a, observe that the data strands for the Cook County CoC (IL-511) and the Cambridge (MA) CoC exhibit very different associations with ZRI / Median Income. A single linear model is too rigid to realistically model 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 overcome 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 Dirichlet process prior (Ferguson, 1973). As atoms from the Dirichlet process are replicated across CoCs, the infinite mixture model forms clusters of CoCs that share similar associations between homeless rates and CoC-level predictors. Pooling data at the cluster level facilitates sharper inference of shared parameters than would be possible if each CoC were endowed with its own parameter. The information-borrowing strategy allows us to overcome the limited sample size of each CoC, which has only seven years of PIT data, and it further provides a well-defined peer group of CoCs based on the shared pattern of variation in homeless rates. While each CoC's homeless rate – conditional on its cluster assignment – is a linear regression in the latent log odds space, integrating over cluster assignments locally in predictor space yields a localized posterior predictive distribution that flexibly models the form of association between homeless rates and CoC-predictors – providing a model-based strategy for inferring potential inflection points in homeless rates. Before introducing the modeling innovation in section 3.2, we 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 in Section 3.1.

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 (2018) 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 Glynn and Fox (2018) framework here.

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

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

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. Sampling from the posterior predictive distribution $p(N^*_{i,t}|N_{i,1:T})$ generates samples of the CoC's population that are informed by the ACS data and provides a mechanism for propagating uncertainty in the CoC populations to predictions about the underlying size of the homeless population.

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,

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

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 distinctly less than $H_{i,t}$, is modeled as a conditionally binomial random variable

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

The parameter $\pi_{i,t} \in [0, 1]$ is the probability that a person who is homeless will be counted as homeless. Refer to Glynn and Fox (2018) for a full discussion of prior choices for $\pi_{i,t} \sim Beta(a_{i,t}, b_{i,t})$ and their consequences for inference on changes in homeless rates. 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.

These three model components are coupled with a two stage binomial thinning. In the first stage (equation 2), the total CoC population is reduced to the total number of homeless, a step that depends critically on the homeless rate $p_{i,t}$. In the second stage (equation 3), the unobserved total number of homeless is reduced to the counted number of homeless, $C_{i,t}$, a step that depends on one's prior beliefs about count accuracy. We adopt the priors utilized by Glynn and Fox (2018) to carry out our analysis.

3.2 Dirichlet process mixture model for $\psi_{i,t}$

The novel modeling contribution of this paper is a Bayesian nonparametric model for $p_{i,t}$ based on the Dirichlet process prior of Ferguson (1973). As outlined in 2, we model the total number of homeless $H_{i,t}$ with a Bayesian logistic regression. Here, we transform $p_{i,t}$ to the real line with a logit transformation

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

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 $F'_{i,t}\beta_{i,t}$ and the regression $X'_{i,t}\phi_i$. We address each component in turn.

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. To induce shared regression coefficients in groups of CoCs, we model ϕ_i with a discrete random measure G, where G itself is drawn from a Dirichlet process prior.

$$\phi_i \sim G \tag{5}$$
$$G \sim DP(\alpha G_0) \tag{6}$$

The Dirichlet process prior for G places prior probability on a countable sequence of p-dimensional vectors $(\phi^{(1)}, \phi^{(2)}, \phi^{(3)}, \ldots)$, each with probability mass $(\omega^{(1)}, \omega^{(2)}, \omega^{(3)}, \ldots)$. The atoms of G, denoted $\phi^{(l)}$, are drawn from base measure G_0 with support on $\mathbb{R}^p$, and the weights $\omega^{(l)}$ are recursively constructed utilizing the stick-breaking representation of Sethuraman (1994). The weights $\omega^{(l)} = \gamma_l \prod_{j=1}^{l-1}(1 - \gamma_j)$ depend on γ_j (for $j = 1, \ldots$), which are drawn independently from a $Beta(1, \alpha)$ distribution. The discrete probability measure for ϕ_i is then $\sum_{l=1}^{\infty} \omega^{(l)} \delta_{\phi^{(l)}}$. One consequence of the discrete probability measure G for the set of all $\{\phi_i\}_{i=1}^{386}$ is that multiple CoCs may share the same atom $\phi^{(l)}$, inducing a partition of CoCs into clusters that share the same relationship between $\psi_{i,t}$ and $X_{i,t}$.

Though the form of $X_{i,t}$ may be customized by the modeler, in this study we include a leading one in $X_{i,t}$ (e.g., $X_{i,t} = \begin{bmatrix}1 & \dots\end{bmatrix}'$). The leading one results in a shared cluster-level intercept – or expected rate of homelessness – that is unrelated to housing costs, economic variables, and poverty. One way of interpreting the cluster-level intercept is as the expected rate of chronic homelessness in a particular group of communities.

The cluster-level regression coefficient ϕ_i models variation in $\psi_{i,t}$ associated with predictors $X_{i,t}$; however, there are many features of a community that are either not directly observed or excluded from $X_{i,t}$. To account for these unobserved local features, we include a CoC-level dynamic latent factor $\beta_{i,t}$ – 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 $\beta_{i,t}$ 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 state-space model

$$\beta_{i,t} = A\beta_{i,t-1} + w_t, \qquad w_t \sim N(0, W_t). \tag{7}$$

The dynamic latent factor model in 7 makes two important contributions: first, $\beta_{i,t}$ 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 $F'_{i,t} = \begin{bmatrix}1 & 0\end{bmatrix}$.

The number of clusters in our Dirichlet process model 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 particular regression lines is tight, and many clusters are needed to explain variation in the 386 CoCs. If the innovation variance $\sigma^2_{\psi_i}$ is large, 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.

$$\sigma^2_{\psi_i} \sim IG(a_\psi, b_\psi) \tag{8}$$

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.

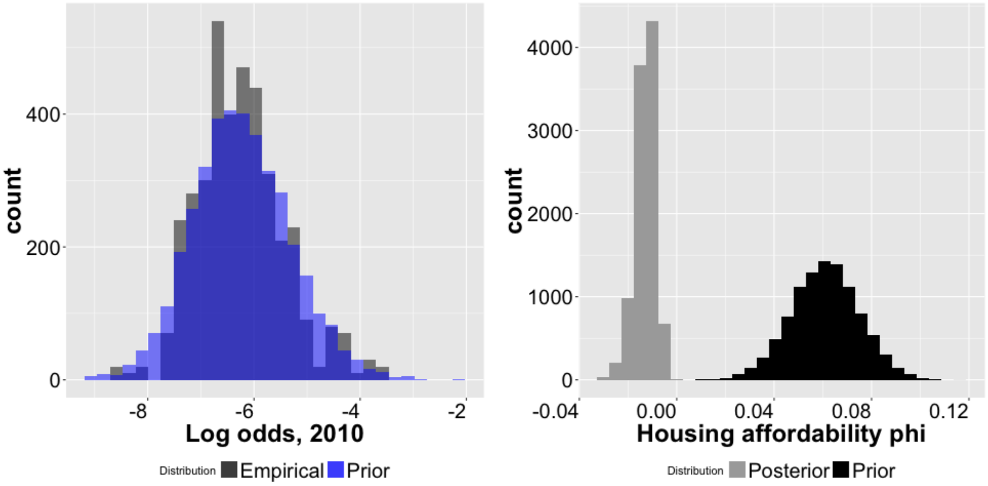
3.3 Prior choices

Prior distributions for $(\beta_{i,0}, \alpha, \sigma^2_{\psi_i})$ and base measure G_0 are chosen by matching the first two moments of the implied prior distribution at time zero 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 distribution of log odds of homelessness in

2010, denoted $\psi_{i,0}$, is unimodal and symmetric with a mean of -6.24 and a variance of 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 for 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 base measure G_0. We choose G_0 to be a $p-$dimensional Gaussian distribution with mean μ_0 and variance Σ_0. In Section 3.2, we noted that the cluster-level intercept may be interpreted as the rate of homelessness that is unrelated to community features. We interpret this as the rate of chronic homelessness in a community and utilize PIT counts from 2010 on chronic homelessness to inform the first element $\mu_0 = -8.28$. Remaining elements of μ_0 are chosen so that $X^{(2)}_{i,0}\mu^{(2)}_0 + \ldots + X^{(p)}_{i,0}\mu^{(p)}_0 = (\psi_{i,0} - 8.28)$ receive equal contributions and $\mu^{(2)}_0 = \ldots = \mu^{(p)}_0 = \frac{\psi_{i,0}-8.28}{X^{(2)}_{i,0}+\ldots+X^{(p)}_{i,0}}$. When we include predictors for housing affordability (measured by ZRI as a share of median income) and the rate of extreme poverty, the predictor vector is $X'_{i,t} = \begin{bmatrix}1 & \frac{ZRI_{i,t}}{MedianIncome_{i,t}} & ExtPoverty_{i,t}\end{bmatrix}$, and the mean of G_0 is $\mu'_0 = \begin{bmatrix}-8.28 & 0.061 & 0.061\end{bmatrix}$.

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,

$$\begin{aligned} Var(\psi_{i,0}) &= E[Var(\psi_{i,0}|\beta_{i,0}, \phi_i, \sigma^2_{\psi_i})] + Var(E[\psi_{i,0}|\beta_{i,0}, \phi_i, \sigma^2_{\psi_i}]) && (9)\\ &= E[\sigma^2_{\psi_i}] + F'_{i,0}Var(\beta_{i,0})F_{i,0} + X'_{i,0}Var(\phi_i)X_{i,0}. && (10)\end{aligned}$$



(a) Moment Matching

(b) Posterior update

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

We begin by fixing $Var(\beta_{i,0}) = 0.1$ to allow for meaningful systematic (as opposed to idiosyncratic) deviations in a community's homeless rate from the homeless rate of the cluster. The variance of G_0, denoted Σ_0 is chosen to encode the belief that our most uncertain component is the intercept, the chronic 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 they do 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 count process is robust and stable from one year to the next, while in other CoCs, the observed count fluctuates significantly due to random local factors such as weather, changes in count methodology, volunteer turnout, and funding levels. We follow Escobar and West (1995) in modeling the concentration parameter of the Dirichlet process with the conventional $\alpha \sim Ga(1, 1)$. We note that prior choices for $Var(\beta_{i,0})$, α and $\sigma^2_{\psi_i}$ significantly impact the 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.

4 Markov Chain Monte Carlo

Our objective is to sample from the posterior distribution

$$p(\phi_{1:K}, Z_{1:386}, \beta_{1:386,1:T} | N_{1:25,1:T}, C_{1:25,1:T}), \tag{11}$$

where $Z_i = k$ is the cluster assignment variable that includes CoC i in the group sharing regression coefficient ϕ_k. Recall that $\psi_{i,t}$ is a parameter in the Bayesian logistic regression that depends on $H_{i,t}$, the latent variable for the size of the total homeless population (see 2). Our computational strategy is to condition on observations $N_{i,t}$ and $C_{i,t}$ while numerically integrating latent variables $H_{i,t}$ and $\psi_{i,t}$ from the joint posterior

$$\begin{aligned}&p(\phi_{1:K}, Z_{1:386}, \beta_{1:386,1:T} | N_{1:25,1:T}, C_{1:25,1:T})\\ &= \int p(\psi_{1:386,1:T}, H_{1:386,1:T}, \phi_{1:K}, Z_{1:386}, \beta_{1:386,1:T}, | N_{1:25,1:T}, C_{1:25,1:T}) dH_{1:386,1:T} d\psi_{1:386,1:T}.\end{aligned}$$

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 infer latent factor sequence $\beta_{i,1:T}$, we rely on forward filtering and backward sampling (FFBS, Carter and Kohn (1994); Fruhwirth-Schnatter (1994)); to make inference on ϕ and Z, we use Neal's algorithm 2 (Neal, 2000). We run our MCMC algorithm for 50,000 iterations and discard the first 25,000 as a burn-in. The MCMC simulation took approximately 12 hours to run on a MacBook Pro.

4.1 Sampling steps

There are eight different sampling steps required in the MCMC algorithm. Step 1 is for latent variable $H_{i,t}$. 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 choose $a_{i,t}$ and $b_{i,t}$ by specifying the prior mean ($E[\pi_{i,t}]$) and variance ($Var(\pi_{i,t})$), which implies that

$$a_{i,t} = E[\pi_{i,t}]\left(\frac{(1-E[\pi_{i,t}])E[\pi_{i,t}]}{Var(\pi_{i,t})} - 1\right), \tag{12}$$

$$b_{i,t} = \frac{Var(\pi_{i,t})}{E[\pi_{i,t}]^2}\left(\frac{a_{i,t}^2}{E[\pi_{i,t}]} + a_{i,t}\right). \tag{13}$$

We follow Glynn and Fox (2018) and specify prior mean $E[\pi_{i,t}]$ and $Var(\pi_{i,t})$ based on the proportion of the homeless population in each CoC that is unsheltered in 2010 and the assumption that 95% of sheltered homeless are counted while 60% of unsheltered homeless are counted.

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 from the Dirichlet process ϕ_k and $Z_i = k$, we sample the latent factor sequence $\beta_{i,1:T}$ utilizing FFBS. Step 5 and Step 6 are from Neal's Algorithm 2, which is closely related to algorithms developed by Bush and MacEachern (1996) and West et al. (1994). Step 7 updates the innovation variance $\sigma^2_{\psi_i}$ by sampling from an inverse gamma full conditional distribution. Step 8 updates the Dirichlet process concentration parameter α by sampling from a mixture of Gamma distributions.

To simplify presentation of the algorithm, we consider the case where $N_{i,t}$ is assumed to be the actual CoC population size. A straightforward modification of this algorithm allows for sampling a synthetic population $N^*_{i,t}$ from the posterior predictive distribution $p(N^*_{i,t}|N_{i,1:T})$ to propagate uncertainty in CoC-level populations to estimates of other model parameters. See MCMC sampling steps 1-5 (and prior choices therein) in Section 5 of Glynn and Fox (2018) for a detailed procedure to sample from $p(N^*_{i,t}|N_{i,1:T})$. Modify the algorithm below by replacing $N_{i,t}$ with the synthetic randomly sampled population $N^*_{i,t}$.

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

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 and t, sample the normally distributed $\psi_{i,t}|\zeta_{i,t}, N_{i,t}, H_{i,t}, Z_i = k, \phi_k, \sigma^2_{\psi_i}$.

4. For each i, sample $\beta_{i,1:T}|\psi_{i,1:T}, Z_i = k, \phi_k, \sigma^2_{\psi_i}$ from a multivariate normal distribution using standard FFBS computations.

5. For each i, sample $Z_i|\phi, \sigma^2_{\psi_i}, \beta_{i,1:T}$ following algorithm 2 in Neal (2000).

6. For each k, sample $\phi_k|Z_{1:386}, \psi_{1:386,1:T}, \beta_{1:386,1:T}, \{\sigma^2_{\psi_i}\}_{i=1}^{386}$ from a multivariate normal distribution.

7. For each i, sample $\sigma^2_{\psi_i}|Z_i = k, \phi_k, \beta_{i,1:T}, \psi_{i,1:T}$ from an inverse gamma distribution.

8. Sample $\alpha|\phi_{1:K}$ from a mixture of Gamma distributions as in Escobar and West (1995).

4.2 Approximate posterior predictive distributions

Inferred relationships between homeless rates and CoC-predictors are best summarized by the posterior predictive distribution of the homeless rate in a new community with predictor-vector X_*. Define $\psi_* = X_*'\phi_*$ to be the contribution of X_* to the log odds of homelessness. The model implied (5,6) posterior predictive distribution for ϕ_* is a mixture of base measure G_0 and the discrete distribution for ϕ learned from the data, which is represented by the Blackwell-MacQueen urn scheme $\frac{\alpha}{\alpha+386}G_0 + \frac{1}{\alpha+386}\sum_{i=1}^{386}\delta_{\phi_i}$. Note that the predicted ϕ_* does not depend on the predictor vector X_*; however, observe in Table 2 that cluster assignments of CoCs clearly depend on levels of the homeless rate, housing affordability, and extreme poverty. Utilizing the standard Blackwell-MacQueen urn scheme to predict the homeless rate in a new community results in unrealistic predictions, as it fails to adequately account for the inferred partition in predictor space and the local characteristics of the community. In other words, when predicting the homeless rate in a new community, it is reasonable to rely heavily on posterior draws from peer communities with similar characteristics. To generate more realistic and local predictions, we construct an approximate posterior predictive distribution using a localized variant of the Blackwell-MacQueen urn scheme: the predicted ϕ_* depends on X_*, which we denote $\phi_*(X_*)$. We fix a window around an element of X_* and utilize draws from the n_{X_*} CoCs with levels of housing affordability and extreme poverty, respectively, within the specified window. The index set for the CoCs local in predictor space is $\mathcal{I} = \{i : (\exists t)|X_{i,t} - X_*| < \epsilon\}$. To examine changes in homeless rates as a function of X_*, we compute the localized posterior predictive distribution

$$p(\phi_*(X_*)|C_{1:386,1:T}, N_{1:386,1:T}, X_*) = \int p(\phi_*(X_*)|\vartheta, X_*)p(\vartheta|C_{1:386,1:T}, N_{1:386,1:T})d\vartheta \qquad (14)$$

where $\vartheta = (\phi_1, \ldots, \phi_{386}, \alpha)$. We draw samples from this approximate posterior predictive distribution with a two step procedure.

1. For the m^{th} MCMC iteration, sample a new $\phi_*^{(m)}(X_*)$ from a modified Blackwell-MacQueen urn scheme that depends on X_*, $\frac{\alpha^{(m)}}{\alpha^{(m)}+n_{X_*}}G_0 + \frac{1}{\alpha^{(m)}+n_{X_*}}\sum_{j\in\mathcal{I}}\delta_{\phi_j^{(m)}}$, where $\mathcal{I} = \{i : (\exists t)|X_{i,t} - X_*| < \epsilon\}$, the index set for the n_{X_*} CoCs with predictor $X_{i,t}$ nearly equal to X_* for at least one t.
2. Construct $\psi_*^{(m)} = X_*'\phi_*^{(m)}(X_*)$ and transform to the homeless rate, $p_*^{(m)} = \frac{1}{1+e^{-\psi_*^{(m)}}}$.

While the conditional distribution of $\psi_*|Z_* = k$ is linear in predictor space, the marginal distribution of ψ_* may exhibit nonlinear associations as a function of CoC-predictor X_*. This flexible functional form allows us to to identify inflection points in the relationship between homeless rates (p_*) and features of a community (X_*), a main objective of the analysis.

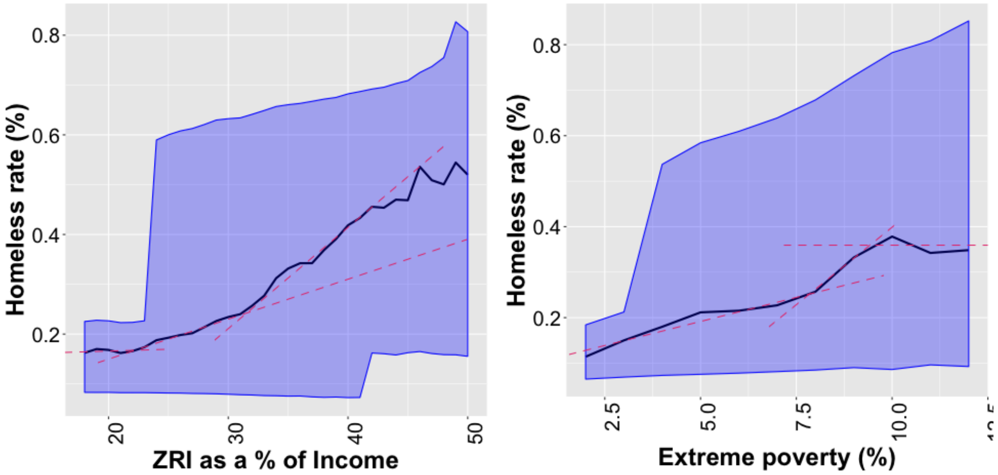
5 Results

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 non-linear associations between homeless rates, housing affordability, and extreme poverty. In Section 5.2, we present findings from our cluster analysis and discuss different types of homelessness. 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 break points in community features after which homeless rates are expected to rapidly increase. Identifying these inflection points can help communities prepare for rapid growth in homeless populations as key metrics of housing affordability and community-wide poverty cross a tipping point. In Figure 3, we summarize the relationship between homeless rates and community features with approximate posterior predictive distributions computed from the modified Blackwell-MacQueen urn scheme outlined in Section 4.2. The general strategy is to fix one community feature (affordability or poverty) to investigate the expected homeless rate as a function of the other. In Figure 3a, we predict the homeless rate as a function of housing affordability (x_*) for a new community with 6.64% of residents living in extreme poverty, the sample average. The predictor vector is $X_* = \begin{bmatrix} 1 & x_* & 6.64 \end{bmatrix}'$. For example, we expect a homeless rate of $\approx 0.41\%$ (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. The 90% predictive interval for the homeless rate spans 0.07% on the low end to 0.68% on the high end when ZRI is 40% of median income. 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. An important feature of Figure 3a is the widening 90% predictive interval when ZRI as a percent of income exceeds 40%. Since there are relatively few CoCs with extreme housing costs, the posterior predictive is informed by less data and the uncertainty interval widens.

Observe that when ZRI as a percent of median income is between 18-32%, the rate of increase in the expected homeless rate is not nearly as sharp as the rate of increase after 32%. In fact, the expected homeless rate is approximately piecewise linear, which is illustrated by the three dashed lines superimposed on the graph: the first line is flat over the range 18-22%; the second line increases from 22-32%; and the third line, beginning at 32%, has the steepest slope of all. The cluster assignments of the Dirichlet process model allow for changes in the structural relationship between housing affordability and homelessness, and the breakpoint in the expected homeless rate when ZRI reaches 32% of median income is learned from the data. The estimated 32% threshold is roughly consistent with the widely debated 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. We construct the predictive distribution in Figure 3a until ZRI reaches 50% of median income. We truncate at 50% because only 9 CoC's have higher relative housing costs, a number we feel is inadequate for robust estimation of the predictive distribution. In order to borrow information locally in X_*, we choose $\epsilon = 3\%$ (Step 1 in Section 4.1), which provides a rolling window of the communities included in the computation and results in local smoothing of the expected homeless rate in Figure 3a.



(a) Housing affordability

(b) Poverty rate

Figure 3: Left: The approximate posterior predictive distribution for homeless rates as ZRI/median income increases. Right: the approximate posterior predictive distribution for homeless rates as rates of extreme poverty increase. The shaded intervals illustrate the 90% predictive uncertainty intervals.

In Figure 3b, we predict the 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 0.24% (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 0.084% to 0.67%. In Albuquerque, NM (7.75% in extreme poverty, ZRI is 28.7% of median income) we estimate that in 2017 the homeless rate was 0.32% – again within the predicted range. Observe that the predictive interval also widens in Figure 3b as extreme poverty increases since there are few CoCs with very high extreme poverty rates. We note two separate breakpoints in the expected homeless rate at 8% and 10% extreme poverty in Figure 3b. When the extreme poverty rate exceeds 8%, the rate of increase sharpens. At 10%, the expected homeless rate reaches a plateau. Although the expected homeless rate flattens after 10%, the upper edge of the predictive interval continues to increase.

5.2 Clusters of CoCs

In our Dirichlet process mixture model of homeless rates, the number of clusters is learned from the data. In every iteration of our MCMC algorithm, both the number of clusters and the cluster membership of each CoC are sampled. Label switching among clusters and the varying dimension of the parameter space make direct inference on any one cluster difficult. For these reasons, we summarize inference on the relationship between community features and homeless rates with approximate posterior predictive distributions as in Section 5.1; however, there is significant interest from a policy perspective in identifying a group of peer 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 partition the set of 386 CoCs.

We find six different clusters; however, most CoCs (381 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 (0.08%), the most affordable housing (27.04%) and the lowest rate of extreme poverty (5.98%). Of clusters one through three, cluster three has, on average, the highest homeless rate (0.60%), the least affordable housing (38.44%), and the highest rate of extreme poverty (7.47%). The largest cluster – both by number of CoCs and by population – is cluster two, which is home to 47% of the U.S. population. While only 15.1% of the total U.S. population lives in cluster three, it contains 47.3% of the homeless included in the 2017 PIT counts.

Cluster	1	2	3	4	5	6
Size (# CoCs)	138	189	54	1	3	1
Share of Total Pop (%)	36.60	47.60	15.10	0.10	0.60	0.10
Share of PIT Count (%)	14.00	38.20	47.30	0.10	0.10	0.20
Homeless Rate (%)	0.08	0.19	0.60	0.42	0.03	0.53
Relative ZRI (%)	-8.91	-2.32	28.54	32.81	-29.49	183.37
Affordability Rate (%)	27.04	29.49	38.44	30.94	25.78	47.11
Poverty Rate (%)	5.98	6.80	7.47	3.96	7.96	3.26

Table 2: Cluster characteristics: 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. Relative ZRI (%) is the 2017 mean ZRI in the cluster as a percentage above (below) the national average. 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 southeast, where the ZRI is 8.91% below the national average. Most of New England, Florida, the mountain west and central United States are assigned to cluster two, where housing costs are on par with the national average – only falling 2.32% below the national average in ZRI. 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 38% of median income on average, are well above the break 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.

Clusters four through six correspond to CoCs that are relatively unique. The sole member of cluster four is El Dorado County CoC, which is unique because it has a high homeless rate but modest housing costs and low poverty rates (see Table 2). Cluster five has three mem-

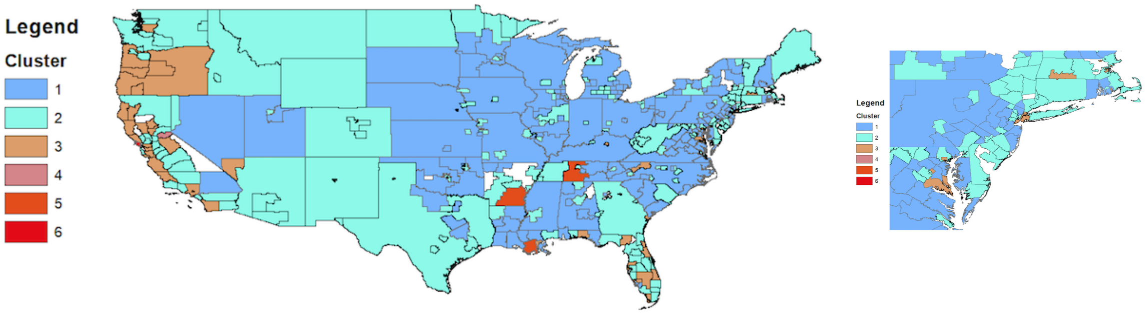


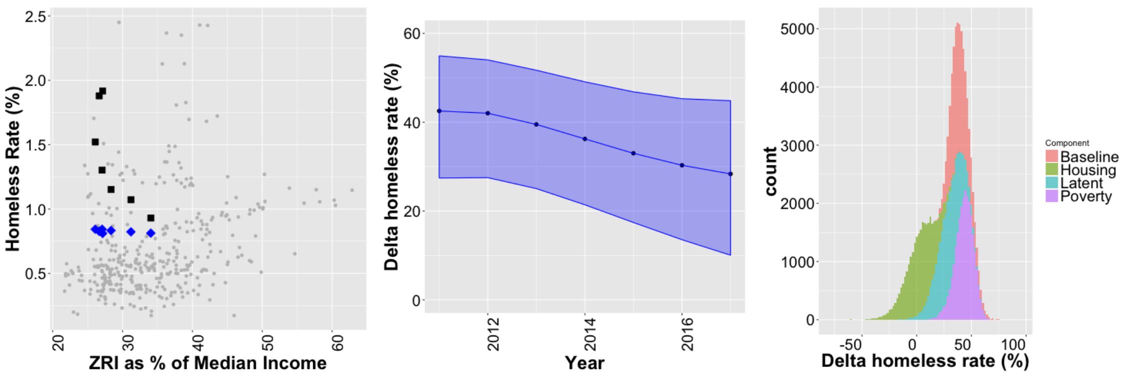
Figure 4: Map of clusters in the continental United States (left) and the northeast corridor (right) from Washington, D.C. to Boston, MA. Clusters exhibit strong spatial structure.

bers in the rural south – the Southeast Arkansas, Houma-Terrebonne/Thibodaux (Louisiana), and Central Tennessee CoCs (see Figure 4). In these communities, the average homeless rate is very low (0.03%) considering the high rate of extreme poverty (7.96%). The sole member of cluster six is the Marin County CoC in the San Francisco Bay area, which stands out for its particularly strong association between the homeless rate and worsening housing affordability. Cluster assignments for the 386 CoCs included in this analysis may be downloaded from https://github.com/G-Lynn/Inflection/.

5.3 CoC-level latent factors

There are many dimensions of a community. Poverty and housing affordability, while important features 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 $\beta_{i,t}|C_{1:386,1:T}, N_{1:386,1:T}$, as the deviation of the homeless rate in CoC i from the rate expected of CoCs with similar features 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 modest rents (approximately 30% of median income). Relative to peer CoCs in cluster three with similar housing costs, the homelessness rate in Atlanta is 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 $\beta_{i,t}$. Since latent factors in Atlanta are adding to the homeless rate beyond the rate expected of peers in cluster three with similar features, the posterior distribution for $\beta_{i,T}|C_{1:386,1:T}, N_{1:386,1:T}$ concentrates on positive values (Figure 5b). We interpret Figure 5b as the percent increase in the predicted homeless rate from a model that includes $\beta_{i,t}$ compared to the predicted rate when $\beta_{i,t} = 0$,



(a) Model Fit (b) Latent factors (c) Components

Figure 5: Atlanta Continuum of Care (GA-500). Left: Estimated homeless rate (squares) in Atlanta, the model fit for the homeless rate excluding latent factors (diamonds), and the homeless rates of other CoCs in cluster three (circles). Middle: Contribution of latent factors in Atlanta to homeless rate from 2011-2017. Right: Components of the 2017 homeless rate.

expressed mathematically as $100 \times \left(1 - \frac{1+\exp\{-\beta_{i,t} - X'_{i,t}\phi_i\}}{1+\exp\{-X'_{i,t}\phi_i\}}\right)$. 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 $\beta_{i,t}|C_{1:386,1:T}, N_{1:386,1:T}$. The latent factor distribution over time provides a mechanism to evaluate the CoC's changing environment for homelessness – including policy interventions.

In Figure 5c, we examine the contribution of each element in $X_{i,t}$ to the predicted homeless rate with a similar strategy. Denote the j^{th} element of vectors $X_{i,t}$ and ϕ_i as $X^{(j)}_{i,t}$ and $\phi^{(j)}_i$. We define the baseline (chronic) homeless rate in a CoC as $\frac{1}{1+\exp\{-\phi^{(1)}_i\}}$, the rate predicted by the shared cluster intercept alone. The baseline rate presented in Figure 5c is the percentage of the predicted homeless rate corresponding to the cluster intercept alone, $100 \times \left(\frac{1+\exp\{-\beta_{i,t} - X'_{i,t}\phi_i\}}{1+\exp\{-\phi^{(1)}_i\}}\right)$. In Figure 5c, observe that the expected baseline homeless rate associated with the cluster intercept is 39% of Atlanta's predicted homeless rate in 2017. The contribution of housing affordability (extreme poverty) is quantified as the percent change in the predicted homeless rate of the full model compared to a model that excludes housing affordability (extreme poverty). The percent change in the predicted homeless rate for predictor j is then $100 \times \left(1 - \frac{1+\exp\{-\beta_{i,t} - X'_{i,t}\phi_i\}}{1+\exp\{-\beta_{i,t} - X'_{i,t}\phi_i + X^{(j)}_{i,t}\phi^{(j)}_i\}}\right)$. In Atlanta in 2017, adding housing affordability to the model only increases the predicted homeless rate by an expected 2.5%. On the other hand, adding the rate of extreme poverty to the model increases the predicted homeless rate by an expected 43%. Including latent factors increases the predicted homeless rate by an expected 28.5%. While these contributions do not sum to 100%, they do indicate the magnitude of the relative contribution associated with each factor.

The posterior distributions for each component presented in Figure 5c provide a tool for HUD and individual CoCs to investigate the largest factors related to their homeless rate. For each CoC, it is possible to construct a version of Figure 5b to (i) establish a baseline homeless rate and (ii) examine the magnitude of increases associated with each of housing affordability, poverty, and latent factors. It is possible to focus policy interventions on mitigating the factors most pertinent to an individual CoC.

6 Discussion

In this paper, we present a Bayesian nonparametric model of community-level homeless rates. The Dirichlet process model shares information across CoCs where homeless rates are similarly related to features of a community, and we utilize an approximate posterior predictive distribution 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. By identifying inflection points in the relationship between homelessness and both housing affordability (as measured by the rent/income ratio) and rate of extreme poverty, we show that these relationships follow a unique 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, but with significantly large populations (cluster 3), are experiencing surges in homelessness related to very high housing costs and extreme poverty. The average housing affordability metric is higher in cluster three (38.44%) 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 27% and 29.5%, respectively – and the majority of the United States is less sensitive to increases in housing costs than those 54 communities in cluster 3. This may explain why, despite increased homelessness in cluster 3 cities like Los Angeles, New York, and Seattle, the nation has been measuring a steady net decline in homelessness since the recession of 2008.

The motivation for prior research on community-level determinants of homelessness has been that factors identified as key drivers of higher (or lower) rates of homelessness can 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 features. 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

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric bayesian model for randomised block designs. Biometrika, 83(2):275–285.

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.

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.

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. (2018). Dynamics of homelessness in urban America. Annals of Applied Statistics, forthcoming.

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.

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

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.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. Statistica sinica, pages 639–650.

West, M., Muller, P., and Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. Aspects of Uncertainty: A Tribute to DV Lindley, pages 363–386.