Submitted to the Annals of Applied Statistics

INFLECTION POINTS IN COMMUNITY-LEVEL HOMELESS RATES

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

Zillow Research [*], *Boston Univ.* [†], *and Univ. of Pennsylvania* [‡]

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 (EPA) distribution to develop a Bayesian nonparametric regression 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 begins to quickly increase once median rental costs exceed 30% of median income, providing a statistical link between homelessness and the U.S. government's definition of a housing cost burden. Our analysis also identifies clusters of communities that exhibit distinct geographic patterns and yields 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). In these rate estimates, HUD uses a naive calculation, dividing the raw counted number of homeless by a community's total population. Due to systematic inaccuracies in homeless count data, this is methodologically problematic and naive estimates should be used with caution (Glynn and Fox, 2019; Hopper et al., 2008); nevertheless, the more than seventeen-fold increase in the HUD-reported rate of homelessness within the state of California suggests that homelessness is critically influenced by features of individual communities.

In this study, we investigate potentially nonlinear relationships between homeless rates and community-level predictors. Quantifying the association

Keywords and phrases: homelessness, housing affordability, Bayesian hierarchical model, random partition model, Ewens-Pitman attraction

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 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 ignores 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 overly 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; however, inferring covariate-dependent clusters from scarce data is a challenging statistical learning problem.

We have two primary objectives in our analysis:

(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 peer groups of communities for development and evaluation of policy interventions.

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

[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)

limited region of predictor space. Estimating the complete response surface 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.

The methodological contribution of the study is EPA regression, a novel Bayesian nonparametric regression framework based on the Ewens-Pitman attraction distribution (Dahl et al., 2017). EPA regression is designed to estimate nonlinear response surfaces when observational units generate data that is both scarce and concentrated in one region of predictor space. EPA regression offers important modeling advantages compared to existing Bayesian nonparametric regression alternatives such as BART (Chipman et al., 2010), Gaussian process regression (Williams and Rasmussen, 1996), and Dirichlet process mixture models (Antoniak, 1974; Escobar and West, 1995). Further, it allows us to achieve both objectives (O_1) and (O_2). EPA regression explicitly models the partition of communities, and cluster probabilities depend on covariates. Unlike BART, which partitions predictor space into rectangular boxes, EPA regression offers a more flexible partition of predictor space. Gaussian process regression does not explicitly model cluster structure, and Dirichlet process mixtures, which do model cluster structure, do not model cluster dependence on covariates.

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.

We identify a period or structural change in homeless rates when housing costs in a community are between 30-34% of median income, with the most likely inflection point occuring at 32%. Once housing costs enter the 30-34% of median income region, the expected homeless rate in a community increases sharply. We also find three dominant modes of variation in homeless rates, with 373 of 381 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 regression model for homeless populations and describe choices for prior distributions; 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 discuss the applied contribution of our findings and related policy considerations.

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 con-

[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.

struct 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.

	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	2,354.98	62,552.42	8.46

TABLE 1

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

3. A Bayesian nonparametric model for homeless counts.

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 latent variables 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 framework here.

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

$$N_{i,t}|\lambda_{i,t} \sim Poisson(\lambda_{i,t}). \tag{3.1}$$

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,

$$H_{i,t}|N_{i,t}, p_{i,t} \sim Binomial(N_{i,t}, p_{i,t}). \tag{3.2}$$

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

$$C_{i,t}|H_{i,t}, \pi_{i,t} \sim Binomial(H_{i,t}, \pi_{i,t}). \tag{3.3}$$

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. The methodological novelty is presented in Section 3.2.

3.2. *EPA regression 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 3.2, 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

$$\psi_{i,t} = log\left(\frac{p_{i,t}}{1-p_{i,t}}\right) = F_i'\beta_{i,t} + X_{i,t}'\phi_i + \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim N(0, \sigma^2_{\psi_i}). \tag{3.4}$$

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.

EPA 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

$$\phi_i = \sum_{k=1}^{q_n} \tilde{\phi}_k \mathbb{1}_{\{Z_i = k\}}, \tag{3.5}$$

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 $X_{i,t}$ (e.g., $X_{i,t} = \begin{bmatrix}1 & ZRI_{i,t}/MedianIncome_{i,t} & ExtPoverty_{i,t}\end{bmatrix}'$). The leading one results in a shared cluster-level intercept or expected rate of homelessness.

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 $f : \mathbb{R}^3 \to (0, 1]$ maps distance between CoCs in covariate space to the unit interval, quantifying the pairwise similarity between two CoCs,

$$f(X_{\omega_j,T}, X_{\omega_i,T}) = \exp\{-\tau ||X_{\omega_j,T} - X_{\omega_i,T}||_2\}. \tag{3.6}$$

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. We find that our results are robust to the choice of τ.

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

$$p(\boldsymbol{\pi}_n | \alpha, \delta, f, \boldsymbol{\omega}) = \prod_{\ell=1}^{n} p_\ell(\alpha, \delta, f, \boldsymbol{\pi}(\omega_1, \ldots, \omega_{\ell-1}), \tag{3.7}$$

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.

$$p_\ell(\alpha, \delta, f, \boldsymbol{\pi}(\omega_1, \ldots, \omega_{\ell-1})) = Pr(Z_{\omega_\ell} = k | \alpha, \delta, f, \boldsymbol{\pi}(\omega_1, \ldots, \omega_{\ell-1})) \tag{3.8}$$

$$= \begin{cases} \left(\frac{\ell-1-\delta q_{\ell-1}}{\alpha+\ell-1}\right) \frac{\sum_{\{\omega_s : Z_{\omega_s} = k\}} f(X_{\omega_\ell,T}, X_{\omega_s,T})}{\sum_{s=1}^{\ell-1} f(X_{\omega_\ell,T}, X_{\omega_s,T})}, & \text{for } k = 1, \ldots, q_{\ell-1} \\ \\ \frac{\alpha+\delta q_{\ell-1}}{\alpha+\ell-1} \text{ for } k = 0 \text{ (e.g., a new cluster)} \end{cases} \tag{3.9}$$

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 nonexchangeability issue by utilizing a prior distribution for permutations and numerically integrating over all possible permutations in our MCMC algorithm (see MCMC details in the Supplement (Glynn et al., 2020)), 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 $(\phi_1, \dots, \phi_n)$ 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 3.4. 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 $n = 381$ CoCs. As the innovation variance $\sigma^2_{\psi_i}$ increases, larger deviations in homeless rates from the cluster-level 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.

$$\sigma^2_{\psi_i} \sim IG(a_\psi, b_\psi) \tag{3.10}$$

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 EPA 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

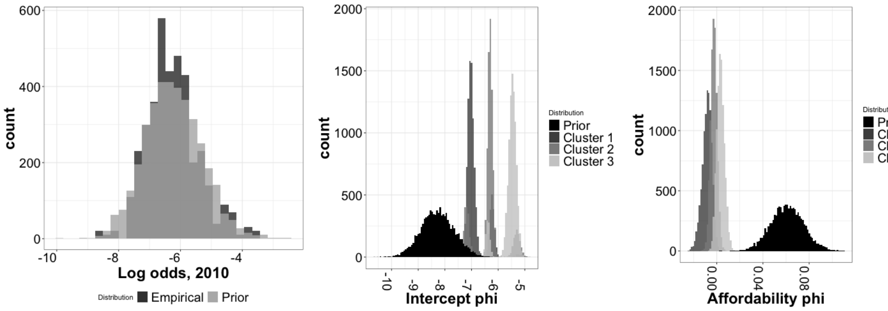
$$\beta_{i,t} = A\beta_{i,t-1} + w_{i,t}, \qquad w_{i,t} \sim N(0, W). \tag{3.11}$$

The dynamic latent factor model in 3.11 makes two important contributions: first, the 2×1 $\beta_{i,t}$ vector provides a mechanism to include (in aggregate) the 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.

We only include the housing affordability and extreme poverty covariates in the EPA regression because previous research has found that related covariate-dependent clustering methods, such as the model of Müller et al. (2011), work best with only a few covariates (Page and Quintana, 2018). Rather than controlling for additional covariates in the EPA regression, we have grouped the aggregate contribution of excluded variables into a single latent factor. This strategy allows us to use only a few covariates, as recommended by Page and Quintana (2018), and still estimate the CoC-specific latent factor, serving as an aggregate, albeit difficult to interpret, latent control variable.

3.3. *Prior choices.* Prior distributions for $(\beta_{i,0}, \alpha, \sigma^2_{\psi_i})$ and hyperparameters μ_0 and Σ_0 in the base measure for ϕ_i 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 1a, which also presents the marginal prior for $\psi_{i,0}$). The expectation of $\psi_{i,0}$, computed by taking the expectation of 3.4, is $E[\psi_{i,0}] = F_{i,0}'E[\beta_{i,0}] + X_{i,0}'E[\phi_i]$. We choose $E[\beta_{i,0}] = \begin{bmatrix} 0 & 0 \end{bmatrix}'$ 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. 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 $\mu_0' = \begin{bmatrix} -8.28 & 0.061 & 0.061 \end{bmatrix}$.



(a) Moment Matching (b) Intercept (c) Affordability

Fig 1: Left: The empirical distribution of log odds of homelessness in 2010 and the implied prior distribution for $\psi_{i,0}$. Middle: the prior and posterior distributions for $\tilde{\phi}_k^{(1)}$, the parameter associated with cluster intercepts. Right: the prior and posterior distributions for $\tilde{\phi}_k^{(2)}$, the parameter associated with cluster housing affordability.

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,

$$Var(\psi_{i,0}) = E[Var(\psi_{i,0}|\beta_{i,0}, \phi_i, \sigma_{\psi_i}^2)] + Var(E[\psi_{i,0}|\beta_{i,0}, \phi_i, \sigma_{\psi_i}^2]) \tag{3.12}$$

$$= E[\sigma_{\psi_i}^2] + F_{i,0}' Var(\beta_{i,0}) F_{i,0} + X_{i,0}' Var(\phi_i) X_{i,0}. \tag{3.13}$$

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 0.4 for the intercept variance is quite diffuse, and, as seen in Figure 1b, the posterior distributions for the intercept in clusters one, two, and three are concentrated on values in the far right tail of the prior. 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 1c, where the posteriors for the housing affordability coefficients in clusters one, two, and three concentrate on values closer to zero. The important takeaway is that the prior distribution is sufficiently diffuse to allow the data a major contribution to the posterior, with clear Bayesian learning. While Figures 1b and 1c present posterior distributions for individual coefficients, we caution readers from interpreting these coefficients too deeply, as there are complicated interactions between coefficients due to the EPA clustering strategy. Instead, we focus on posterior predictive distributions presented in Section 5. 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.

4. Markov chain Monte Carlo.

Our objective is to sample from the posterior distribution

$$p(\tilde{\boldsymbol{\phi}}, Z_{1:n}, \beta_{1:n,1:T} | N_{1:n,1:T}, C_{1:n,1:T}). \tag{4.1}$$

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. To guarantee reproducible inference, we ran multiple MCMC chains initialized at different parameter values. Posterior distributions and functionals from each chain generated identical inferences. Trace plots and effective sample size calculations for individual parameters give us additional confidence in our posterior estimates. The MCMC simulation is developed in R (R Core Team, 2017), and it took approximately 24 hours to run on a MacBook Pro.

A major focus of our analysis is computing the posterior predictive distribution for the homeless rate in a new community,

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

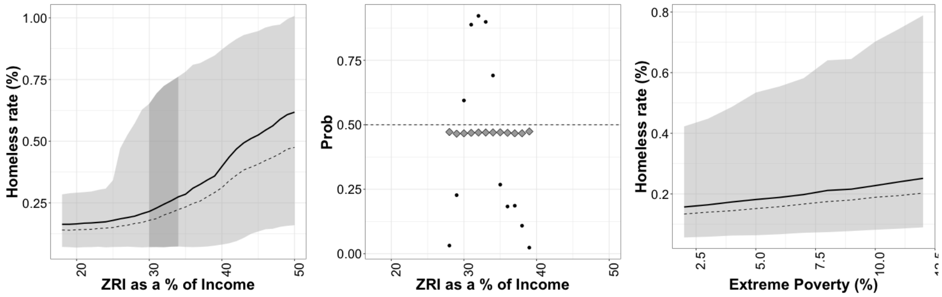
Section 1 of the Supplement (Glynn et al., 2020) presents MCMC sampling steps and constructs the posterior predictive from MCMC samples.

5. Results.

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 2, we summarize the relationship between homeless rates, housing affordability, and extreme poverty with the posterior predictive distribution computed in the Supplement (Glynn et al., 2020).

In Figure 2a, we predict the homeless rate as a function of housing affordability when extreme poverty is 7%, approximately the sample average.

For example, we expect a homeless rate of ≈ 0.40% (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 6.25% and ZRI consumed 39.28% of median income. The estimated homeless rate in San Diego in 2017 was 0.36% – right in the middle of the predicted range.



(a) Housing affordability (b) Inflection Prob. (c) Poverty rate

Fig 2: Cross-sections from the homeless rate surface. Left: Posterior predictive distribution for homeless rates as ZRI/median income increases when extreme poverty is 7%. The shaded intervals illustrate the 90% predictive uncertainty intervals, and the dashed line represents the benchmark case in which the naive homeless rate ($p_{i,t}^{Naive} = C_{i,t}/N_{i,t}$) is used for the analysis. The dark shaded region denotes the 30-34% region where inflection points are likely. Middle: Posterior probability of inflection points in the expected homeless rate as a function of housing affordability. The shaded diamonds present the prior probability of an inflection point, reflecting our prior belief that each point has a roughly equal probability of being an inflection point. Right: Posterior predictive distribution for the homeless rate as a function of extreme poverty when housing affordability is 28%. The shaded intervals illustrate the 90% predictive uncertainty intervals, and the dashed line represents the benchmark case in which the naive homeless rate ($p_{i,t}^{Naive} = C_{i,t}/N_{i,t}$) is used for the analysis.

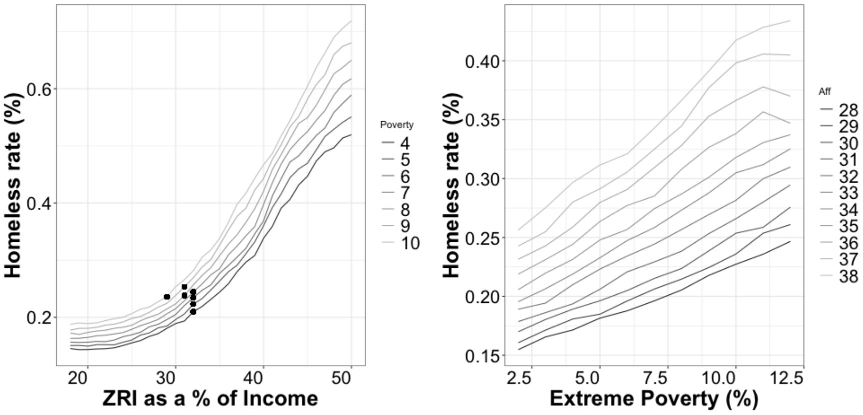
To identify inflection points in housing affordability, we numerically evaluate the second derivative of $\tilde{p}(x_*) := E[p_{n+1,T}|C_{1:n,1:T}, N_{1:n,1:T}]$ for covariate vector $X_*' = \begin{bmatrix}1 & x_* & 7\end{bmatrix}$. To estimate the location of inflection points, we compute the posterior probability that the second derivative $\tilde{p}''(x_*)$ exceeds threshold κ, corresponding to structural changes in the slope of $\tilde{p}(x_*)$. This probability is computed with posterior samples in equation 5.1.

$$P\left(\tilde{p}''(x_*) > \kappa\right) \approx \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}_{\left\{\frac{\tilde{p}^{(m)}(x_*+h)-2\tilde{p}^{(m)}(x_*)+\tilde{p}^{(m)}(x_*-h)}{h^2}>\kappa\right\}} \tag{5.1}$$

We classify x_* as an inflection point if the probability that $(\tilde{p}''(x_*) > \kappa)$ exceeds 0.5. When we apply this procedure with $\kappa \geq 0.00125$ and $h = 10$, a series of likely inflection points emerge from 30-34%, with 32% being the most likely (see Figure 2b). In Figure 2a, we have marked the 30-34% region with a dark shaded uncertainty interval. Observe that when ZRI as a percent of median income is between 18-30%, the rate of increase in the expected homeless rate is not nearly as sharp as the rate of increase from 34 - 50%. A clear structural change occurs in this 30-34% range, which is particularly noteworthy because the U.S. government defines a housing cost burden when a family spends more than 30% of its household income on housing costs (HUD, 2018). When families become acutely cost burdened, we find that the expected homeless rate sharply increases.

In Figure 2c, 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 2c 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 more than 0.6%. In Albuquerque, NM (8.1% in extreme poverty, 28.1% for ZRI/median income) we estimate that in 2017 the homeless rate was 0.32% – again within the predicted range. Observe in Figure 2c 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.

To benchmark the homeless rate model in Section 3 against the standard method for calculating homeless rates, we computed the posterior predictive distribution in Figures 2a and 2c using the naive homeless rate, $p_{i,t}^{Naive} = C_{i,t}/N_{i,t}$. The naive homeless rate is the current standard in both academic research (see e.g. Byrne et al. (2013)) and policy analysis, and it is currently used by HUD in its Annual Homeless Assessment Report (AHAR) to Congress. In equation 3.4, we let $\psi_{i,t}^{Naive} = \log\left(\frac{p_{i,t}^{Naive}}{1-p_{i,t}^{Naive}}\right)$. Our comparison analysis ignores the underlying count data and two-stage binomial thinning model in equations 3.2 and 3.3 and assumes that the homeless rates are directly observed. The dashed lines in Figures 2a and 2c illustrate the naive homeless rate as a function of both housing affordability and extreme poverty. Observe that the functional form in the naive estimates is nearly identical to the homeless rates estimated in our full model, with the difference being the naive homeless rates are less than the implied rates from our model. This is expected since our full model accounts for the imperfect accuracy in HUD's point-in-time count data, following the method of Glynn and Fox (2019). An important takeaway from Figure 2a is that the posterior predictive of the naive homeless rate depends only on modeling advances in the EPA regression and is not impacted by any prior choices on count accuracy.



(a) Rate v. Aff

(b) Rate v. Pov

Fig 3: The expected homeless rate from the posterior predictive distribution, $E[p_{n+1,T}|C_{1:n,1:T}, N_{1:n,1:T}]$. Left: Homeless rate v. Affordability, stratified by poverty. The black points represent the most likely inflection point for each level of extreme poverty. Right: Homeless rate v. extreme poverty, stratified by housing affordability.

To compare the Glynn and Fox (2019) model with the current EPA model, we utilized the Widely Applicable Information Criterion (Vehtari et al., 2017). To make the fairest comparison between the log-odds regression model in Glynn and Fox (2019) and the EPA regression model, we ignored the problem of count accuracy and compared the model fit strictly using the naive homeless rate calculation as the response. The WAIC from the present EPA regression (-1015.36) is significantly lower than the WAIC from the Glynn and Fox model (215.42), indicating a superior model fit from the EPA regression.

In Figure 3a, we present the expected homeless rate as a function of housing affordability, stratified by extreme poverty. The most likely inflection points at different levels of extreme poverty are marked by black points. As the level of extreme poverty increases from 4% to 10%, the inflection point in housing affordability slightly decreases from 32% to 29%. From an applied perspective, this makes sense. As a larger share of a CoC contends with extreme poverty, the homeless rate inflection point occurs at a lower housing cost. Figure 3b illustrates the expected homeless rate as a function of extreme poverty, stratified by housing affordability. Observe that the nearly linear relationship between homeless rates and extreme poverty holds across a wide range of housing affordability values, and the slope steepens as housing affordability increases from 28-38%. Interactions between covariates and localized Bayesian learning are illustrated in the full homeless rate surface, presented in Figure 2 in the Supplement (Glynn et al., 2020).

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 cooccurences 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 Hubert and Arabie (1985) to compute a point estimate of the partition $\hat{\boldsymbol{\pi}}_{381}$ following the approach of Fritsch and Ickstadt (2009).

We find six different clusters; however, most CoCs (373 of 381) 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 (25.51%) and the lowest rate of extreme poverty (5.91%). Of clusters one through three, cluster three has (on average) the highest homeless rate (0.65%), the least affordable housing (37.41%), and the highest rate of extreme poverty (7.30%). The largest cluster – both by number of CoCs and by population – is cluster two, which is home to 50.16% of the U.S. population. While only 13.85% of the total U.S. population lives in cluster three, it contains 45.76% of the homeless included in the 2017 PIT counts.

Although the model contains no specific mechanism for spatial patterns in homeless rates, there is clear spatial structure in our cluster assignments. Observe in Figure 4 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

Cluster	1	2	3	4	5	6
Size (# of CoCs)	135	190	48	6	1	1
Share of Total Pop (%)	34.43	50.16	13.85	1.10	0.32	0.14
Share of PIT Count (%)	13.46	40.20	45.76	0.19	0.16	0.23
Homeless Rate (%)	0.09	0.19	0.65	0.04	0.09	0.36
Affordability Rate (%)	25.51	28.41	37.41	26.78	22.71	32.17
Poverty Rate (%)	5.91	6.76	7.30	7.11	4.38	5.22

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.

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 37.4% of median income on average, are well above the inflection point range of 30-34% 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, which contain a total of eight CoCs, may not be robust to the addition (or removal) of CoCs to the data set. For this reason, we focus our interpretation on clusters one through three and do not draw conclusions from clusters four through six.

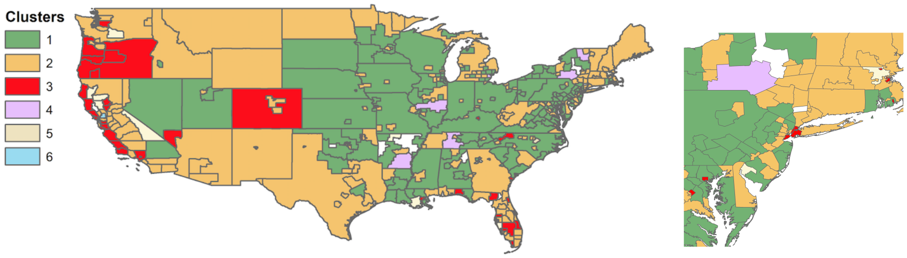
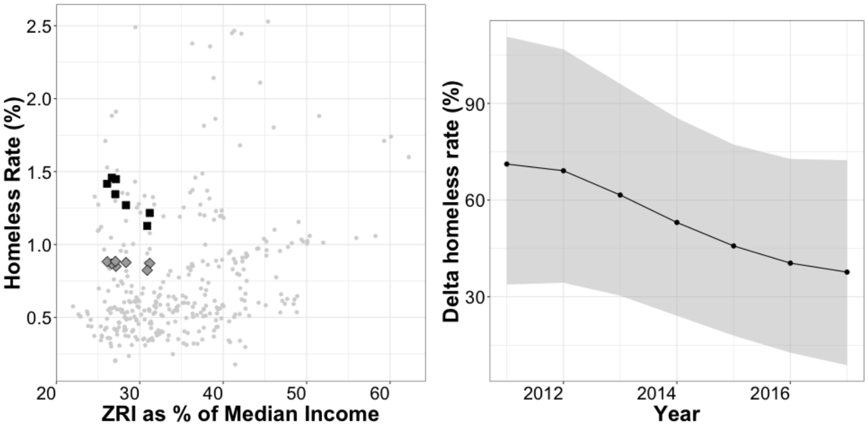


Fig 4: Map of clusters in the continental United States (left) and the north-east corridor (right) from Washington, D.C. to Boston, MA. Clusters exhibit strong spatial structure.

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. Recall from Section 3.2 that $F_i' = \begin{bmatrix}1 & 0\end{bmatrix}$.

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 (0.94%) for a CoC with housing costs at 31% 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 11.5% 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



(a) Model Fit

(b) Latent factors

Fig 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.

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 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 31% 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.

6. Discussion.

In this paper, we present a Bayesian nonparametric model of community-level homeless rates. The EPA regression 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 exhibits a structural change when ZRI as a percentage of median income is in the 30-34% range, a finding that statistically connects community-level homelessness and 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 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. 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 (37.41%) than the 30-34% region we identify, which partly explains rapid growth in the homeless populations of many of these CoCs.

Despite statistical associations between housing affordability, extreme poverty, and homelessness, the present analysis does not permit causal inference. As a result, the implications of our results for crafting public policy may be limited. Significantly modifying policy without first establishing a strong causal link between housing affordability, extreme poverty, and homelessness would be a mistake. Causal inference on the relationship between community-level characteristics and homelessness is an important direction for future research. While our results may not be directly used to determine policy, the inflection point analysis is important to improve homeless population monitoring and forecasting systems.

Our findings statistically link homelessness and the HUD guideline of using a rent-to-income ratio of 30% as a standard measure of affordability. This 30% rule–which has its origins in the so-called "Brooke Amendment" to the 1968 Fair Housing Act–has been criticized as being essentially arbitrary in nature and for its failure to account for other factors that affect affordability, such as family size and composition and geographic variation in the cost of other goods and services besides housing (U.S. Department of Housing and Urban Development, 2014). Despite these criticisms, our results suggest broad increases in housing vulnerability in communities where the rent-to-income ratio exceeds 30%. While we are unable to assert causality, and it is possible that the 30-34% finding reflects rather than validates the current HUD guideline, statistical evidence for wide-spread housing vulnerability as renters become increasingly cost burdened is meaningful for improved monitoring and forecasting. Our results serve as an important first step in developing improved early warning systems for monitoring housing market conditions and forecasting their impact on the demand for homeless services.

In addition, identification of distinct clusters of communities suggests there is potential value in implementing strategically differentiated policy interventions based on community characteristics, which would be a departure from current approaches. In prior research on the relationship between community characteristics and rates of homelessness, factors identified as key drivers of higher (or lower) rates of homelessness may have been subsequently 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 points to the potential need for multiple policy responses that target the needs and structural determinants of homelessness in 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.

While the development of EPA regression is motivated by modeling homeless rates, we believe it may be a useful method in other application areas, particularly those challenged by data that is sparse in both time and space. For example, crime, real estate transactions, and onset of rare disease occur infrequently, and data is scarce at the community level. To estimate the response surface, it is often necessary to pool data across communities with starkly different characteristics. EPA regression offers a principled statistical framework for sharing information across related communities in many applications, particularly those challenged by data that is sparse in both time and space.

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.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. Ann. Appl. Stat., 4(1):266–298.

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., 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.

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

Glynn, C., Byrne, T. H., and Culhane, D. P. (2020). Supplement to Inflection Points in Community-Level Homeless Rates. Annals of Applied Statistics.

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.

Hubert, L. and Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1):193–218.

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. (2018). Calibrating covariate informed product partition models. Statistics and Computing, 28(5):1009–1031.

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.

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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.

U.S. Department of Housing and Urban Development (2014). Rental Burdens: Rethinking Affordability Measures.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. Statistics and computing, 27(5):1413–1432.

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

Williams, C. K. and Rasmussen, C. E. (1996). Gaussian processes for regression. In Advances in neural information processing systems, pages 514–520.

Chris Glynn
Zillow Research
1301 2nd Ave.
Seattle, WA 98101 E-mail: christophergl@zillowgroup.com

Thomas H. Byrne
School of Social Work
Boston University
Boston, MA 02215 E-mail: tbyrne@bu.edu

Dennis P. Culhane
School of Social Policy & Practice
University of Pennsylvania
Philadelphia, PA 19104 E-mail: culhane@upenn.edu