Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

DYNAMICS OF HOMELESSNESS IN URBAN AMERICA

By Chris Glynn[†] and Emily B. Fox[‡]

University of New Hampshire [†] *and University of Washington* [‡]

The relationship between housing costs and homelessness has important implications for the way that city and county governments respond to increasing homeless populations. Though many analyses in the public policy literature have examined inter-community variation in homelessness *rates* to identify causal mechanisms of homelessness (Byrne et al., 2013; Lee et al., 2003; Fargo et al., 2013), few studies have examined time-varying homeless *counts* within the same community (McCandless et al., 2016). To examine trends in homeless population counts in the 25 largest U.S. metropolitan areas, we develop a dynamic Bayesian hierarchical model for time-varying homeless count data. Particular care is given to modeling uncertainty in the homeless count generating and measurement processes, and a critical distinction is made between the counted number of homeless and the true size of the homeless population. For each metro under study, we investigate the relationship between increases in the Zillow Rent Index and increases in the homeless population. Sensitivity of inference to potential improvements in the accuracy of point-in-time counts is explored, and evidence is presented that the inferred increase in the rate of homelessness from 2011-2016 depends on prior beliefs about the accuracy of homeless counts. A main finding of the study is that the relationship between homelessness and rental costs is strongest in New York, Los Angeles, Washington, D.C., and Seattle.

1. Introduction. Counts of people experiencing homelessness in cities such as Seattle, Los Angeles, and New York reveal alarming year-over-year increases in the raw numbers of enumerated individuals. In addition to rising counts of homeless, rental costs in these cities are significantly increasing as well. The relationship between housing costs and homelessness is a topic of great public importance and has received considerable attention (Hanratty, 2017; Fargo et al., 2013; Byrne et al., 2013; Stojanovic et al., 1999; O'Flaherty, 1995; Sclar, 1990).

Several challenges exist in quantifying the impact of increased rental

*Part of this research was conducted while Glynn was a postdoctoral researcher at the University of Washington, where he was supported by a research gift from Zillow. The work was also supported by AFOSR Grant FA9550-16-1-0038, and NSF CAREER Award IIS-1350133. The authors thank Surya Tokdar, Svenja Gudell, Cory Hopkins, Melissa Allison, Thomas Byrne, and Dennis Culhane for helpful discussions.

Keywords and phrases: homelessness, housing costs, missing data, state-space

costs on the size of the homeless population. The first challenge is that point-in-time homeless counts often occur on a single night in January and are thus subject to significant sampling variability. The second challenge is that the accuracy of the count itself is not the same from one year to the next. Differences in the number of volunteers, weather, and count methodologies lead to counts that are difficult to compare year-over-year.

These facts beg the question: are homeless populations across the county increasing? Or do the reported counts simply represent a higher fraction of the homeless population? Changing count accuracy over time directly impacts inferred trends in the size of the homeless population. In light of this fact, we investigate the impact of different count accuracy trajectories on the inferred change in homelessness rates from 2011-2016.

Inference on the relationship between trends in rental costs and trends in the homeless population is related to other trend analyses with data quality challenges (Tokdar et al., 2011; Coles and Sparks, 2006; Cornulier et al., 2011; Kery and Royle, 2010). Although we observe the number of *counted* homeless, we do not observe the true size of the homeless population. Plant-capture methods (Laska and Meisner, 1993; Schwarz and Seber, 1999) have demonstrated that homeless counts systematically understate the size of the total homeless population (Hopper et al., 2008). One strategy to include the uncounted number of homeless in the analysis is to build a mechanism for the imperfect counting process into the statistical model, as McCandless et al. (2016) have done with plant-capture data from Edmonton, Canada.

In this paper, the total size of the homeless population is imputed, and uncertainty in the total homeless population and count accuracy is propagated to our assessment of the relationship between rental costs and homelessness. Our goal is to *jointly* model the collection of homeless count time series from the 25 largest metropolitan areas in the United States. In contrast to McCandless et al. (2016), who treat time-indexed counts as exchangeable, we directly model temporal dependence. We develop a Bayesian dynamic modeling framework to investigate the relationship between the number of homeless and rent costs subject to different prior beliefs about count accuracy over time.

The data in our analysis comes from three sources: the U.S. Census Bureau; the U.S. Department of Housing and Urban Development (HUD); and the housing website Zillow. The data include the total population, point-in-time homeless counts, and the Zillow Rent Index (ZRI) for continuums of care that service the 25 largest metro areas from 2011-2016. The continuum of care (CoC) is an administrative unit that seeks to integrate local homeless services, counts, and data under a single community organization (Office

of Community Planning and Development, Dept. of Housing and Urban Development, 2009).

Numerous previous studies have utilized inter-community variation in homelessness *rates* to identify potential causal mechanisms of homelessness (Fargo et al., 2013; Byrne et al., 2013; Raphael, 2010; Lee et al., 2003; Early and Olsen, 2002; Quigley and Raphael, 2001; Quigley et al., 2001a; Troutman et al., 1999; Hudson, 1998; Grimes and Chressanthis, 1997; Honig and Filer, 1993; Burt, 1992; Bohanon, 1991; Appelbaum et al., 1991; Quigley, 1990). Studies that model homelessness rates, defined as $\frac{\text{total homeless}}{\text{total population}}$, assume that both the numerator and denominator are observed without error. In practice, there is significant uncertainty in both the numerator and denominator in any such homelessness rate calculation. To account for that uncertainty, we directly model time-varying *counts* within the same community. Working with time series of count data has two advantages. First, statistical models of counts more aptly characterize the sampling variability in the observed data; and second, focusing on within community variation over time avoids drawing conclusions from data generated across different municipal and state governments, climates, and social structures. A major contribution of our work is the development of a statistical framework that enables researchers, policymakers, and local continuum coordinators to address five specific questions for each metro:

(Q1) When adjusting for increases in count accuracy and total population growth, is the *rate* of homelessness increasing?

(Q2) If ZRI increases by $x\%$, what are the predicted increases in the counted and total number of people experiencing homelessness?

(Q3) How can we quantify uncertainty in the reported point-in-time counts?

(Q4) Given that C homeless are counted and count accuracy is imperfect, what is the expected range in the total number of people experiencing homelessness at a point in time?

(Q5) What is the one-year-ahead forecast of the total homeless population in 2017?

We identify New York, Los Angeles, Seattle, and Washington, D.C. as metros where (i) the inferred rate of homelessness significantly increased from 2011-2016 and (ii) there exists a strong relationship between housing costs and homelessness. We present evidence that the inferred change in the homelessness rate from 2011-2016 is sensitive to the trajectory of count

accuracy. This point is emphasized to encourage researchers, policymakers, and continuum leaders to carefully quantify their beliefs and uncertainty about count accuracy.

The prior beliefs that we incorporate in this analysis are informed by existing literature and discussions with count coordinators, volunteers, and homelessness experts from around the country. Incorporating the expert opinions of count coordinators in every metro in the sample will lead to a more informed study. Our goal in this paper is to advance the statistical methodology utilized by researchers to analyze data on homelessness. We view this as a demonstration of a predictive modeling framework that will benefit from a partnership between private companies with relevant data, HUD, and local continuums of care.

Homelessness is a complex problem with many potential contributing factors. Though previous research has consistently identified the cost of rental housing as a significant predictor of homelessness (see, e.g., Corinth (2015); Byrne et al. (2013); Quigley et al. (2001b)), we recognize that apartment vacancy rates, social safety net measures, changes in affordable/supportive housing policy, and local unemployment rates could contribute to increased homelessness as well. There are many paths to homelessness with complicated interactions among a host of factors; however, data scarcity at the metro level precludes robust estimation of many correlated covariate effects. In the presence of limited data, it is essential to focus on a sparse predictive model. For this reason, we utilize changes in rental costs as our sole predictor and emphasize that we cannot draw causal conclusions from our analysis.

In Section 2, we discuss the data used in our analysis and necessary pre-processing steps to account for geographic mismatches between counties and continuums of care. Section 3 describes the Bayesian dynamic model that hierarchically shares information across all metros under study. Efficient information sharing, both locally in time and hierarchically across all metros, facilitates sharper inference on the relationship between rental costs and homelessness. Our hierarchical dynamic model allows us to estimate local relationships between homelessness and rental costs whereas the cross-sectional regression model in Byrne et al. (2013) estimates a single global effect. We discuss prior information and how that information translates to prior distributions for model parameters in Section 4. Model fitting with a custom Markov chain Monte Carlo algorithm is discussed in Section 5, and Section 6 presents results and addresses questions (Q1) - (Q5). Section 7 concludes with a discussion of our findings.

2. Data. The data in our study comes from three different sources: the U.S. Census Bureau, HUD, and the housing website Zillow. For the continuums of care in the 25 largest metros, we observe a collection of three time series that correspond to (i) the total number of people living in the metro, (ii) the *counted* number of homeless in the continuum(s) of that metro, and (iii) the ZRI for the metro.

For the total population data, we use county-level population estimates reported by the U.S. Census Bureau (U.S. Census Bureau, 2016). The homeless counts are the number of individuals experiencing homelessness (both sheltered and unsheltered) at a point-in-time as reported by HUD (U.S. Department of Housing and Urban Development, 2016). While the first two series of interest are fairly self-explanatory, the use of ZRI warrants further discussion.

According to Zillow's description in Bun (2012), ZRI is computed so that it does not depend on the set of homes that are currently for rent. Using proprietary statistical models trained on market rent data, Zillow estimates the market rent for every home, regardless of whether that home is currently for rent. The rent estimates of individual homes account for home attributes, physical characteristics, prior sales, tax assessments, and geographic location. The ZRI is computed to be the median of the market rent estimates for the housing stock in each metro.

One potential concern is that Zillow's estimated rent levels are skewed high as a consequence of training data from a biased sample of high quality homes. To mitigate the impact of this potential bias, we focus on percent differences in ZRI from one year to the next, ΔZRI, and do not directly utilize the level of ZRI itself.

One of the challenges in working with the HUD point-in-time data is that the jurisdiction of the HUD-defined continuums of care do not always agree with the boundaries of cities or counties. Often, each county will have a single continuum; however, cases exist where this is not true. In some counties, there may be more than one continuum (e.g, Cook County, IL and Fulton County, GA have two). Other times, there may be multiple counties in a single continuum (e.g., the Denver, CO continuum spans seven different counties, and the New York City continuum spans five). Table 1 maps the 25 metros under study to the underlying HUD continuum(s) of care. In each metro, if the continuum does not match up with a single county, we construct a synthetic unit of analysis by following one of two approaches. If a county includes multiple continuums, we aggregate homeless totals reported by each continuum in the county. If a continuum includes multiple counties, we aggregate population totals reported by the different counties that make

TABLE 1

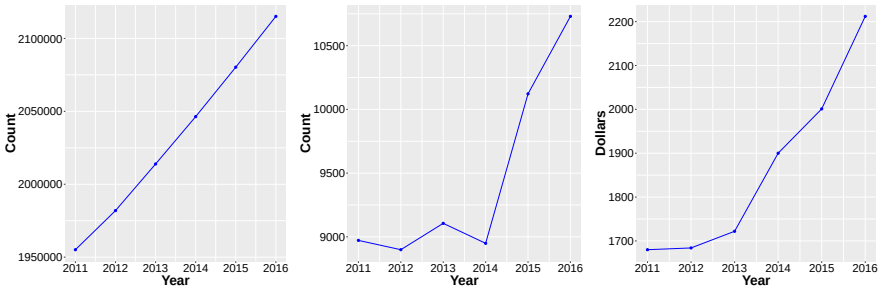
HUD continuums of care and state counties that correspond to the 25 largest metropolitan areas under study. In cases where more than one continuum of care is in a county, we aggregate homeless counts to form a synthetic continuum for that county. When a single continuum spans multiple counties, we construct a population-weighted ZRI measure and aggregate total population figures across the multiple counties.

	Metro area	HUD continuum of care	Counties
1	New York, NY	NY-600	New York, Bronx, Queens, Kings, Richmond
2	Los Angeles-Long Beach-Anaheim, CA	CA-600, CA-606, CA-607, CA-612	Los Angeles
3	Chicago, IL	IL-510, IL-511	Cook
4	Dallas-Fort Worth, TX	TX-600	Dallas
5	Philadelphia, PA	PA-500	Philadelphia
6	Houston, TX	TX-700	Harris, Fort Bend
7	Washington, DC	DC-500	District of Columbia
8	Miami-Fort Lauderdale, FL	FL-600	Miami-Dade
9	Atlanta, GA	GA-500, GA-502	Fulton
10	Boston, MA	MA-500	Suffolk
11	San Francisco, CA	CA-501	San Francisco
12	Detroit, MI	MI-501, MI-502	Wayne
13	Riverside, CA	CA-608	Riverside
14	Phoenix, AZ	AZ-502	Maricopa
15	Seattle, WA	WA-500	King
16	Minneapolis-St Paul, MN	MN-500	Hennepin
17	San Diego, CA	CA-601	San Diego
18	St. Louis, MO	MO-500, MO-501	St. Louis
19	Tampa, FL	FL-501	Hillsborough
20	Baltimore, MD	MD-501, MD-505	Baltimore
21	Denver, CO	CO-503	Adams, Arapahoe, Boulder, Broomfield, Denver, Douglas, Jefferson
22	Pittsburgh, PA	PA-600	Allegheny
23	Portland, OR	OR-501	Multnomah
24	Charlotte, NC	NC-505	Mecklenburg
25	Sacramento, CA	CA-503	Sacramento

up a single continuum and construct a population-weighted ZRI metric.

We also focus on year-over-year changes in the metro-specific ZRI rather than the absolute level of the ZRI itself. This standardizes the analysis of rental markets across metros. The result is a data set of synthetic continuums that properly record the counted number of homeless, total population, and changes in rent levels in each metro. Because the ZRI is only available after October 2010, the time series for these three quantities are observed from 2011 - 2016 at an annual frequency. The homeless count data and ZRI are recorded each January, but the intercensal total population estimates from the Census Bureau are dated July 1. There is a six month temporal mismatch in both the homeless count and ZRI and the total metro population series. To account for this mismatch, we linearly interpolate the census population estimates to align the data so that the average of year t and $t-1$ is an estimate of the population in January for metro i. Figure 1 presents these three time series for the All Home King County continuum in Seattle, WA.

The count of homeless in Seattle/King County has dramatically increased since 2014 (Figure 1b); however, the total population (Figure 1a) also significantly increased over that same time period. The King County,



(a) Total population (b) Homeless count (c) ZRI (median rent)

Fig 1: Data from the All Home King County (WA) continuum of care from 2011 - 2016. Left: the total population in King County has rapidly increased in recent years. Increased population creates increased demand for rental housing and community services. Middle: The number of homeless counted in King County has dramatically increased since 2014. Right: The median rent, as measured by the ZRI, demonstrates the same basic pattern of increases as the count of people experiencing homelessness.

WA data demonstrate the need for modeling the homelessness rate to control for increases in the total population. The ZRI for King County, shown in Figure 1c, has similarly increased.

In order to properly calculate the homelessness rate, it is necessary to account for time-variation in the count accuracy. We define the count accuracy to be the probability that a person who is homeless will be accounted for in the homeless count. If the count accuracy improves over time, more homeless are likely to be counted. In a scenario where the homeless count has improved, an increase in the number of homeless counted does not necessarily imply that the total size of the homeless population has increased. The count may simply represent a higher fraction of the total homeless population.

To account for the homeless not included in the HUD-reported count data, we impute the total homeless population in each metro from 2011-2016 and examine the impact of different trajectories in the count accuracy on the inferred homelessness rate. Figure 2 illustrates the distribution of the unobserved total number of homeless over time in King County, WA if we assume that the count accuracy does not improve with time and approximately 80% of homeless are included in the count.

It is reasonable to assume time variation in the count accuracy: in many continuums, the count accuracy may incrementally improve each year; in

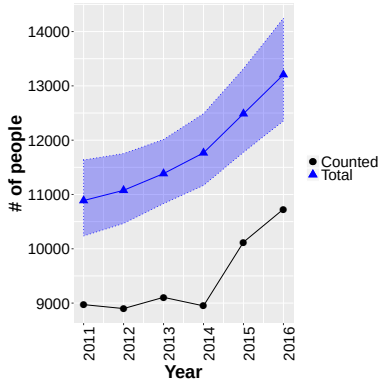


Fig 2: Counted number of homeless (points) and imputed mean (triangles) and 95% predicted interval (shaded region) of the total number of homeless in King County. In this illustration, we assume the expected count accuracy is constant over time when inferring the distribution of the total number of homeless.

some continuums, the count accuracy could degrade over time due to lack of funding; in others, the accuracy may jump at a single year. A primary objective of our study is to assess the impact of different trajectories in the count accuracy on the relationship between homelessness rates and changes in ZRI. The model and prior distribution for different trajectories of the count accuracy will be discussed further in Sections 3.3 and 4.1.

3. Model. In this section, we develop a joint statistical model for collections of population-level and subpopulation counts. For each metro, we model (i) the number of homeless counted, (ii) the true number of homeless, and (iii) the total number of people living in the metro. Of the three quantities, only two are observed: the homeless counted and the total number of people. The true number of people experiencing homelessness is not observed, and we treat it as missing data. The total population of a metro (as reported by the Census) is modeled as a noisy observation of the true total population.

Figure 3 is the graphical representation of our dynamic Bayesian hierarchical model. The random variable $N_{i,t}$ is the total number of people that live in metro i in year t. The $N_{i,1:T}$ variables depend on the dynamic process governing population growth, $\lambda_{i,1:T}$. The expected increase in population from one year to the next in metro i is modeled by parameter ν_i, and the global population growth is modeled by parameter $\bar{\nu}$. Section 3.1

develops the total population model in greater detail.

The total number of homeless, $H_{i,t}$, depends on $N_{i,t}$ and the probability of being homeless, $p_{i,t}$. The log odds of homelessness, $\psi_{i,t} = \log\left(\frac{p_{i,t}}{1-p_{i,t}}\right)$, is modeled by a dynamic process that depends on changes in ZRI. The relationship between change in ZRI and the log odds of homelessness is modeled hierarchically by parameter ϕ_i with global mean $\bar{\phi}$. The full model for $H_{i,t}$ and the dynamics of $\psi_{i,t}$ are discussed in Section 3.2. The counted number of homeless, $C_{i,t}$, depends on $H_{i,t}$ and the probability that a homeless person is counted, $\pi_{i,t}$. We call $\pi_{i,t}$ the count accuracy and discuss the count data generating process in Section 3.3.

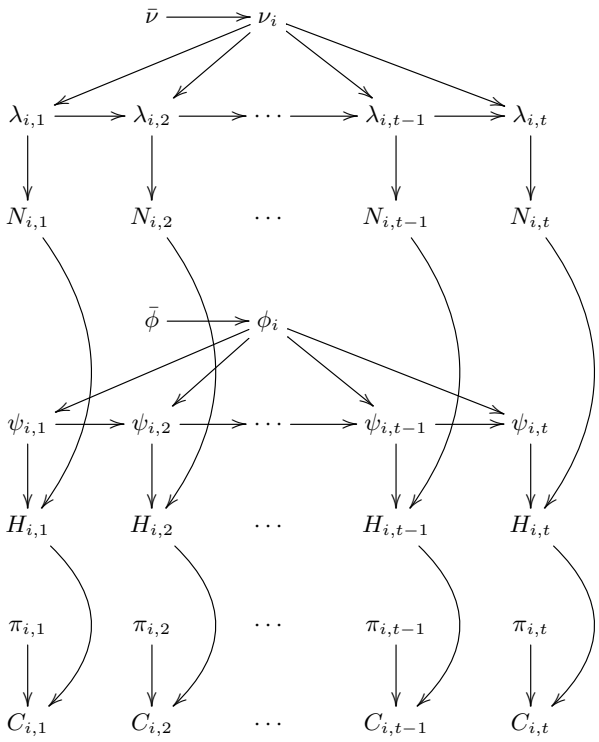


Fig 3: Graphical model of the continuum-level homeless population. In metro i in year t, the total population is modeled by $N_{i,t}$, the homeless population is modeled by $H_{i,t}$, and the number of homeless counted is modeled by $C_{i,t}$. The dynamical processes and associated parameters are outlined in Sections 3.1 - 3.3.

3.1. *Total population model.* Significant interest lies in homelessness *rates* that facilitate comparison across different metros. The total population of the metro, or the denominator in a rate calculation, is uncertain. The intercensal population estimates reported annually by the U.S. Census Bureau are noisy. To properly quantify the uncertainty in the estimated homelessness rate of each metro, it is necessary to account for the uncertainty in the total population size.

A secondary reason for modeling the total population is that it facilitates forecasting. In order to forecast the size of the homeless population in future years, it is necessary to know the size of the future total population. A dynamic model for the total population enables a model-based forecast of the homeless population.

The total population size for metro i in year t, $N_{i,t}$, is modeled as a time-indexed Poisson random variable that allows for growth and decay in the population of each metro.

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

$$\lambda_{i,t} = \bar{\lambda}_i \theta_{i,t} \tag{3.2}$$

The Poisson rate, $\lambda_{i,t}$, is the product of a static scale factor, $\bar{\lambda}_i$, and a latent time-varying component $\theta_{i,t}$. The dynamics of $\lambda_{i,t}$ are driven by a dynamic process on the unit interval, $\theta_{i,t} \in (0, 1)$. Modeling $\lambda_{i,t}$ as the product of the scaling factor and the dynamic term $\theta_{i,t}$ provides an intuitive and computationally tractable dynamic model for Poisson counts. An auxiliary Poisson-Binomial thinning step for efficient computation is discussed in Section 5.1. The unit-interval-constrained dynamic process $\theta_{i,1:T}$ is constructed with the logistic transformation and a real-valued stochastic process $\eta_{i,1:T}$.

$$\theta_{i,t} = \frac{e^{\eta_{i,t}}}{1 + e^{\eta_{i,t}}} \tag{3.3}$$

$$\eta_{i,t} = \eta_{i,t-1} + \nu_i + v_{i,t}, \qquad v_{i,t} \sim N(0, \sigma^2_{\eta_i}) \tag{3.4}$$

The nonstationary $\eta_{i,1:T}$ process is a random walk with a metro-specific drift term, ν_i. The drift component is aimed at modeling population dynamics in cities like Seattle and Detroit. In Seattle, the population is rapidly growing which would correspond to a positive drift ($\nu_i > 0$). On the other hand, the population in Detroit has recently decreased, which would correspond to negative drift ($\nu_i < 0$). To borrow information across metros, we model the drift components hierarchically. The parameter $\bar{\nu}$ may be interpreted as

the expected drift in population across all metros.

$$\nu_i = \bar{\nu} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2_{\nu_i}) \tag{3.5}$$

$$\bar{\nu} \sim N(0, \sigma^2_{\bar{\nu}}) \tag{3.6}$$

Because $\eta_{i,1:T}$ is nonstationary, the Poisson marginals $p(N_{i,t}|\lambda_{i,t}), \ldots, p(N_{i,T}|\lambda_{i,T})$ are not identically distributed. Consequently, $N_{i,1:T}|\lambda_{i,1:T}$ is a nonstationary process for the total population counts. While the PoINAR method of Aldor-Noiman et al. (2016) would be suitable for stationary population modeling, the expected total populations in these metro areas are clearly changing over time. A second modeling alternative would be to transform the large counts with a natural logarithm and model the transformed response with a Gaussian dynamic model. Despite the computational simplicity of such a model, we prefer to directly model the count data with discrete, time-varying distributions to more aptly characterize the uncertainty in the observed data.

3.2. *Homeless population model.* In a metro with $N_{i,t}$ total residents, some small fraction of the residents will be homeless. For this reason, it is natural to model the total number of people experiencing homelessness, $H_{i,t}$, with a time-indexed binomial distribution. The binomial parameter $p_{i,t}$ is the unobserved probability that a person in metro i is homeless in year t (i.e., the homelessness rate).

$$H_{i,t}|N_{i,t}, p_{i,t} \sim Binomial(N_{i,t}, p_{i,t}) \tag{3.7}$$

One of our primary objectives is to include ZRI as a covariate in the dynamic model for the homeless probability $p_{i,t}$. We achieve this by modeling the log odds of homelessness.

$$p_{i,t} = \frac{e^{\psi_{i,t}}}{1 + e^{\psi_{i,t}}} \tag{3.8}$$

$$\psi_{i,t} = \psi_{i,t-1} + \phi_i \Delta ZRI_{i,t} + w_{i,t}, \qquad w_{i,t} \sim N(0, \sigma^2_{\psi}) \tag{3.9}$$

The dynamic process that controls the homelessness rate, $\psi_{i,1:T}$, linearly depends on the year-over-year rate of change in the ZRI, $\Delta ZRI_{i,t}$.

$$\Delta ZRI_{i,t} = \frac{ZRI_{i,t} - ZRI_{i,t-1}}{ZRI_{i,t-1}} \tag{3.10}$$

The regression coefficient ϕ_i models the relationship between change in rent levels and change in homelessness rates. As a concrete example, if ZRI increases by 1% in continuum i from one year to the next, the expected log odds of homelessness will increase by $.01\phi_i$.

The connection between increased rental costs and homelessness rates is well established in the homelessness literature (Hanratty, 2017; Fargo et al., 2013; Byrne et al., 2013; Stojanovic et al., 1999; O'Flaherty, 1995; Sclar, 1990)). To model this positive relationship, the prior distribution for regression coefficient ϕ_i favors values greater than zero but no truncation at zero is forced. The data inform the degree to which increasing rent levels are associated with higher homelessness rates. The parameter ϕ_i is modeled hierarchically across metros to borrow strength and provide a more robust estimation of the relationship between rent increases and homelessness.

$$\phi_i \sim N(\bar{\phi}, \sigma^2_{\phi_i}) \tag{3.11}$$

$$\bar{\phi} \sim N(m_{\bar{\phi}}, \sigma^2_{\bar{\phi}}) \tag{3.12}$$

As noted at the beginning of Section 3, $H_{i,t}$, is not observed. Only the imperfect homeless count, $C_{i,t}$, is observed. In our study, we treat $H_{i,t}$ as missing data and impute it to estimate each ϕ_i. By modeling the relationship between $\Delta ZRI_{i,t}$ and the imputed $H_{i,t}$, we obtain a more reliable quantification of the uncertainty in the posterior distribution for ϕ_i and $\bar{\phi}$.

3.3. *Homeless count model.* Plant-capture studies and postcount surveys have demonstrated that homeless counts systematically understate the number of people experiencing homelessness (Hopper et al., 2008). While single night counts are imperfect, it is not clear that there exist feasible alternatives. Logistics, expenses, and privacy concerns preclude volunteers and continuums from counting every person without a home.

We model the imperfection in the homeless counts with a binomial thinning step. Of the true number of homeless, $H_{i,t}$, only $C_{i,t}$ of them are counted. It is $C_{i,t}$ that we observe.

$$C_{i,t} \sim Binomial(H_{i,t}, \pi_{i,t}) \tag{3.13}$$

$$\pi_{i,t} \sim Beta(a_{i,t}, b_{i,t}) \tag{3.14}$$

The parameter $\pi_{i,t}$ is the probability that a person experiencing homelessness is counted, and it is modeled with a beta distribution. In other words, $\pi_{i,t}$ is the accuracy of the count. If $\pi_{i,t} = 1$, the count in metro i in year t is perfectly accurate and every homeless person is counted.

Observe that we model $\pi_{i,t}$ and $\pi_{i,t-1}$ as independent random variables. While the count accuracy is surely time-varying and may exhibit trends, we believe there is no clear dependence of $\pi_{i,t}$ on $\pi_{i,t-1}$. Factors driving count accuracy such as weather and volunteer turnout are unrelated across

years. Even the count methodology utilized by a continuum may change from one year to the next. As an example, in 2017, the All Home King County continuum of care overhauled its count methodology to enhance the accuracy (Beekman, 2016). Rather than sending volunteers to known areas where homeless congregate, as in previous years, volunteers covered each census tract in the county. In addition, volunteers were lead by guides who were either currently or recently homeless themselves. As a result, the number of homeless counted in January 2017 was significantly higher than the number counted in January 2016. Due to changes in methodology, it is not necessarily accurate to conclude that the size of the homeless population, $H_{i,t}$, dramatically increased. By providing a mechanism for changes in count accuracy in each metro from one year to the next, it is possible to more reliably assess the local relationship between increased rental costs and the homeless population.

The count accuracy itself is an unknown quantity, and since we do not observe $H_{i,t}$, it is not possible to learn $\pi_{i,t}$. Instead of trying to learn $\pi_{i,t}$, we marginalize it out so that $C_{i,t}|H_{i,t} \sim \text{Beta-binomial}(H_{i,t}, a_{i,t}, b_{i,t})$. Despite the lack of an underlying dynamic model for the count accuracy, we examine the impact of different time trends in $E[\pi_{i,t}]$ on posterior inference for ϕ_i. The trends are achieved through specification of the $a_{i,t}$ and $b_{i,t}$ parameters. Given the sequences of expected values and variances for $\pi_{i,t}$ – which we assume are provided by the agencies conducting the counts – the hyperparameters $a_{i,t}$ and $b_{i,t}$ may be computed from (3.15) and (3.16).

$$a_{i,t} = E[\pi_{i,t}]\left(\frac{(1-E[\pi_{i,t}])E[\pi_{i,t}]}{Var(\pi_{i,t})} - 1\right) \tag{3.15}$$

$$b_{i,t} = \frac{Var(\pi_{i,t})}{E[\pi_{i,t}]^2}\left(\frac{a_{i,t}^2}{E[\pi_{i,t}]} + a_{i,t}\right) \tag{3.16}$$

By modeling each $\pi_{i,t}$ with an independent beta distribution, it is possible to easily achieve different types of accuracy trajectories. We discuss three trajectories of specific interest in Section 4.1.

4. Prior Distributions. With limited data for each metro, it is critically important to elicit well-informed prior distributions. Information from data that predates 2011, existing literature, and the expert opinion of homeless count coordinators have been combined to elicit prior distributions for four components of the model: (i) the count accuracy that is formalized through $\pi_{i,1:T}$ (Section 4.1); (ii) the relationship between homelessness and rising rental costs, which is modeled with the regression coefficient ϕ_i (Sec-

tion 4.2); (iii) the dynamic process $\eta_{i,1:T}$ that governs the total population (Section 4.3); and (iv) the dynamic process $\psi_{i,1:T}$ that governs the total homeless subpopulation dynamics (Section 4.4).

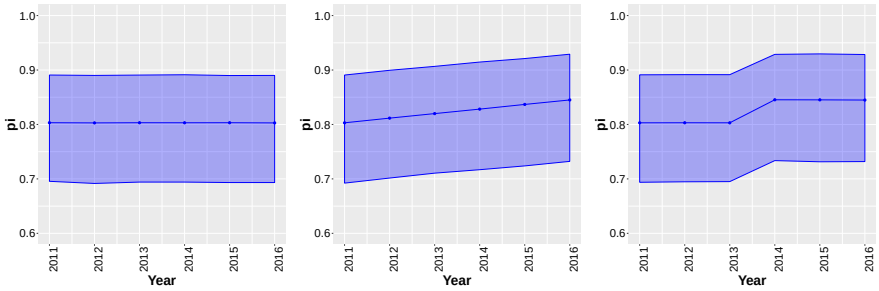
4.1. *Priors for count accuracy.* We base our prior distribution for count accuracy on a study by Hopper et al. (2008), who report evidence that 60-70% of unsheltered individuals in New York were visible and included in the city's 2005 count. They discuss one plant-capture study where only 59% of participants were counted. The number of homeless used in our study includes *sheltered* homeless as well. Hopper et al. (2008) note that counts of sheltered homeless are more reliable than the counts of unsheltered homeless. To elicit our prior for count accuracy, we compute a weighted average of accuracy for sheltered and unsheltered populations, respectively. We use homeless counts from 2010, $C_{i,0}$, to compute this weighted average in year $t = 0$.

$$E[\pi_{i,0}] = (0.95)\,\frac{C_{i,0}^{\text{sheltered}}}{C_{i,0}} + (0.6)\,\frac{C_{i,0}^{\text{unsheltered}}}{C_{i,0}}. \tag{4.1}$$

Our prior expectation is that the probability that a sheltered homeless person is included in the homeless count is 0.95, which allows for a small discrepancy between the true number and counted number of sheltered homeless due to administrative and other count errors that may occur. From the Hopper study, we believe the probability that an unsheltered homeless person is included in the homeless count to be approximately 0.6. Because each metro has a different proportion of sheltered and unsheltered homeless, each metro is assigned a unique baseline prior distribution for count accuracy based on the 2010 data. We develop prior distributions for $\pi_{i,1:T}$ that exhibit different expected trajectories: constant, linear, and step functions in time. In each of these cases, $Var(\pi_{i,t}) = .0015$ is chosen so that reasonable prior mass covers the $E[\pi_{i,t}] \pm 0.10$ interval.

The constant case corresponds to a count that utilizes relatively consistent procedures and resources from one year to the next. In this case, the mean and variance of the accuracy are constant over time (for all t, $E[\pi_{i,t}] = E[\pi_{i,0}]$). With $E[\pi_{i,t}]$ and $Var(\pi_{i,t})$, calculation of $a_{i,t}$ and $b_{i,t}$ follows directly from (3.15) and (3.16). The prior for $\pi_{i,1:T}$ with constant count accuracy in King County, WA is presented in Figure 4a.

The linear case corresponds to a count where the accuracy incrementally improves by a fixed amount (called δ_i) until it reaches one, as shown in (4.2). We assume that δ_i is known and is ideally specified by the agency conducting the count. Alternatively, δ_i can be adjusted to examine sensitivity of



(a) Constant accuracy (b) Linear gains (c) Step in accuracy

Fig 4: Different prior beliefs about the trajectory of $\pi_{i,1:T}$ in King County, WA. The solid lines are the expected count accuracy over time, and the shaded interval corresponds to the 99% prior uncertainty interval. Left: constant count accuracy. Middle: incremental (linear) increases in count accuracy. Right: Step in count accuracy.

inference to different accuracy scenarios. This is the approach adopted here. As an example, to consider an increase of $\bar{\delta}$ in the accuracy of the unsheltered homeless count, δ_i is computed as in (4.3). We assume that sheltered homeless are always counted with 95% accuracy, and the improvement in accuracy of $\bar{\delta}$ applies only to the unsheltered count.

$$E[\pi_{i,t+1}] = \min(E[\pi_{i,t}] + \delta_i, 1) \tag{4.2}$$

$$\delta_i = \bar{\delta}\left(\frac{C_{i,0}^{\text{sheltered}}}{C_{i,0}}\right) \tag{4.3}$$

In the step scenario, the accuracy dramatically increases at a specific point in time due to improved count methodology. This is observed in practice with the All Home King County continuum as discussed in Section 3.3. In this case, we assume that the year of change for metro i, τ_i, is known. For $t < \tau_i$, $E[\pi_{i,t}] = E[\pi_{i,0}]$. For $t \geq \tau_i$, $E[\pi_{i,t}] = E[\pi_{i,0}] + \delta_i$. A step in $E[\pi_{i,t}]$ occurs at time τ_i. Figure 4c illustrates a hypothetical step in count accuracy for King County in 2014. It is possible that there could be multiple steps for each metro and that there exists sequences $\tau_i^1, \tau_i^2, \ldots$ and $\delta_i^1, \delta_i^2, \ldots$ where steps of different size occur in different years. Because we do not know τ_i for each metro, we do not investigate the step scenario further; however, with consultation of each local coordinator, this may be a very promising area of future work.

4.2. *Priors for ϕ_i and $\bar{\phi}$.* We use previous work by Byrne et al. (2013) to form the basis of our prior distribution for $\bar{\phi}$. Byrne et al. (2013) found that in metropolitan continuums, when median rent increased by \$100, the expected homelessness rate increased by 6.34%. The average log odds of homelessness across all continuums in 2010 was $\bar{f}_0 := -5.5$. The average ZRI across continuums in 2010 was \$1534. So a \$100 increase in median rent would translate to a percent change in ZRI of $\frac{100}{1534} \approx 6.5\%$. This leads to calculation of the expectation of $\bar{\phi}$ based on $\bar{f}_0$, the 6.5% increase in ZRI, and the expected increase in the homelessness rate of 6.34%:

$$\frac{1+\exp\left\{-\bar{f}_0\right\}}{1+\exp\left\{-\bar{f}_0-\frac{\$100}{\$1534}m_{\bar{\phi}}\right\}}=1.0634. \tag{4.4}$$

We calculate that $m_{\bar{\phi}} = 0.94$. Because of differences in methodology and data, we use $\sigma^2_{\bar{\phi}} = 0.005$ so that there is reasonable prior uncertainty about $\bar{\phi}$. We let $\sigma^2_{\phi_i} = 0.05$ so that there is modest shrinkage of each local effect toward the global mean $\bar{\phi}$. To examine the prior uncertainty in the relationship between increases in ZRI and increases in homelessness implied by our choices of $m_{\bar{\phi}}$, $\sigma^2_{\bar{\phi}}$, and $\sigma^2_{\phi_i}$, we simulate from the marginal prior distribution for percent changes in the homelessness rate (see Figure 5).

Although we inform our prior using the results from Byrne et al. (2013), whose methodology we are trying to advance, notice a few things in Figure 5. One is that our prior is diffuse, and it becomes increasingly diffuse with larger percent increases in ZRI. Second, the inferred posterior concentrates on different values than the prior, indicating that we are indeed learning from data. Third, very little posterior mass is placed on values less than zero, providing evidence for the positive relationship between rising rents and homelessness. The conclusion is that using the Byrne et al. result is a useful way to center our prior.

4.3. *Prior for $\eta_{i,1:T}$.* The sampling distribution for the total population, $N_{i,t}$, depends on $\eta_{i,t}$ through the Poisson rate, $\lambda_{i,t}$ (refer to (3.1) - (3.3)). Recall that $\lambda_{i,t}$ is the product of a scaling factor, $\bar{\lambda}_i$, and a dynamic process on the unit interval, $\theta_{i,t}$. We let the expectation of $\lambda_{i,0}$ be the 2010 population, $N_{i,0}$. This is achieved by fixing $\bar{\lambda}_i = 2 \times N_{i,0}$ and $E[\theta_{i,0}] = 0.5$ (i.e., $E[\eta_{i,0}] = 0$). The prior variance of $\eta_{i,0}$ is fixed to be 0.0001, as we are confident that the Poisson rate of the total population in 2010 is the observed total population.

We let $\nu_i \sim N(\bar{\nu}, 0.01)$ and $\bar{\nu} \sim N(0, 0.005)$. The innovation variance of the $\eta_{i,1:T}$ process is fixed to be 0.0001 so that ν_i primarily drives changes in

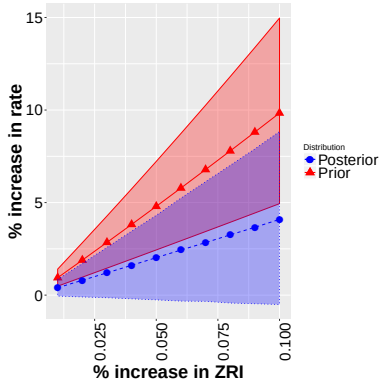


Fig 5: Implied prior and posterior distribution of % change in homelessness rate with increases in ZRI for an arbitrary metro i. The triangle-marked (dashed) line is the prior (posterior) mean, and the shaded regions with solid (dotted) boundaries mark the 95% prior (posterior) credible interval.

the Poisson rate. The implied marginal distribution of $N_{i,1:T}$ in King County, WA is presented in Figure 6a. Observe that the distribution is centered at the 2010 King County population and allows significant uncertainty over the six year period. While the prior variance on each $\eta_{i,t}$ is relatively small, the large magnitude of the scaling factor, $\bar{\lambda}_i$, results in a relatively diffuse marginal distribution for $N_{i,t}$.

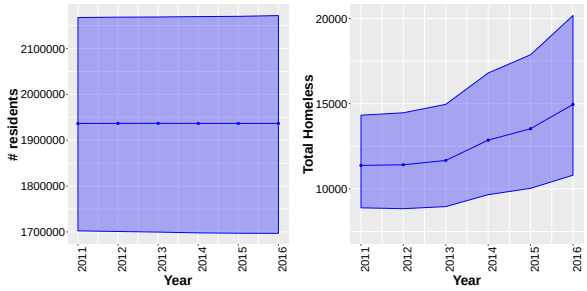
4.4. *Prior for* $\psi_{i,1:T}$. We utilize the counted number of homeless in 2010 to specify the prior expectation $E[\psi_{i,0}]$. The conditionally binomial sampling distribution for $H_{i,0}|\psi_{i,0}, N_{i,0}$ in (3.7) yields the expectation

$$E[H_{i,0}|\psi_{i,0}, N_{i,0}] = \frac{1}{1+e^{-\psi_{i,0}}} N_{i,0}. \tag{4.5}$$

Solving for $\psi_{i,0}$ results in (4.6).

$$\psi_{i,0} = \log\left(\frac{E[H_{i,0}|\psi_{i,0}]/N_{i,0}}{1 - E[H_{i,0}|\psi_{i,0}]/N_{i,0}}\right) \tag{4.6}$$

Because we observe the noisy total population $N_{i,0}$ from the Census estimate in 2010, we can compute a value for $\psi_{i,0}$ given the expectation $E[H_{i,0}|\psi_{i,0}]$. Though we do not observe $E[H_{i,0}|\psi_{i,0}]$, we use an approximation to center the prior distribution of $\psi_{i,0}$ and compensate for the approximation with moderate prior uncertainty. We approximate the expected



(a) Total (b) Homeless

Fig 6: Marginal prior distributions for $N_{i,t}$ and $H_{i,t}|ZRI_{1:T}$ in King County, WA. The solid lines are the the prior means and the shaded regions are the 95% prior uncertainty intervals. Left: implied prior distributions for the total population, $N_{i,1:T}$. Right: Prior for total homeless population, $H_{i,1:T}|ZRI_{1:T}$. The upward trend in the implied prior for the total homeless population is due to observed increases in ZRI.

total number of homeless in 2010 as an expected inflation of the observed count, $E[H_{i,0}|\psi_{i,0}] \approx E\left[\frac{1}{\pi_{i,0}}\right] C_{i,0}$. $C_{i,0}$ is the 2010 homeless count value and $E[1/\pi_{i,0}]$ is the expectation of the reciprocal count accuracy in 2010. The multiplier $E[1/\pi_{i,0}]$ is evaluated by Monte Carlo simulation. This leads to a prior expectation as defined in (4.7).

$$E[\psi_{i,0}] := \log\left(\frac{E\left[\frac{1}{\pi_{i,0}}\right] C_{i,0}/N_{i,0}}{1 - E\left[\frac{1}{\pi_{i,0}}\right] C_{i,0}/N_{i,0}}\right) \tag{4.7}$$

The time zero variance is chosen to be $\sigma^2_{\psi_0} = 0.01$, as this provides a three standard deviation interval of ± 0.3 around the prior mean. The result is that $\psi_{i,0} \sim N(E[\psi_{i,0}], \sigma^2_{\psi_0})$. One reason that the prior specification of $\psi_{i,0}$ is important is that the implied prior distribution for the total homeless population, $H_{i,t}$, depends on $\psi_{i,t}$ (refer to (3.7) and (3.8)). The implied prior distribution for $H_{i,1:T}|ZRI_{1:T}$ is presented in Figure 6b. The 95% prior interval spans a very reasonable range for each $H_{i,t}$.

The innovation variance of the dynamic process is fixed to be $\sigma^2_{\psi} = 0.001$ (see (3.9)). We observe in synthetic data experiments that the innovation variance of $\psi_{i,1:T}$ must be small in order to accurately learn ϕ_i and $\bar{\phi}$. If σ^2_{ψ} is large relative to $Var\left(\phi_i \Delta ZRI_{i,t}\right)$, changes in the homelessness rate are modeled as noise in $\psi_{i,t}$ rather than driven by changes in ZRI. The ratio

$\frac{Var(\phi_i \Delta ZRI_{i,t})}{\sigma^2_\psi}$ can be thought of as a signal-to-noise ratio for $\psi_{i,t}$.

We also observe in synthetic data experiments that reliably inferring $\psi_{i,1:T}$ and ϕ_i requires that a metro's homelessness rate exceed 0.05% of the total population. Because $p_{i,t}$ is the logistic transformation of $\psi_{i,t}$ (see (3.8)), the derivative $\frac{dp_{i,t}}{d\psi_{i,t}} \to 0$ as $|\psi_{i,t}|$ increases. Flat tails of $p_{i,t}$ as a function of $\psi_{i,t}$ mean that in metros with very low homeless rates, practically observed changes in homeless counts are consistent with a wide range of changes in ZRI. Under such conditions, it is not possible to reliably estimate ϕ_i. Inference on ϕ_i degrades along the continuum of decreasing $\psi_{i,t}$, but we set a limit based on our empirical studies. We do not trust inference for metros where the homelessness rate is less than 0.05%, or when $\psi_{i,t} < -7.6$.

5. Markov chain Monte Carlo. Our objective is to sample from the posterior distribution

$$p(H_{1:25,1:T}, \eta_{1:25,1:T}, \psi_{1:25,1:T}, \phi_{1:25}, \bar{\phi}, \nu_{1:25}, \bar{\nu} | N_{1:25,1:T}, C_{1:25,1:T}). \tag{5.1}$$

To sample from the posterior, we develop a custom Pólya-Gamma Gibbs sampler for dynamic Bayesian logistic regression (Polson et al., 2013; Windle et al., 2013, 2014). The Pólya-Gamma augmentation strategy allows us to harness a forward filtering and backward sampling (FFBS) algorithm that is commonly used to fit Bayesian dynamic models (Fruhwirth-Schnatter, 1994; Carter and Kohn, 1994). We found that a burn-in of 25,000 samples and 50,000 samples collected after burn-in were sufficient for reproducible inferences. The MCMC simulation took approximately 4 hours to run on a MacBook Pro.

5.1. *Sampling steps.* There are ten different sampling steps required in the MCMC algorithm. The first step is for an auxiliary random variable whose only purpose is to facilitate computation when $N_{i,t}|\tilde{\lambda}_i, \theta_{i,t} \sim Poisson(\tilde{\lambda}_i \theta_{i,t})$ (refer to (3.1) and (3.2)). To construct this marginal distribution, we model the auxiliary $Z_{i,t} \sim Poisson(\tilde{\lambda}_i)$ and the observed $N_{i,t}$ conditionally binomial, $N_{i,t}|Z_{i,t}, \theta_{i,t} \sim Binomial(Z_{i,t}, \theta_{i,t})$. The Binomial-Poisson thinning strategy results in the desired marginal distribution for $N_{i,t}$ and a computationally tractable method for making inference on $\eta_{i,t}$, ν_i, and $\bar{\nu}$. The full conditional for the auxiliary $Z_{i,t}$ is shown in Step 1.

Step 2 and Step 6 use Pólya-Gamma data augmentation to allow a forward filtering backward sampling strategy. Step 3 and Step 7 sample the auxiliary Pólya-Gamma variables $\omega_{i,t}$ and $\zeta_{i,t}$. The collection of auxiliary variables $Z_{1:25,1:T}$, $\omega_{1:25,1:T}$, and $\zeta_{1:25,1:T}$ are numerically integrated out from the posterior by discarding posterior samples. Each sampling step is outlined below.

1. For each i, t, sample the auxiliary $Z_{i,t}$ from a shifted Poisson by first sampling
   $j = Z_{i,t} - N_{i,t} | \tilde{\lambda}_i, \theta_{i,t} \sim Poisson\left((1-\theta_{i,t})\tilde{\lambda}_i\right)$ and then fixing $Z_{i,t} = j + N_{i,t}$.
2. For each i, sample the dynamic process that governs total population growth, $\eta_{i,1:T} | N_{i,1:T}, \omega_{i,1:T}$, with an FFBS algorithm.
   (a) compute forward filtered distribution $\eta_{i,t} | N_{i,1:t}, Z_{i,1:t}, \nu_i, \omega_{i,1:t} \sim N(m_{i,t}, S_{i,t})$
      - $S_{i,t} := \left(\omega_{i,t} + \frac{1}{S_{i,t-1}+\sigma_\eta^2}\right)^{-1}$
      - $m_{i,t} := S_{i,t}\left(N_{i,t} - \frac{1}{2}Z_{i,t} + \frac{m_{i,t-1}+\nu_i}{S_{i,t-1}+\sigma_\nu^2}\right)$
   (b) sample recursively $\eta_{i,t} | \eta_{i,t+1}, N_{i,1:t}, \omega_{i,1:t} \sim N(\tilde{m}_{i,t}, \tilde{S}_{i,t})$
      - $\tilde{S}_{i,t} := \left(\frac{1}{S_{i,t}} + \frac{1}{\sigma_\eta^2}\right)^{-1}$
      - $\tilde{m}_{i,t} := \tilde{S}_{i,t}\left(\frac{m_{i,t}}{S_{i,t}} + \frac{\eta_{i,t+1}-\nu_i}{\sigma_\eta^2}\right)$
3. For each i, t, sample the auxiliary Pólya-Gamma random variates to augment the total population variable, $\omega_{i,t} | Z_{i,t}, \eta_{i,t} \sim PG(Z_{i,t}, \eta_{i,t})$.
4. For each i, sample the parameter controlling expected population growth in metro i, $\nu_i | \bar{\nu}$, $\eta_{i,1:T} \sim N(\tilde{m}_{\nu_i}, \tilde{\sigma}^2_{\nu_i})$.
   - $\tilde{\sigma}^2_{\nu_i} := \left(\frac{1}{C_0+\sigma_\eta^2} + \frac{T-1}{\sigma_\eta^2} + \frac{1}{\sigma^2_{\nu_i}}\right)^{-1}$
   - $\tilde{m}_{\nu_i} := \tilde{\sigma}^2_{\nu_i}\left(\frac{\eta_{i,1}}{C_0+\sigma_\eta^2} + \frac{1}{\sigma_\eta^2}\sum_{t=2}^{T}(\eta_{i,t} - \eta_{i,t-1}) + \frac{\bar{\nu}}{\sigma^2_{\nu_i}}\right)$
5. Sample the expected total population growth globally across metros, $\bar{\nu} | \nu_{1:25} \sim N\left(\left(\frac{N}{\sigma^2_{\nu_i}} + \frac{1}{\sigma^2_{\bar{\nu}}}\right)^{-1} \frac{1}{\sigma^2_{\nu_i}} \sum_{i=1}^{25} \nu_i, \left(\frac{N}{\sigma^2_{\nu_i}} + \frac{1}{\sigma^2_{\bar{\nu}}}\right)^{-1}\right)$.
6. For each i, sample the dynamic process for the log odds of homelessness, $\psi_{i,1:T} | N_{i,1:T}, H_{i,1:T}, \phi_i, \omega_{i,1:T}$, with an FFBS algorithm.
   (a) compute forward filtered distribution $\psi_{i,t} | N_{i,1:t}, H_{i,1:t}, \zeta_{i,1:t} \sim N(f_{i,t}, q_{i,t})$
      - $q_{i,t} := \left(\zeta_{i,t} + \frac{1}{q_{i,t-1}+\sigma_\psi^2}\right)^{-1}$
      - $f_{i,t} := q_{i,t}\left(H_{i,t} - \frac{1}{2}N_{i,t} + \frac{f_{i,t-1}+\phi_i \Delta ZRI_{i,t}}{q_{i,t-1}+\sigma_\psi^2}\right)$
   (b) sample recursively $\psi_{i,t} | \psi_{i,t+1}, N_{i,1:t}, H_{i,1:t}, \zeta_{i,1:t}, \phi_i \sim N(\tilde{f}_{i,t}, \tilde{q}_{i,t})$
      - $\tilde{q}_{i,t} := \left(\frac{1}{q_{i,t}} + \frac{1}{\sigma_\psi^2}\right)^{-1}$

- $\tilde{f}_{i,t} := \tilde{q}_{i,t}\left(\frac{f_{i,t}}{q_{i,t}} + \frac{\psi_{i,t+1}-\phi_i \Delta ZRI_{i,t}}{\sigma^2_\psi}\right)$

7. For each i, t, sample the auxiliary Pólya-Gamma random variates to augment the total homeless variable, $\zeta_{i,t}|N_{i,t}, \psi_{i,t} \sim PG(N_{i,t}, \psi_{i,t})$.
8. For each i, sample the parameter governing the relationship between change in ZRI and change in homelessness in metro i, $\phi_i|\psi_{i,1:T}, \bar{\phi} \sim N(m_{\phi_i}, \Sigma_{\phi_i})$.
   - $\Sigma_{\phi_i} := \left(\frac{(\Delta ZRI_{i,1})^2}{\sigma^2_{\psi_0}+\sigma^2_\psi} + \frac{\sum_{t=2}^T (\Delta ZRI_{i,t})^2}{\sigma^2_\psi} + \frac{1}{\sigma^2_\phi}\right)$
   - $m_{\phi_i} := \Sigma_{\phi_i}\left(\frac{\bar{\phi}}{\sigma^2_{\phi_i}} + \frac{\Delta ZRI_{i,1}(\psi_{i,1}-f_{i,0})}{\sigma^2_{\psi_0}+\sigma^2_{\psi_i}} + \frac{\sum_{t=2}^T \Delta ZRI_{i,t}(\psi_{i,t}-\psi_{i,t-1})}{\sigma^2_{\psi_i}}\right)$
9. Sample the global mean parameter for the change in ZRI and change in homelessness,
   $\bar{\phi}|\phi_{1:25} \sim N\left(\left(\frac{25}{\sigma^2_{\phi_i}} + \frac{1}{\sigma^2_{\bar{\phi}}}\right)^{-1}\left(\frac{1}{\sigma^2_{\phi_i}}\sum_{i=1}^{25}\phi_i + \frac{m_{\bar{\phi}}}{\sigma^2_{\bar{\phi}}}\right), \left(\frac{25}{\sigma^2_{\phi_i}} + \frac{1}{\sigma^2_{\bar{\phi}}}\right)^{-1}\right)$.
10. For each i, t, sample the total number of people experiencing homelessness in metro i and year t, $H_{i,t}$, from $p(H_{i,t}|N_{i,t}, C_{i,t}, p_{i,t}, a_{i,t}, b_{i,t}) \propto \frac{\Gamma(H_{i,t}+1)}{\Gamma(C_{i,t}+1)\Gamma(H_{i,t}-C_{i,t}+1)}\frac{\Gamma(C_{i,t}+a_{i,t})\Gamma(H_{i,t}-C_{i,t}+b_{i,t})}{\Gamma(H_{i,t}+a_{i,t}+b_{i,t})}\frac{\Gamma(a_{i,t}+b_{i,t})}{\Gamma(a_{i,t})\Gamma(b_{i,t})}\ldots$
    $\times \binom{N_{i,t}}{H_{i,t}} p_{i,t}^{H_{i,t}}(1-p_{i,t})^{(N_{i,t}-H_{i,t})}$.

Sampling $H_{i,t}$ in Step 10 incorporates the beta-binomial marginal distribution for homeless counts, and it does not depend on $\pi_{i,t}$ but instead on $a_{i,t}$ and $b_{i,t}$. Sampling $H_{i,t}$ requires sampling from a discrete distribution with support $[C_{i,t}, N_{i,t}]$. This large range creates a computational bottleneck as it involves evaluating densities at each value in the support. In practice, though, posterior probability is concentrated on values much closer to the lower end of the support. It is possible to speed up computation by setting a threshold after which the support is truncated. Once posterior probability falls below 1×10^{-8}, we stop evaluating the densities and truncate the support.

5.2. *Posterior Predictive Distributions.* To examine the predicted increase in total homeless populations associated with increases in ZRI, we utilize the posterior predictive distribution for the total homeless population in each metro, $H_{i,t}|C_{1:25,1:T}, N_{1:25,1:T}$. The main quantify of interest is the distribution of the increase in the homeless population when the observed change in ZRI, $\Delta ZRI_{i,t}$, increases by an $x > 0$. The increase is modeled by $\left(H^x_{i,t} - H_{i,t}\right)|C_{1:25,1:T}, N_{1:25,1:T}$, which is the difference between the predicted homeless total for a change in ZRI of $\Delta ZRI_{i,t} + x$ and the baseline

prediction at $\Delta ZRI_{i,t}$.

We draw samples from this posterior with a three step procedure that approximates the integral:

$$\begin{aligned} & p\left(H_{i,t}^{x}-H_{i,t}|N_{1:25,1:T},C_{1:25,1:T}\right) \\ & =\int p\left(H_{i,t}^{x}-H_{i,t}|\psi_{i,t},\phi_{i},N_{1:25,1:T},C_{1:25,1:T}\right)p\left(\psi_{i,t},\phi_{i}\right)d\psi_{i,t}d\phi_{i}. \end{aligned} \tag{5.2}$$

The procedure relies on the m^{th} posterior sample of (i) the relationship between ZRI and homelessness, $\phi_i^{(m)}$, (ii) the log odds of homelessness, $\psi_{i,t}^{(m)}$, and (iii) the Census reported estimate of the total population $N_{i,t}$. The procedure is detailed below.

1. Construct the m^{th} sample of log odds of homelessness where $\Delta ZRI_{i,t}$ is increased by x.

$$\psi_{i,t}^{(m),x}=\psi_{i,t}^{(m)}+\phi_{i}^{(m)}x \tag{5.3}$$

2. Generate a prediction for the total homeless population at $\Delta ZRI_{i,t}+x$ by sampling

$$H_{i,t}^{(m),x}\sim Binomial(N_{i,t},p_{i,t}^{x}) \tag{5.4}$$

   where $p_{i,t}^{x}$ is the same logistic transformation of $\psi_{i,t}^{x}$ as in (3.8).
3. Compute the difference

$$H_{i,t}^{(m),x}-H_{i,t}^{(m)} \tag{5.5}$$

We go one step further and also examine the predicted change in counted homeless under increased ZRI, $\left(C_{i,t}^{*,x}-C_{i,t}^{*}\right)|C_{1:25,1:T},N_{1:25,1:T}$. Samples from this distribution are drawn by thinning the m^{th} MCMC samples $H_{i,t}^{(m),x}$ and $H_{i,t}^{(m)}$ with a beta-binomial step

$$C_{i,t}^{(m),*,x}\sim Beta-Binomial\left(H_{i,t}^{(m),x},a_{i,t},b_{i,t}\right) \tag{5.6}$$

and computing the difference $C_{i,t}^{(m),*,x}-C_{i,t}^{(m),*}$.

5.3. *Reproducible MCMC inference.* We verify that our MCMC simulation generates reproducible inference about the relationship between increases in homelessness and increases in ZRI by examining the posterior distribution for ϕ_i. Ten different MCMC simulations are run, and inferences

from two simulations j and j' are compared by computing $|E[\phi_i^{(j)}] - E[\phi_i^{(j')}]|$. Figure 7 illustrates the largest deviation across simulations by computing $\max_j |E[\phi_i^{(j)}] - E[\phi_i^{(1)}]|$ for each metro. Each point in the histogram corresponds to the largest difference in posterior mean in reference to the first simulation for each of the 25 metros. The small values of these maximum differences in Figure 7 give us confidence that our MCMC simulation generates reproducible inferences.

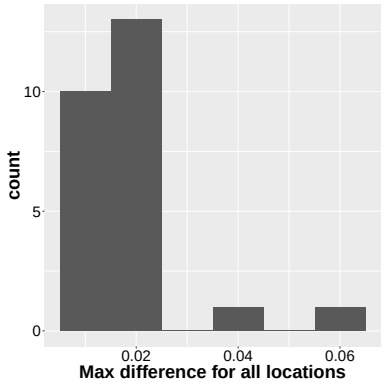


Fig 7: The maximum difference in posterior means of $\phi_1, \ldots, \phi_{25}$ for 10 different MCMC simulations. Each of the 25 points in the histogram corresponds to $\max_j |\phi_i^{(j)} - \phi_i^{(1)}|$, for each of the 25 metros denoted by i. The superscript index j denotes the MCMC simulation. The very small differences indicate that our MCMC simulations generate reproducible inference in ϕ_i.

6. Results. We seek to answer five questions, (Q1) - (Q5). Each of these questions is answered (in order) in Sections 6.1 - 6.5. In Section 6.1, we examine changes in homelessness rates from 2011-2016 across all metropolitan areas. In Section 6.2, the inferred relationship between increased ZRI and increases in homelessness is presented. Posterior predictive distributions for additional homeless counts are presented in Section 6.3, and the imputed distributions for the total number of homeless in each metro are presented in Section 6.4. Section 6.5 discusses our forecasts for the total homeless populations in 2017.

6.1. *Percent changes in the homelessness rate.* The inferred increases in homelessness rates from 2011 - 2016 are illustrated in Figure 8a. We present results under two scenarios for the trajectory of the count accuracy: (i) the mean of the count accuracy is constant over time (i.e. $\bar{\delta}$ in (4.3) is

zero); and (ii) the mean of the unsheltered count accuracy increases by 2% annually until it reaches 100% (i.e. $\bar{\delta} = .02$).

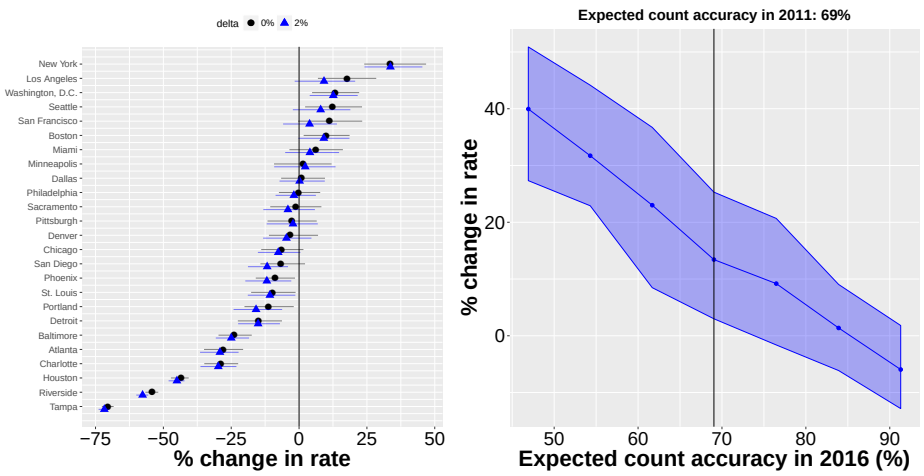
Metros where the rate of homelessness almost certainly increased from 2011-2016 when $\bar{\delta} = 0$ include New York, Los Angeles, Washington, D.C., Seattle, San Francisco, and Boston. For these cities, the 95% posterior credible interval for the percent change in the homelessness rate is bounded below by 0%, giving confidence that these homelessness rates did in fact increase. For each of these metros, the posterior mean increase in the homelessness rate exceeds 10%. In response to its growing homeless population, the City of Seattle has declared an official state of emergency (Beekman and Broom, 2015). We adopt this moniker and characterize these metros as in similar states of emergency.

Metros where the homelessness rate almost certainly decreased when $\bar{\delta} = 0$ include Phoenix, St. Louis, Portland, Detroit, Baltimore, Atlanta, Charlotte, Houston, Riverside, and Tampa. For these cities, the 95% posterior credible interval for the percent change in the homelessness rate is bounded above by 0%, giving confidence that the homelessness rate did in fact decrease. For all but Phoenix and St. Louis, the posterior mean decrease in the homelessness rate exceeded 10%, and it seems real progress has been made in reducing homelessness in this group.

A third group of cities exists where the percent change in the homelessness rate has neither significantly increased nor decreased in either scenario. The 95% posterior credible interval for the change in the homelessness rate includes zero in Miami, Minneapolis, Dallas, Philadelphia, Sacramento, Pittsburgh, Denver, Chicago, and San Diego. The situation remains largely unchanged in this group, and the current homelessness rate is the status quo.

Observe in Figure 8a that New York and Los Angeles exhibit different sensitivities to change in the count accuracy over time. In New York, a city with a predominantly sheltered population, the inferred percent increase is essentially unchanged between the two scenarios. In Los Angeles, a warm-weather city with a large unsheltered population, the difference between the $\bar{\delta} = 0$ and $\bar{\delta} = 0.02$ cases is large, as demonstrated by separation of the posterior means. Equation (4.3) demonstrates that, in metros with large unsheltered populations, a $\bar{\delta}$ increase in the accuracy of an unsheltered homeless count leads to large changes in the overall count accuracy, $\pi_{i,t}$.

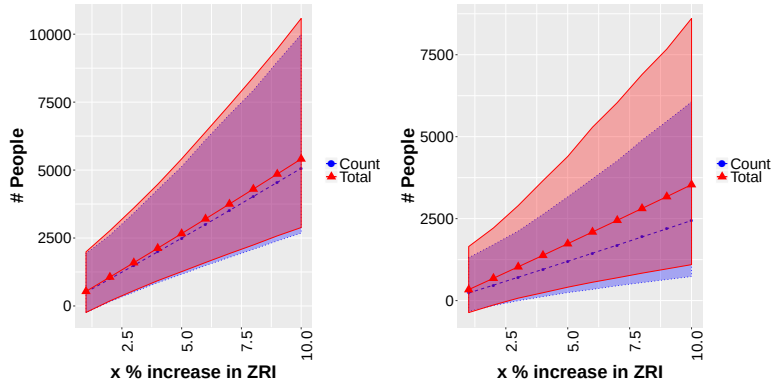
In Figure 8b, we examine how inference on the change in homelessness rates from 2011 - 2016 can change with different values of $\bar{\delta}$. We focus on Los Angeles, above considered in a "state of emergency" and see that the posterior distribution for the 2011-2016 change in the homelessness rate



(a) Rate Change 2011-2016 (b) LA Rate Sensitivity

Fig 8: Left: Posterior distribution in percent change in homelessness rate from 2011 to 2016. The middle point in each segment is the posterior mean and the line segment encompasses the 95% posterior credible interval. For each metro, there is a posterior presented when the count accuracy is modeled with a constant mean over time ($\bar{\delta} = 0$) and a posterior when the count accuracy of unsheltered homeless improves by 2% each year ($\bar{\delta} = .02$). Right: Sensitivity of the percent increase in homelessness rate from 2011 - 2016 to different choices of $\bar{\delta}$ in Los Angeles. The vertical line marks the count accuracy in 2011

depends on the count accuracy. If the expected count accuracy in LA in 2011 was 0.69 and it remained unchanged from 2011 - 2016 (i.e., $\bar{\delta} = 0$), Figure 8b shows that the homelessness rate almost certainly increased over that time, with a posterior expected increase of 13.4%. On the other hand, if LA had improved its counting method over this time frame, the increased homeless counts could be explained away by the increased count accuracy rather than an actual increase in the homelessness rate. For example, if the expected count accuracy in LA in 2011 was 0.69 and it incrementally increased to 0.91 over that 6 year window (i.e., $\bar{\delta} = 6\%$), it is likely that the homelessness rate decreased, with a posterior expected change in the homelessness rate of -5.9%. That is, the increased homeless count in LA is entirely explained by improvements in the count accuracy, and the overall homeless rate likely decreased over the six year window. The conclusion again is that any inferences drawn about changes in homeless populations



(a) New York (b) Los Angeles

Fig 9: Posterior predictive distributions of increased homeless counts $(C^{*,x}_{i,t} - C^{*}_{i,T}|C_{1:25,1:T}, N_{1:25,1:T})$ and total homeless populations $(H^{x}_{i,t} - H_{i,T}|C_{1:25,1:T}, N_{1:25,1:T})$ associated with increases in ZRI for both New York and Los Angeles when $\bar{\delta} = 0$. The dashed line and shaded interval with dotted boundaries correspond to the posterior mean and 95% predictive interval of increases in the homeless count. The triangle-marked line and interval with solid boundaries correspond to the posterior mean and 95% predictive interval of increases in the total homeless population.

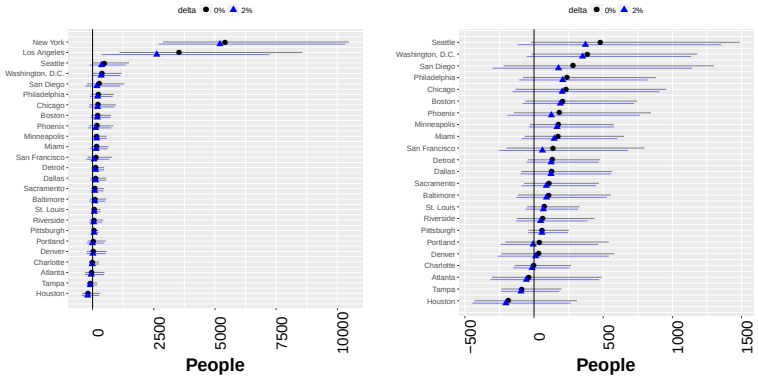
are highly sensitive to assumptions about the count accuracy. Sensitivity analyses similar to the one presented in Figure 8b are presented for each metro in Glynn and Fox (2018), an online supplement to this article.

6.2. *Rental costs and homelessness.* To examine the predicted increase in homelessness and homeless counts as ZRI increases, we focus on the posterior predictive distributions $\left(H^{x}_{i,t} - H_{i,t}\right) |C_{1:25,1:T}, N_{1:25,1:T}$ and $\left(C^{*,x}_{i,t} - C^{*}_{i,t}\right) |C_{1:25,1:T}, N_{1:25,1:T}$. Section 5.2 provides complete details for sampling from these distributions.

We find that, for a fixed percent increase in ZRI of $x = 10\%$, the predicted increase in homelessness is largest in New York and Los Angeles (see Figure 10). Predicted increases in homeless counts are robust to whether we set $\bar{\delta} = 0$ or $\bar{\delta} = .02$, as one would hope; however, the predicted count increases map to different increases in total homelessness under different prior beliefs about $\pi_{i,t}$. In Figure 9, the posterior predictive distributions for the increase in total and counted homeless are illustrated for different increases in ZRI in New York and Los Angeles.

In New York, the large sheltered population and high count accuracy imply that the distributions of increased counts and total homeless populations are nearly identitical (Figure 9a). If the ZRI in New York increases by $x = 10\%$, given 2016 levels of homelessness, we expect that the homeless population will increase by 5,413 people, with 95% posterior probability of the homelessness increase in New York being more than 2,896 people and less than 10,523 people. In Los Angeles, the lower overall count accuracy implies more separation between the distributions of increased counted and total homeless (Figure 9b). Under the same $x = 10\%$ increase in ZRI in Los Angeles, we expect that 3,536 people will become homeless, with 95% posterior probability of more than 1,106 people and less than 8,554 people.

Figure 10a summarizes the predicted increase in the total homeless population when ZRI increases by $x = 10\%$ across all metros. The distributions of increases presented in Figure 10 account for the different sizes of metros with binomial sampling as shown in (5.4) and (5.6) (i.e. the values $N_{i,t}$ are larger for larger metros). We expect the largest increases to occur in the largest metros (New York and Los Angeles), and this is confirmed by our analysis of the data. For the increase in the homeless population associated with increases in ZRI, we report the one-sided 95% posterior credible interval to shed light on the right tail of the distribution.



(a) 25 largest metros (b) Excluding NY and LA

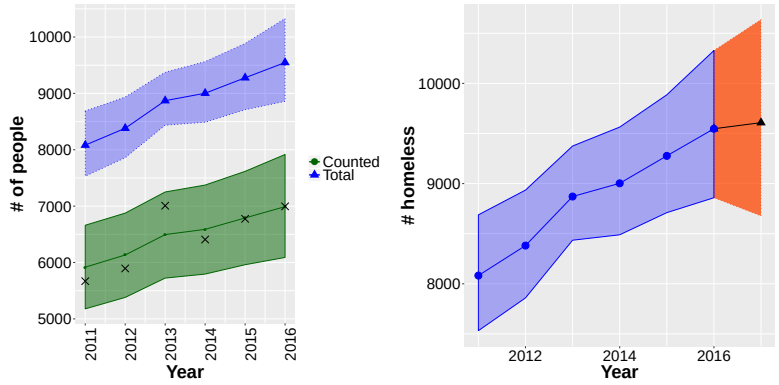
Fig 10: Predicted increase in homeless population when ZRI increases by $x = 10\%$ in 2016. The points are the posterior mean of $\left(H^x_{i,T} - H_{i,T}\right) | C_{1:25,1:T}, N_{1:25,1:T}$. The line segment spans the one-sided (right-tail) 95% posterior credible interval. In the left panel, results are presented for all metros. In the right panel, New York and Los Angeles are excluded for more careful inspection of the remaining 23 metros.

We note Seattle and Washington, D.C. as metros of interest, even though the 95% credible intervals for the predicted increase in the homeless population cross the zero threshold in Figure 10b. In Seattle the posterior mean increase in the homeless population is 479 people and the posterior probability that the increase is positive is 0.942. A similar story emerges in Washington, D.C., where the posterior mean increase in homeless population is 386 people and the posterior probability that the increase is positive is 0.941. With at least 0.94 posterior probability of an increase in the homeless population, Seattle and Washington D.C. exhibit a meaningful statistical relationship between increased ZRI and increased homelessness. The posterior probability of an increase in the homeless population when ZRI increases by 10% is less than 0.9 for all metros but New York, Los Angeles, Washington, D.C., and Seattle. While the expected increases in homelessness in some metros may be large (e.g., San Diego), the variance in the posterior predictive distribution precludes us from confidently concluding that the predicted changes in the population are strictly positive. See Table 2 for posterior probabilities of homeless increases associated with ZRI increases of 10% in each metro. Predicted increases in homelessness as a function of increases in ZRI, as shown in Figure 9 for New York and Los Angeles, are available for each metro in the supplement (Glynn and Fox, 2018).

6.3. *Additional homeless counts.* The number of homeless that HUD reports in each continuum is from an annual point-in-time (PIT) count conducted in January. An important limitation of PIT counts is that no standard errors or other measures of uncertainty are reported, leaving decision makers without important context. One way of assessing sampling variability in PIT data is to construct posterior predictive distributions for the outcome of additional counts, allowing count coordinators to quantify uncertainty in the PIT data.

In this section, we report the posterior predictive distribution for this hypothetical second homeless count. The prediction, denoted by $C^*_{i,t}|C_{1:25,1:T}, N_{1:25,1:T}$, conditions on both the observed counts and the census reported total populations in all metros. Figure 11a presents the predicted outcome from additional homeless counts for San Francisco, a metro with one of the larger increases in the homelessness rate from 2011-2016. Observe that the posterior mean of $C^*_{i,t}|C_{1:25,1:T}, N_{1:25,1:T}$ is a filtered and retrospectively smoothed quantity. The smoothing is apparent in 2013, when the HUD reported count appears to be an outlier relative to prior and subsequent HUD reported counts. Though the 2013 posterior mean is pulled slightly upward toward the reported count, the model does not overfit the data. In

the remaining years, the posterior mean closely tracks the reported HUD counts. In 2016, the HUD reported count of homeless in San Francisco was 6,996. If a second count were conducted in 2016, we expect the counted number of homeless would have been 6,991, with 95% posterior probability of being more than 6,090 and less than 7,916.



(a) Predicted Homeless (b) 2017 Forecast

Fig 11: Predicted homeless totals for San Francisco, CA. Left: Predicted number of counted homeless in additional (hypothetical) counts and predicted number of total homeless in San Francisco. The 'x'marks are the actual HUD reported counts, and the solid line with points is the filtered and retrospectively smoothed mean of the posterior predictive distribution $C^*_{i,1:T}|C_{1:25,1:T}, N_{1:25,1:T}$. The line with triangles is the mean of the posterior predictive distribution $H_{i,1:T}|C_{1:25,1:T}, N_{1:25,1:T}$, and the associated shaded region is the 95% predictive interval. Right: Predicted number of total homeless in San Francisco, CA in 2017, $H_{i,T+1}|C_{1:25,1:T}, N_{1:25,1:T}, ZRI_{i,T+1}$ The line with triangles is the mean of the out-of-sample prediction for 2017, and the shaded region with dotted boundaries is the out-of-sample predictive interval.

The predictive distribution $C^*_{i,t}\ |C_{1:25,1:T}, N_{1:25,1:T}$ provides policymakers and resource constrained counting agencies with a principled and data-driven way of conducting synthetic "additional"homeless counts to quantify uncertainty in the PIT data. The posterior predictive distributions for additional 2016 counts in all metros are summarized in Table 2. Each metro has its own version of Figure 11a in the online supplement (Glynn and Fox, 2018).

TABLE 2

Summary of posterior distributions across metros for 2016. The number of counted homeless reported by HUD is presented the first column. The Synthetic counts column corresponds to the posterior predictive distribution for a second hypothetical count in metro i in 2016, $C^*_{i,T}|C_{1:25,1:T}, N_{1:25,1:T}$. The Total homeless column corresponds to the posterior predictive distribution for the total number of homeless in metro i in 2016, $H_{i,T}|\ C_{1:25,1:T},\ N_{1:25,1:T}$. The Forecasted homeless (2017) column is the posterior predictive distribution for the total homeless population in 2017, $H_{i,T+1}|\ C_{1:25,1:T},\ N_{1:25,1:T}$. In all cases, the first number reported is the posterior mean, with the 95% posterior predictive interval in parenthesis. The Prob column is the posterior probability that the predicted change in the homeless population when rent rises 10% is greater than zero given that $\delta = 0$: $P(\left(H^{10}_{i,T} - H_{i,T} > 0\right)\ |C_{1:25,1:T}, N_{1:25,1:T})$.

Metro	HUD	Synthetic count	Total homeless	Forecast (2017)	Prob
New York	73,523	74,272 (66,027, 80,959)	79,348 (75,317, 84,222)	79,404 (72,716, 86,775)	0.999
Los Angeles	46,874	45,315 (39,144, 51,621)	65,585 (60,565, 70,919)	67,285 (60,145, 75,034)	0.993
Chicago	6,841	7,231 (6,399, 8,025)	8,209 (7,665, 8,833)	8,223 (7,434, 9,087)	0.849
Dallas	3,810	3,688 (3,288, 4,056)	4,180 (3,981, 4,449)	4,301 (3,888, 4,769)	0.829
Philadelphia	6,112	6,045 (5,386, 6,626)	6,645 (6,316, 7,074)	6,705 (6,109, 7,373)	0.892
Houston	4,031	4,709 (4,202, 5,137)	5,565 (5,245, 5,708)	5,907 (5,387, 6,433)	0.114
Wash., D.C.	8,350	8,167 (7,291, 8,858)	8,719 (8,415, 9,216)	8,907 (8,176, 9,733)	0.941
Miami	4,235	4,274 (3,787, 4,739)	4,919 (4,615, 5,265)	4,993 (4,515, 5,514)	0.879
Atlanta	4,546	5,011 (4,388, 5,618)	5,689 (5,221, 6,172)	5,781 (5,137, 6,471)	0.404
Boston	6,240	6,289 (5,585, 6,858)	6,687 (6,345, 7,110)	6,807 (6,208, 7,459)	0.892
San Francisco	6,996	6,991 (6,090, 7,916)	9,547 (8,859, 10,328)	9,610 (8,669, 10,644)	0.747
Detroit	2,612	2,871 (2,515, 3,216)	3,112 (2,851, 3,397)	3,121 (2,778, 3,495)	0.890
Riverside	2,165	2,574 (2,281, 2,846)	3,518 (3,376, 3,582)	3,638 (3,325, 3,959)	0.709
Phoenix	5,702	5,940 (5,273, 6,593)	6,997 (6,597, 7,497)	7,192 (6,513, 7,954)	0.815
Seattle	10,730	10,604 (9,350, 11,889)	13,207 (12,355, 14,245)	13,688 (12,311, 15,273)	0.943
Minneapolis	3,056	3,379 (2,957, 3,763)	3,631 (3,328, 3,953)	3,771 (3,348, 4,227)	0.920
San Diego	8,669	8,933 (7,795, 10,114)	11,899 (11,084, 12,898)	12,164 (10,908, 13,595)	0.817
St. Louis	1,713	1,750 (1,544, 1,940)	1,902 (1,778, 2,057)	1,939 (1,731, 2,167)	0.833
Tampa	1,817	2,068 (1,819, 2,296)	2,579 (2,397, 2,711)	2,634 (2,366, 2,908)	0.171
Baltimore	3,488	3,578 (3,174, 3,965)	4,045 (3,784, 4,348)	4,080 (3,685, 4,512)	0.783
Denver	5,728	6,026 (5,318, 6,727)	6,685 (6,210, 7,301)	6,813 (6,135, 7,603)	0.578
Pittsburgh	1,156	1,332 (1,163, 1,466)	1,423 (1,297, 1,517)	1,458 (1,292, 1,626)	0.832
Portland	3,914	4,044 (3,537, 4,559)	5,175 (4,780, 5,614)	5,267 (4,669, 5,927)	0.598
Charlotte	1,818	2,055 (1,804, 2,272)	2,248 (2,065, 2,402)	2,313 (2,058, 2,583)	0.489
Sacramento	2,500	2,611 (2,299, 2,918)	3,195 (2,975, 3,433)	3,300 (2,948, 3,682)	0.834

6.4. *Imputed total number of homeless.* Imperfect count accuracy leads to count totals that are less than the size of the total homeless population. By modeling the mechanism of count accuracy, we are able to include the uncounted number of homeless in our estimate of the size of the total homeless population. In this section, we predict the total number of homeless in each metro and year. We report the posterior distribution $H_{i,t}|C_{1:25,1:T}, N_{1:25,1:T}$. Observe that the posterior distribution does not condition on the count accuracy parameter, $\pi_{i,t}$. The count accuracy has been integrated out; however, the variance of $H_{i,t}|C_{1:25,1:T}, N_{1:25,1:T}$ is inextricably linked to the prior variance of the count accuracy, $\pi_{i,t}$. In this analysis, we fixed the prior variance to be 0.0015 so that prior mass would span the ± 0.1 interval. Though it is appealing to specify a diffuse prior for count accuracy, we found in practice that such a prior does not provide sufficient regularization. In settings with overly diffuse priors for count accuracy, inference for ϕ_i was not reproducible across MCMC simulations. This highlights the importance of

reliable prior information about count accuracy as it pertains to estimating the relationship between trends in ZRI and homelessness.

In Figure 11a, observe that, because the counting process is imperfect, the expected total number of homeless is more than the counted number of homeless. In San Francisco, we expect that, in 2016, there were 9,547 people experiencing homelessness, with 95% posterior probability that there were more than 8,859 and fewer than 10,328. Table 2 presents the posterior mean and 95% credible interval for the total number of homeless in 2016 for each metro.

6.5. *Forecasts for 2017.* Resources to address the needs of a homeless population are budgeted well in advance of the January point-in-time count. In order to allocate resources in communities with growing (shrinking) homeless populations, a forecast of the next year's total homeless population is needed. In this section, we forecast the total homeless population in each metro in January 2017.

Our forecasts of the homeless population for 2017 take into account both predicted increases in the 2017 total population and the January 2017 ZRI value. We report the one-year-ahead forecast

$$H_{i,T+1}|C_{1:25,1:T}, N_{1:25,1:T}, ZRI_{1:25,T+1}. \tag{6.1}$$

For true out-of-sample forecasting when $ZRI_{i,T+1}$ is not yet available, we could utilize Zillow's forecasted ZRI for metro i, $\hat{ZRI}_{i,T+1}$.

Figure 11b illustrates the year-ahead forecast in San Francisco. Forecasting the total homeless population one year in the future requires forecasting both the year-ahead total metro population and log odds of homelessness. The uncertainty in these component year-ahead forecasts accumulates, and the result is an uncertainty interval for the 2017 year-ahead forecast that increases relative to the intervals presented from 2011-2016. This is observed in Figure 11b as the predictive interval in 2017 fans out relative to the uncertainty interval from 2011-2016.

We predict that 9,610 people experienced homelessness in San Francisco on any given January night in 2017, with 95% posterior probability of more than 8,669 and less than 10,644 people. Although the January 2017 ZRI decreased 3.6% relative to its January 2016 value, we still expect a slight increase in San Francisco's total homeless population. The increase is largely driven by the model-based forecasted increase in San Francisco's total population. Each metro has a figure corresponding to Figure 11b presented in the supplement (Glynn and Fox, 2018). The forecasted mean and 95% posterior predictive intervals for each metro are shown in Table 2.

7. Discussion. We presented statistical evidence that the relationship between rental costs and homelessness depends on one's beliefs about the time-varying accuracy of homeless counts. We highlight this fact to encourage public policy researchers, policymakers, and continuum leaders to carefully quantify their beliefs and uncertainty about count accuracy or the inferences drawn from studies relying on these counts. While the prior beliefs about count accuracy that we elicit in this paper are informed by existing literature and our discussions with count coordinators and homelessness experts from around the country, we believe that collecting expert opinions from every continuum can lead to a more robust and informed study. We encourage other researchers in this area to explicitly model variation in count accuracy when conducting their own analyses.

We use ZRI as a frequently updated measure of metro-level market rent. While ZRI is frequently updated and responsive to changing market conditions, it is a potentially biased measure of market rent given that it is computed from an incomplete sample of rental homes. To limit the impact of ZRI's absolute level on our analysis, we focus on year-over-year percent differences in ZRI, a strategy that allows us to investigate the relationship between changes in homelessness to changes in ZRI.

We found in synthetic data experiments that making accurate inference on the relationship between ZRI and homelessness with the model outlined in this paper requires homelessness rates that exceed 0.05% of the total population (Section 4.4). To work with counts as large as possible, this implies that sheltered and unsheltered homeless totals should be combined in a single analysis of a metro's total homeless population. Furthermore, reliable estimates of the entire homeless population are significantly aided by a metro having either a large sheltered population or high count accuracy of the unsheletered population; utilizing data on unsheltered homeless populations alone does not yield reliable results. This observation fits with our broader theme of data quality.

Modeling count accuracy is another place where we have directly addressed data quality challenges. Acknowledging that a continuum's homeless counts are imperfect and that the accuracy varies from one year to the next should not be viewed in any way as a failure of the count coordinators or volunteers. We view quantifying the count accuracy as an important step in accounting for the uncertainty inherent in such a difficult undertaking.

In our analysis, we have used the time-varying count accuracy to impute the size of the total homeless population. We believe it is natural in this application to think of the total homeless population as missing data. Assuming that the total population size is the observed count has two flaws.

First, it understates the size of the homeless population. Second, it leads to overly confident estimates of regression coefficients $\phi_1, \ldots, \phi_{25}$. Imputing the missing homeless population size naturally resolves both of these problems.

In this application, proper uncertainty quantification is critical. Counties, city governments, shelters, and health care providers are likely to benefit from an expected range of the homeless population size when they budget resources. By reporting the 95% posterior credible intervals on the total homeless population predictions and the 2017 forecasts, we emphasize the uncertain size of homeless populations now and in the future.

Metro-specific estimates of the relationship between rental costs and homelessness allow for each metro to make more informed policy decisions about affordable housing initiatives. We provide evidence that the homelessness rate significantly increased from 2011 - 2016 in six metros: New York, Los Angeles, Washington, D.C., Seattle, San Francisco, and Boston. While we are unable to conclude that increased rent *causes* homelessness, we found that large increases in ZRI are *statistically associated* with increases in the homeless population in four of those six metros: New York, Los Angeles, Washington, D.C., and Seattle. We believe that our results provide context for policy discussions that are happening in cities across the United States.

Homelessness is a product of many complicated and intertwined factors. Though an increase in rental costs is surely an important factor, vacancy rates, supportive services, affordable housing, affordable healthcare, and lack of economic opportunity may contribute as well. To investigate the complicated interactions of these factors at the local level with our current modeling framework, more frequent metro-level homeless counts are needed. In addition to counts occurring more often, estimating the relationship of these factors to increased homelessness also requires researchers to have more informed prior beliefs about count accuracy. One potential way of obtaining additional information on count accuracy is with post-count surveys at soup kitchens and other service providers, as in Hopper et al. (2008).

Due to limited information about count accuracy, we make several assumptions to aid our analysis. Most notably, we assume a prior distribution for count accuracy in each metro. The prior mean is determined by the proportion of unsheltered homeless in 2010 and the assumption that 60% of unsheltered homeless are counted while 95% of sheltered homeless are counted. Naturally, the estimated size of the total homeless population is sensitive to these assumptions, with lower expected accuracy inflating the estimate of the total homeless population. Prior variance of the count accuracy prior also impacts how sharply we can estimate the relationship between homelessness and ZRI. The more diffuse the prior distribution for count accuracy,

the more diffuse the posterior distribution for our ϕ_i regression coefficients. While we believe the assumptions that we made to be reasonable based on existing literature and discussions with homeless experts and count coordinators, we again emphasize that stronger prior information is needed about count accuracy in every metro to sharpen this analysis.

There are additional limitations of our current approach. One is that it does not account for relocation in homeless populations. It is possible that people experiencing homelessness move to cities with more services and away from cities with fewer services. In this scenario, increases in the population of one metro are driven by decreases in another. Our present approach does not take into account network effects, and we assume that homeless relocation patterns are not a significant driver of trends across metros. At present, we do not know of data that would allow us to further investigate network effects. A second is that we rely on January count data. It is likely that seasonal patterns in homelessness exist, though our current annual data set does not provide insight into such seasonal fluctuations. Modeling network effects and seasonal fluctuations in homeless populations are important areas of future work.

SUPPLEMENTARY MATERIAL

Metro-level supporting figures
(doi: COMPLETED BY THE TYPESETTER; .pdf). As supplementary material, we present figures for each of the 25 largest metros in the United States from 2011-2017. For each metro, we present

a the posterior predictive distribution for homeless counts, $C^*_{i,1:T}|C_{1:25,1:T}, N_{1:25,1:T}$, and the imputed total homeless population size, $H_{i,1:T}|C_{1:25,1:T}, H_{1:25,1:T}$;
b the predictive distribution for the total homeless population in 2017, $H_{i,2017}|C_{1:25,1:T}, N_{1:25,1:T}$;
c the posterior distribution of increase in the total homeless population with increases in ZRI; and
d the sensitivity of the inferred increase in the homelessness rate from 2011 - 2016 to different annual changes in count accuracy.

References.

Aldor-Noiman, S., Brown, L., Fox, E., and Stine, R. (2016). Spatio-temporal low count processes with application to violent crime events. Statistica Sinica, 26:1587–1610.

Appelbaum, R. P., Dolny, M., Dreier, P., and Gilderbloom, J. I. (1991). Scapegoating rent control: Masking the causes of homelessness. Journal of the American Planning Association, 57(2):153–164.

Beekman, D. (2016). King countys homeless count could soar with new method of tallying. Seattle Times. [Online; accessed 05/29/2017].

Beekman, D. and Broom, J. (2015). Mayor, county exec declare state of emergency over homelessness. Seattle Times. [Online; accessed 04/15/2017].

Bohanon, C. (1991). The economic correlates of homelessness in sixty cities. Social Science Quarterly.

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

Burt, M. M. (1992). Over the edge: The growth of homelessness in the 1980s. Russell Sage Foundation.

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.

Coles, S. and Sparks, R. (2006). Extreme value methods for modelling historical series of large volcanic magnitudes. Statistics in volcanology, 1:47–56.

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

Cornulier, T., Robinson, R. A., Elston, D., Lambin, X., Sutherland, W. J., and Benton, T. G. (2011). Bayesian reconstitution of environmental change from disparate historical records: hedgerow loss and farmland bird declines. Methods in Ecology and Evolution, 2(1):86–94.

Early, D. W. and Olsen, E. O. (2002). Subsidized housing, emergency shelters, and homelessness: An empirical investigation using data from the 1990 census. Advances in Economic Analysis & Policy, 2(1).

Fargo, J. D., Munley, E. A., Byrne, T. H., Montgomery, A. E., and Culhane, D. P. (2013). Community-level characteristics associated with variation in rates of homelessness among families and single adults. American Journal of Public Health, 103(S2):S340–S347.

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). Supplement to Dynamics of homelessness in urban America. Annals of Applied Statistics.

Grimes, P. W. and Chressanthis, G. A. (1997). Assessing the effect of rent control on homelessness. Journal of Urban Economics, 41(1):23–37.

Hanratty, M. (2017). Do local economic conditions affect homelessness? impact of area housing market factors, unemployment, and poverty on community homeless rates. Housing Policy Debate, pages 1–16.

Honig, M. and Filer, R. K. (1993). Causes of intercity variation in homelessness. The American Economic Review, pages 248–255.

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.

Hudson, C. G. (1998). Estimating homeless populations through structural equation modeling. J. Soc. & Soc. Welfare, 25:136.

Kery, M. and Royle, A. J. (2010). Hierarchical modelling and estimation of abundance and population trends in metapopulation designs. Journal of Animal Ecology, 79(2):453–461.

Laska, E. M. and Meisner, M. (1993). A plant-capture method for estimating the size of

a population from a single sample. Biometrics, pages 209–220.
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.
McCandless, L. C., Patterson, M. L., Currie, L. B., Moniruzzaman, A., and Somers, J. M. (2016). Bayesian estimation of the size of a street-dwelling homeless population. Journal of Modern Applied Statistical Methods, 15(1):15.
Office of Community Planning and Development, Dept. of Housing and Urban Development (2009). Continuum of care 101. https://www.hudexchange.info/resources/documents/CoC101.pdf. [Online; accessed 04/23/2018].
O'Flaherty, B. (1995). An economic theory of homelessness and housing. Journal of Housing Economics, 4(1):13–49.
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. (1990). Does rent control cause homelessness? taking the claim seriously. Journal of policy analysis and Management, 9(1):89–93.
Quigley, J. M. and Raphael, S. (2001). The economics of homelessness: The evidence from north america. European Journal of Housing Policy, 1(3):323–336.
Quigley, J. M., Raphael, S., and Smolensky, E. (2001a). Homeless in america, homeless in california. Review of Economics and Statistics, 83(1):37–51.
Quigley, J. M., Raphael, S., and Smolensky, E. (2001b). Homeless in america, homeless in california. The Review of Economics and Statistics, 83(1):37–51.
Raphael, S. (2010). Housing market regulation and homelessness. How to house the homeless, pages 110–140.
Schwarz, C. J. and Seber, G. A. (1999). Estimating animal abundance: review iii. Statistical Science, pages 427–456.
Sclar, E. D. (1990). Homelessness and housing policy: a game of musical chairs. American Journal of Public Health, 80(9):1039–1040.
Stojanovic, D., Weitzman, B. C., Shinn, M., Labay, L. E., and Williams, N. P. (1999). Tracing the path out of homelessness: The housing patterns of families after exiting shelter. Journal of community Psychology, 27(2):199–208.
Tokdar, S. T., Grossmann, I., Kadane, J. B., Charest, A.-S., and Small, M. J. (2011). Impact of beliefs about atlantic tropical cyclone detection on conclusions about trends in tropical cyclone numbers. Bayesian Anal., 6(4):547–572.
Troutman, W., Jackson, J. D., and Ekelund, R. B. (1999). Public policy, perverse incentives, and the homeless problem. Public Choice, 98(1):195–212.
U.S. Census Bureau (2016). County population totals tables: 2010-2016. https://www.census.gov/data/tables/2016/demo/popest/counties-total.html. [Online; accessed 04/2/2017].
U.S. Department of Housing and Urban Development (2016). Pit and hic data since 2007. https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/. [Online; accessed 04/2/2017].
Windle, J., Carvalho, C. M., Scott, J. G., and Sun, L. (2013). Efficient data augmentation in dynamic models for binary and count data. ArXiv e-prints.
Windle, J., Polson, N. G., and Scott, J. G. (2014). Sampling Polya-Gamma random variates: alternate and approximate techniques. ArXiv e-prints.

Chris Glynn
Paul College of Business and Economics
University of New Hampshire
Durham, NH 03824 E-mail: christopher.glynn@unh.edu

Emily B. Fox
Statistics and Computer Science Depts.
University of Washington
Seattle, WA 98195 E-mail: ebfox@uw.edu