Connecting HUD Continuum of Care Point-in-Time Homeless Counts to US Census Geographies: 2005 to 2017

Zack W. Almquist[1] Nathaniel E. Helwig[2] Yun You[3]

[1] Corresponding Author; Departments of Sociology, School of Statistics, and Minnesota Population Center; University of Minnesota; almquist@umn.edu

[2] Departments of Psychology and School of Statistics; University of Minnesota; helwig@umn.edu

[3] School of Statistics; University of Minnesota; youxx143@umn.edu

Abstract

Connecting HUD Continuum of Care Point in Time Homeless Counts to US Census Geographies: 2005 to 2017

Housing and Urban Development (HUD) has initiated a required point-in-time count of the homeless across the United States in an administrative unit known as Continuum of Care (CoC). This unit unfortunately does not lineup with more commonly used administrative and functional units used by the US Census Bureau and other major survey firms. To rectify this issue we have employed modern methods of spatial disaggregation, matching and imputation to generate county data. Further, we employ a spatiotemporal model to explore the spatial and temporal change of homeless in the US from 2007-2017. In addition, we employ this model to back cast CoC and county level estimates for 2005 and 2006 which currently only have publicly available spatial data, but not count data. We finish with a brief discussion and overview of how our method works, and potential uses for the data in social science and policy research.

Keywords: Homeless, Homelessness, small area estimation, spatiotemporal modeling

1 Introduction

Measuring the homeless population is an important and ongoing issue in the United States. The count of homeless populations in any given area impacts US federal allocation of funds (Lucas, 2017) as well as private donations, both of which are designed to maximize their impact in the fight against homelessness. In recent years we have seen both an overall decline in homelessness in the US, and a simultaneous increase in concentration of homeless populations in a handful of areas (Corinth, 2017). In 2016, there was approximately 540,566 homeless in the US – representing about a 2% decline since 2015. Although most recently, in 2017, there has been an increase to approximately 553,742 homeless in the US which represents about a 1% increase since 2016. Much of these gains have been concentrated in major counties like Los Angeles and New York that have seen massive increases in homeless populations over this period (See Figure 9 to visualize this trend).

To fully understand these changes, and their relationship to other social processes of interest, one must be able to correctly estimate counts of the homeless and to connect these counts to larger social science databases, such as US Census demographic and economic data, or the National Opinion Research Center's General Social Survey (GSS). Our paper is one such method of doing this; we employ modern methods of spatial disaggregation, matching and imputation to generate a series of county level datasets from currently collected homeless counts, which are collected at the Housing and Urban Development (HUD)'s administrative unit known as Continuum of Care (CoC). Once these counts are linked to US Census geographies (i.e., counties) we can readily connect these to other important demographic and economic surveys and administrative data, such as the American Community Survey (ACS), which measures social and economic indicators for the US, or the IRS migration data which has provided in and out internal migration counts for the United States annually since 1990.

The history of attempting to enumerate the US homeless population is long, however, counting the homeless in the US really began in earnest starting in the 1970s and 80s as it took on great political importance. In the 1980s, homelessness in the US began to receive a national spotlight when advocates described an epidemic encompassing up to millions of people (e.g. Hombs and Snyder, 1986). This political relevance, also, engendered a sense of contention which spurred on a debate surrounding how to measure the homeless population in the US. These assertions pushed the US Federal government into action to acquire high quality counts of the homeless population. In 1983 and 1984, the Department of Housing and Urban Development began conducting point-in-time (PiT) surveys (Bobo, 1984). The first study was limited in nature and only sampled shelters in 60 areas (Burt and Cohen, 1989). This was later followed up by two major studies: (1) the US Department of Agriculture (USDA), which funded a study in 1987 to enumerate a national sample of homeless in the US, and (2) the prototype method for the systematic local-enumerations, based on stratified samples of microgreographies piloted in the city of Chicago. The USDA study attempted to count the homeless and record their demographics in 20 major cities across the US, and is considered the first nationally representative dataset for homelessness in the US (Poulin et al., 2008). The systematic local-enumeration method, pioneered in Chicago, was conducted and written up by Rossi et al. (1987) and greatly influenced the modern day methods for estimating the size and composition of the homeless population. Next, as part of its 1990 census, the US Census Bureau conducted an intensive study of five major cities in the US – collectively known as the S-Night sample (where "S" stands for both street and shelter) (Barrett et al., 1990). This experiment run by the US Census Bureau is the basis for the modern day HUD-collected PiT survey. Starting in 1999, Congress directed HUD to create a vehicle for national reporting of the homeless. This requirement would eventually lead to the creation of Continuum of Care (CoC) geographies/communities and the PiT counts conducted since 2005 (US Department of Housing and Urband Development Office of Community Planning and Development, 2009).

Overall, these efforts have resulted in publicly available datasets which support efforts to better understand how the homeless population is spatially distributed as well as how that distribution has evolved in the US since 2007. These data exist because of a Congressional mandate, and two key HUD initiatives: (1) Continuum of Care requirement for local organizations to organize for homeless care services, and (2) the required Point in Time count required of each CoC since 2007. Each CoC is required to do a biannual (ideally annual) Point in Time count during a night of the last 10 days in January, and must follow the federal definition of homelessness (see Section 4.3 for more details). Thus, there exist spatial/temporal estimates of the homeless population of the US. However, CoC's are currently defined by either a single city, a city and surrounding county, a region, or a state US Department of Housing and Urband Development Office of Community Planning and Development (2009). Of the CoC definitions, only "a single city" has a census analog, and even that analog often does not map to the Census spatial aggregates in a straightforward manner (see Almquist and Butts (2015)). The size and scope of CoCs is largely driven by the density of local organizations/care providers and thus tends not to map onto the common spatial units employed by social scientists, which are defined by the US Census Bureau according to political and demographic characteristics (for details on the US Census spatial aggregates see for Almquist (2010) for example).

This problem of lack of alignment between CoCs and US Census geographies limits the use and scope of the HUD PiT homeless counts and requires a solution to unify the two spatial aggregates. To this end we use modern statistical methods of imputation and small area estimation to disaggregate the PiT count data year by year to provide social science and policy researchers a unified spatial temporal dataset of the homeless population in the US which is connected to US County administrative units. We focus on developing a core dataset for the US Counties which is one of the most important units of study when considering more micro spatial level analysis of the US. For example, this is the common unit for understanding issues of internal migration (e.g., Molloy et al., 2011) or economic forecasts (e.g., Porter, 2003). Further, the county is the unit for which the most reliable economic data from the American Community Survey (ACS) is available Starsinic (2005). The ACS is a yearly economic and social survey conducted by the US Census Bureau every year since 2010.

In this paper we have developed a baseline county level dataset for the homeless PiT count from 2005 to 2017 produced by HUD. This includes all 3,143 counties in all 50 states and DC for all 13 years (11 years of observed data and 2 years of fully imputed data). Because the CoC requirement for reporting did not start until 2007, we have back imputed the years via a spatial temporal model fit to the 2007 to 2017 datasets. Finally, we have performed a smoothing spline analysis of variance framework for understanding both the raw and disaggregated data. The paper is laid out as follows: first we briefly discuss the motivation for county level homeless count data, and then we discuss recent research which employees the PiT data in academic research; then we discuss the limited literature on small area estimation of homeless populations in the US; next, we review the definition of Continuum of Care, PiT and US Census data; we then go over our procedure for connecting CoC data to US County data; following this section, we engage in spatiotemporal modeling of the resulting dataset and original datasets; and finally we end the paper with a discussion and example of the how this collection of data could be used. Included in this work is an R package with all spatial and homeless count data discussed in the paper and the geographic crosswalk of CoCs to counties for use by social science researchers and policy makers (`https://github.com/SSDALab/CoCHomeless`).

2 Motivation for Homeless Counts at the County Level

Governmental agencies, social scientists and others are often interested in comparing change in the homeless population with the US at various geographic levels (e.g., National, State, County, etc) with other measures known to correlate with homelessness (e.g., rents or measures of poverty, see e.g., Burt et al., 2007). One method for understanding such a change and its relationship to important spatial metrics is to work with a stable set of spatial geometries which are used by other Federal Agencies (e.g., The US Census Bureau, the Bureau of Labor Statistics, etc), non-profits (e.g., NORC, RTI, RAND) and private industry (e.g., Zillow, Redfin). For example, yearly measures of economic, demographic social information collected by the US Census can only be estimated down to the county level (US Census Bureau, 2017) – small areal estimates are possible in five-year aggregates for some measures, but these are known to have large standard errors (Council et al., 2007). Much of the housing information, e.g. the Zillow Rent Index (ZRI) used by Glynn and

Fox (2017) for example is only available nationally and at the county level. This work allows for researchers, non-profits and others to readily connect the CoC data these important social, economic, etc. indicators for understanding year-by-year change in the homeless population.

3 PiT Literature Review

While the PiT count is potentially under utilized, it has started to become an important tool in understanding and researching the issues of policies surrounding the homeless. Important key studies include: Burt et al. (2007) is possibly the first study to analyze the PiT count in academic setting, and demonstrates the usefulness of CoCs to replicate and extend the community-level model of homelessness developed by Lee et al. (2003). Fargo et al. (2013) use the PiT to demonstrate that demographic, economic and social factors (i.e., Community-level factors) accounted for 25-50% of the variance in homelessness rates in a multi-level model framework.Corinth (2017); Byrne et al. (2014) to show the permanent supportive housing decreases the homeless population over time. While this is review is not exhaustive, it covers the key cases where this data has been used so far and represents a set of cases where a county level analysis could also be useful.

3.1 Alternative Methods for Connecting CoC Data to Census Data

Byrne et al. (2013) provided a method for attaching US Census data also measured at the county level to CoC spatial aggregates (the inverse of this article). Byrne et al. (2013) matched CoCs to US county centroids and then allocated demographic and economic county data to the CoCs based on the following typology: (1) a CoC only matches one county, (2) a CoC matches multiple counties, (3) multiple CoCs match a single county. The authors treated (1) as a direct match (same as in this paper); in the case of (2), they aggregate up based on a population weighting scheme; and in (3) they aggregate the CoCs up to a single county level (which is same procedure we use). This method focuses on using the CoC as the level of analysis rather than the county. The advantage of such an approach is that it allows for the CoC to be the outcome variable of interest which is important for studying the CoC administrative unit; however, there are two main disadvantages of such a procedure. First, CoC's are not a standard spatial unit for most social science data collection and so connecting homeless counts as either an outcome or as a predictor or covariate is not readily available and thus requires sophisticated alterations to make them useful, i.e. one cannot just match on the federal information processing (FIPs) id. Second, there are many cases where the County is the appropriate unit of analysis rather than the CoC as counties are a stable administrative unit that researchers, policy makers and the general population have intuition, and for which government and non-government resources and policies are implemented.

3.2 City Level Spatial Models and Counts of Homeless

There exist a number of one off small area estimation case studies of homeless population in major US cities (e.g. Los Angeles; Berk et al. (2008)), as well as intensive surveys of the homeless in some major cities (e.g., US Census S-Night count). Berk et al. (2008) uses the pre-CoC mandated PiT count for LA in 2004 to build estimates of county, city and tract level homelessness for Los Angeles County. The researchers have sampled data for tracts at two waves and employ a random forest approach to modeling and aggregating tract sample data (totals are estimated using a Horvitz Thompson estimator). Others have used the US Census S-Night sample to understand the measurement of homeless populations and make small area estimates (e.g., Lee and Price-Spratlen, 2004; Bentley, 1995; Wright and Devine, 1992). There also exist a handful of longitudinal studies such as that conducted by Link et al. (1995) to look at five year prevalence rate homelessness. Though given the prevalence of the problem of homelessness there is surprisingly limited use of the nationwide homeless count data made available by HUD in academic research – we believe this is due to it not being linked to major social science datasets such as the US Census. Notable exceptions include Byrne et al. (2013) and subsequent follow up research on community factors impact on homelessness.

4 Continuum of Care and the Point in Time Count

In 1994, the US Department of Housing and Urban Development (HUD) began requiring each "community" to come to together to submit a comprehensive Continuum of Care (CoC) application rather than allowing applications from individual providers in a community. These local coalitions provide for stable administration for homeless services across the US and an incentive structure to centralize homelessness planning in specified areas. HUD has been working with local CoC agencies (CoC) to provide estimates of homeless persons in the nation since 2007. Each CoC adapts the standard approach required by HUD called homeless point-in-time (PIT) count to estimate the number of homeless individual in its jurisdiction.

This point-in-time count approach is conducted in such a way that staff and volunteers go to homeless shelters and unsheltered places (e.g., street, abandoned buildings) to count homeless persons under a list of specific requirements. Typically this is conducted on a single night during the last 10 days of January. For details and exceptions, please refer to the PIT Count Methodology Guide (U.S. Department of Housing and Urban Development, 2014) by HUD. It is required that in order to get the McKinney-Vento funding from HUD, CoCs have to conduct homeless PIT counts as part of their application at least biennially (ideally annually) and provide their best estimates of the number of homeless persons within their geographical areas. The PiT count must follow the federal definition of homeless,

> “[a]n individual or family living in a supervised publicly or privately operated shelter designated to provide temporary living arrangement (including congregate shelters, transitional housing, and hotels and motels paid for by charitable organizations or by federal, state, or local government programs for low-income individuals).” [24 CFR 578.3 of the Homeless Definition Final Rule]

In practice, the PiT count consist of sheltered homeless counts (e.g., the number of homeless persons who stay at emergency shelters, transitional housing projects, and so on) and unsheltered homeless counts (e.g., the number of homeless persons that sleep on streets and any other places that are not meant for human habitation). In addition to conduct homeless PIT counts, volunteers and staff are also trained to interview homeless persons in order to collect information that is more detailed on demographic characteristics. When a CoC covers a large geographic area, it becomes very costly or even unfeasible to interview every single homeless person encountered. In such case, a sampling approach is allowed to select subjects to be interviewed. In a complete dataset of homeless PIT counts provided by HUD each year, it consists of the total counts, followed by sheltered and unsheltered counts, and counts for various demographic characteristics such as gender, race, age group, veteran, chronic homeless, and so on. Although the homeless PIT counts are available since 2007, few studies have made use of such datasets to enhance the understanding of homelessness across the nation.

The PiT Count Methodology Guide provides instructions for different sampling strategies on minimizing biases in the results when interviewing homeless persons for demographic information. However, it does not cover details of sampling on a subpopulation of a certain characteristic (e.g., youth, veterans, females). The research by Golinelli et al. (2015) shows how sampling the homeless youth from few locations or site types (even sites with the highest homeless youth concentration) can result in biases in the demographic characteristics for the population estimates. When sampling is needed, it usually calls for case-by-case strategy due to the nature of human behaviors. The effect of their case study though is to raise the point that different sampling strategies need to be considered when obtaining a sample for homelessness studies.

4.1 Limitations of the Point in Time Count Data

The PiT count is the only major undertaking to enumerate the homeless population in the US on a yearly basis; however, the measure itself has some inherent limitations. For example, it is thought that up to 30 percent of the homeless migrate from colder areas (e.g., Chicago) to warmer areas (e.g., Los Angeles) during winter months (Burt, 2001). Given that the PiT is collected typically at the end of January – it is plausible that these data over estimate the homeless population warmer areas and underestimate the homeless population in colder areas – though it is worth pointing out that for resource allocation this may be an optimal strategy. Because the PiT is only collected at one point annually there is no way to measure seasonal variation within a given year, e.g. one might hypothesize that there are more homeless during the summer months than the winter months across the US, but the PiT would not be able to provide any insight into that question. The PiT does, however, provide a good measure of change over time because of its consistent measurement in a given year.

4.2 Limitations of Spatially-based Census/Survey Data

Two core methods for conducting the PiT count of unsheltered homeless people are endorsed by HUD: "*directly counting people in public places or screening those using selected services to determine whether they are homeless and without shelter* " (U.S. Department of Housing and Urban Development, 2014; Hopper et al., 2008). Hopper et al. (2008) notes that, "Counts of visibly homeless individuals miss unsheltered people who remain out of sight during the counts." Further, Hopper et al. (2008) conducted a survey based on a version of capture-recapture methods which suggested that standard PiT estimates could be underestimation the population by as much as 15-30%. This issue should have minimal impact on the PiT's ability to provide information on the temporal change of the homeless, e.g., if one assumes the underestimation is constant in time. This estimator could be further improved with standard regression methods (Lohr, 2009, pp. 429). Metraux et al. (2016) and others have also explored issues related to the PiT count focusing on the differences between the homeless population that uses social services and those that do not. Metraux et al. (2016) demonstrated that nonusers were typically harder to find and more likely to reside at the periphery of metropolitan areas.

4.3 The Point in Time Counts from 2005 to 2017 (2007 to 2017)

Starting in 2007, HUD required all CoC's to report PiT counts once every two years and provides PiT counts from 2007 to current date (2017) for public consumption at the CoC level (Housing and Urban Development, 2017). HUD began the formal process of designating the areal units (i.e., CoC administrative districts) in 2005 and thus provides CoC polygon data in ESRI shapefile format for years, 2005 to 2017. To provide a complete set of estimates (matching the spatial information provided by HUD) we generate complete imputations for the 2005 and 2006 CoC data based on the SSANOVA model and the spatial information provided by HUD. We do this for both the CoC spatial aggregates and the county aggregates. In the following subsection we discuss the CoC data made available by HUD.

4.4 Descriptive Statistics and Visualizations of CoC Data

CoC's represent HUD designated aggregates for facilitating provider funding and other outreach for the homeless. The CoC's provide the fundamental spatial unit and coverage of the PiT counts for the US. Overall, there are 41 different categories of homeless demographics available by 2017, but only 12 categories which can be longitudinally tracked over all 10 years reported by HUD. In the following sections we will describe how connect this data to lower level US Census data, i.e. counties and the necessary assumptions we had to make to perform this matching.

5 Overview of US Census Geography and Demographic Data

The US Census Bureau breaks down the US into a series of "core" geographies. The first is the nation as a whole; second is the states and territories (not discussed in this). States are further broken down by administrative units known as counties. Counties are then broken up into tracts, block groups and blocks where a block in an urban setting is about the size of city block (though one should not be confused by the naming convention, a Census block is not necessarily the size of a city block; Almquist (2010)). Because of the nature and size of most CoCs we will only attempt to disaggregate the CoCs into county level data. Further imputation into tract level might be possible with access to the raw PiT data if it is made available to the public. In Figure 1 we have plotted all the counties in the year 2010 alongside the CoC boundaries in 2010. One can see that sometimes the boundaries are perfectly aligned, while often CoCs represent distinct combinations of counties. In the following section we will discuss our methods for disaggregation and imputation of counties which have no coverage.

[Figure 1 about here.]

6 CoC to US Census Geographies

To begin our disaggregation process of the homeless PiT counts from the CoC level to the county level, we performed a spatial alignment with the counties and CoCs and either aggregated the CoCs (see Figure 3) which was a rare occurrence or disaggregated the counties (see Figure 2) which was the more common case.

We disaggregated by using the *simple population density weighting method* as we have no other information than the counts of homeless (and few other categories, e.g. sheltered or not sheltered) for the CoC geographies. The estimator is where we re-allocate the count between the counties that make up the CoC by their population density in 2010 ($D_i = \frac{population_i}{area_i}$)[1], i.e. $\hat{C}_i = \frac{D_i \cdot C_k}{\sum_i D_i}$, where $\hat{C}_i$ is our estimator of the count of homeless for county i in CoC k where homeless count C_k, is indexed on the number of counties contained within a given CoC, and D_i is the population density of any given spatial unit. [2] We use population density so as to allocate the homeless counts based on the evidence from Culhane et al. (1996), who demonstrates that homeless are more likely to reside in or near major cities. For example, the Metropolitan Denver CoC comprises seven counties where we see that a density based allocation scheme places almost 50% of the homeless in Denver county (as we would expect) as compared to an areal estimation which only allocates 3% to Denver county (see Table 1 for details). Further, this method satisfies the *pycnophylactic* or volume-preserving property, which requires the preservation of the initial data as is desired in this - note that we extend this method slightly by using a procedure to round a vector of real numbers to count data while preserving their sum as required by the partitioning algorithm. Other methods could be considered, but are hard to justify with the limited information available (see Chiang (2013) for a discussion of various alternative simple weighting metrics for spatial disaggregation). However, for about 4% to 10% of the US counties (on any given year) there exist no corresponding CoC. For these counties we imputed the data and this is discussed in the following section.

[Table 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

6.1 Multiple Imputation for Missing Cases

For counties which have no CoC information we employ methods of multiple imputation (MI) which are commonly used in the Social Sciences for managing the issue of missing data (Rubin, 1996). The basic logic of multiple imputation is the employment of a (conditional) probability model fit to the observed data and used to generate $m > 1$ complete datasets. One can then use all m datasets for analysis to account for the noise generated by the MI process or used to generate a stable imputation for the missing case (for further details see Rubin, 1996). Here, we employ a Bayesian version of a spatial Poisson generalized linear regression with a gaussian prior as described by Finley et al. (2015). We use only population counts and area of the county to model the total homeless population. By using a Bayesian approach we can sample

[1] Population is benchmarked on the most accurate period, i.e. the last census. Areas are measured in square km for each county.

[2] An alternative simple estimator is to re-allocate the count between the counties that make up the CoC by their individual area, i.e. $\hat{C}_i = \frac{A_i \cdot C_k}{\sum_i A_i}$, where $\hat{C}_i$ is our estimator of the count of homeless for county i in CoC k where homeless count C_k, is indexed on the number of counties contained within a given CoC, and A_l is the area of any given spatial unit. This estimate is also provided in the R package.

from the posterior for making making "predictions" on the missing counties. Our imputation algorithm was to fit all known county data (i.e., matched and disaggregated data) with the Bayesian spatial Poisson model and then checking standard MCMC diagnostics (see for a review Cowles and Carlin, 1996) and performing simple predictive checks. Once convergence was assessed, including for comparing spatial distribution of the fit data, see Figure 4 for example.

[Figure 4 about here.]

For the final estimate of counties with no mapping to a CoC we use the mean of the posterior predictive distribution of the fitted model state by state. Finally, we followed the same rounding procedure as in the areal disaggregation case. All code is made available via the accompanying R package which may be downloadable from `https://github.com/SSDALab/CoCHomeless` using `devtools` (Wickham and Chang, 2015) package.

6.1.1 Imputation of CoC 2005 and 2006

In Section 7 we introduce a nonparametric model for analyzing the variance of the data. We use this model to also back-cast the missing 2005 and 2006 homeless count data for which we have the CoC spatial information and nothing else. These details are discussed in Section 7 and made available in the R package.

6.1.2 Descriptives of County Level Data

Thus far we have mapped the CoC data from 2007 to 2017 onto the 2010 US County boundary files. We can now look at the total homeless population, the sheltered and unsheltered homeless populations by county.

Further, we may observe the county level characteristics by year for all the various homeless counts which are longitudinal (e.g., total homeless counts). In Table 2 we focus on the top 50 counties in this 11 year period (Table 2). We notice that Los Angeles is by far the dominant county with New York County and other Counties in the New York metro rounding out the core of the top 50 counties. Other notable counties are King County (Seattle), WA, San Diego County (San Diego City), CA, and Clark County (Las Vegas City), NV which round out the top 50 counties. In the next section we will compare a spatial temporal model of the county and CoC cases to better understand the limits of this disaggregation.

[Table 2 about here.]

7 Spatiotemporal Modeling

In the following section we will employ smoothing spline analysis of variance (SSANOVA) to understand the spatial temporal dimensions of homelessness in the US. Further we will also compare the CoC to County level results and discuss the limitations and advantages of employing one over the other in the discussion.

7.1 Model Overview

To model spatiotemporal patterns in the data, we use a smoothing spline analysis of variance approach, which is a nonparametric regression framework useful for discovering unknown trends in data (see Gu, 2013; Helwig et al., 2015; Helwig and Ma, 2015). The homeless counts are highly positively skewed, so we fit the model to the log10 transformed homeless counts. Specifically, we assume that

$$y_i = \eta_0 + \eta_s(\text{latitude}_i, \text{longitude}_i) + \eta_t(\text{year}_i) + \epsilon_i \tag{1}$$

where $y_i = \log_{10}(1 + n_i)$ with n_i denoting the number of homeless, η_0 is an unknown intercept term, $\eta_s(\cdot, \cdot)$ is the unknown spatial main effect function, which describes spatial trends in the homelessness counts as a function of the input $(\text{latitude}_i, \text{longitude}_i) \in \mathbb{R}^2$, $\eta_t(\cdot)$ is the unknown temporal main effect function, which describes temporal trends in the homelessness counts as a function of the input $\text{year}_i \in \{2007, \dots, 2017\}$, and $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$ is an unknown error term.

7.2 Fitting and Results: CoC Data

The model in Equation (1) was fit using the locations of all 403 unique CoCs as knots. The smoothing parameters were selected by minimizing the Generalized Cross-Validation (GCV) criterion (Wahba and Craven, 1978). The fit model explains about 92% of the variation in the log10 transformed counts (i.e., $\text{cor}(y_i, \hat{y}_i)^2 = 0.919$) and 90% of the variation in the counts (i.e., $\text{cor}(n_i, 10^{\hat{y}_i} - 1)^2 = 0.902$). This can be visualized in Figure 5, and includes the back casting of the two missing years.

[Figure 5 about here.]

7.3 Fitting Results: County Data

The model in Equation (1) was fit using the locations of 774 bin-sampled counties as knots. The smoothing parameters were selected by minimizing the GCV criterion (Wahba and Craven, 1978). The fit model explains about 71% of the variation in the log10 transformed counts (i.e., $\text{cor}(y_i, \hat{y}_i)^2 = 0.712$) and 79% of the variation in the counts (i.e., $\text{cor}(n_i, 10^{\hat{y}_i} - 1)^2 = 0.795$). This can be visualized in Figure 6, and includes the back casting of the two missing years.

[Figure 6 about here.]

7.4 Comparison of CoC and County Year by Year

We can dig into the SSANOVA results by year and compare the county level and CoC level year by year. The first thing to observe is the distribution or histogram plots (Figure 7). In general we find what we would expect which is slightly more dispersed data for the county than the CoC level, but relatively similar shape and peaks each year. Similarly, if we visualize the data as heat map, we see that county data is more dispersed (as we would expect by the disaggregation), but the basic trend and spikes are maintained (see Figure 8). Overall, it appears that the disaggregation of the CoCs to counties does what we would want which is to push the mass to the respective counties in as principled a manner as can be done with such little information. This data is also the basis for the back-casting imputation which we use to complete the CoC and county years.

[Figure 7 about here.]

[Figure 8 about here.]

8 Discussion and Future Directions

Here, we have introduced methods for connecting the US homeless PiT count for US Counties from 2005 to 2017; we have further provided the resulting datasets as an R package on github (`https://github.com/SSDALab/CoCHomeless`). To do this we have employed simple methods of spatial disaggregation (i.e., population density weighting), imputation (prediction from a spatial Bayesian GLM) and fit a SSANOVA to the data to explore the full spatial and temporal dimensions of the two datasets. We have also employed the SSANOVA to back cast the data for 2005 and 2006 which are not made available by HUD, because these periods occurred before the required reporting period for CoCs. If we look at the resulting mean visualizations of the CoC (Figure 9) and County (Figure 10) SSANOVA results we can see that the county formulation provides a much smoother spatial and temporal plot. We see that the spatiotemporal model works reasonable for the CoC data, but is limited because of sparseness of the spatial information. The spatial predictions in Figure 9 are very smooth because there is little to no data to work with in many of the spatial locations (particularly in the Midwest). The SSANOVA model is smoothing the data, and if there is simply no data to smooth it will just interpolate. Whereas, the spatiotemporal model works well for the county level data, where there are 31,430 (3,143 counties by 10 years) data points (and thousands of unique spatial locations). This produces a very detailed—less smooth—picture of the spatial trend, because there is much more information to work with here, see Figure 10. These results appear to be more useful from an SSANOVA perspective, as there is more smoothing and therefore less interpolation. Another key feature of the SSANOVA results on the county level data is we can see a two year lag from 2008 spike in the number of total homeless as we might expect following the 2008 financial collapse and resulting foreclosures. Further, the county results have the advantage that they can be directly linked with other publicly available county data such as the IRS Migration data (Molloy et al., 2011) or the US Census ACS and other economic surveys (United States Census Bureau, 2013). As Burt et al. (2007) and others have demonstrated, community level factors such as basic demographics, economic indicators, et cetra can provide a lot of insight into the changing nature of homelessness in the US. For example, Glynn and Fox (2017) and others demonstration that rental pricing can be used to track homeless changing is an important area of research and one that could be studied continuously over time – the practicality of such an ongoing study would be greatly simplified by this county level data. Further, we can ask the inverse question of the effect of homelessness on say county level GDP over time. Finally, this dataset could be used to suggest areas where resource allocation needs to be strengthened within a given state or region based on changes that are occurring at the county level. Altogether this dataset and its features are very rich and potentially highly useful for social science and policy workers who need county level homeless data.

[Figure 9 about here.]

[Figure 10 about here.]

9 References

Almquist, Z. W. (2010). US Census spatial and demographic data in R: The UScensus2000 suite of packages. *Journal of Statistical Software*, 37(6):1–31.

Almquist, Z. W. and Butts, C. T. (2015). Predicting regional self-identification from spatial network models. *Geographical analysis*, 47(1):50–72.

Barrett, D. F., Anolik, I., and Abramson, F. H. (1990). The 1990 census shelter and street night enumeration. *Atlanta*, 21134:2r161.

Bentley, D. (1995). *Measuring homelessness: A review of recent research.* Institute of Urban Studies.

Berk, R., Kriegler, B., Ylvisaker, D., et al. (2008). Counting the homeless in los angeles county. In *Probability and statistics: Essays in honor of David A. Freedman*, pages 127–141. Institute of Mathematical Statistics.

Bobo, B. (1984). A report to the secretary on the homeless and emergency shelters. *Washington, DC: US Department of Housing and Urban Development*.

Burt, M. R. (2001). Homeless families, singles, and others: Findings from the 1996 national survey of homeless assistance providers and clients. *Housing Policy Debate*, 12(4):737–780.

Burt, M. R. and Cohen, B. E. (1989). Differences among homeless single women, women with children, and single men. *Social problems*, 36(5):508–524.

Burt, M. R., Pearson, C., and Montgomery, A. E. (2007). Community-wide strategies for preventing homelessness: Recent evidence. *The Journal of Primary Prevention*, 28(3-4):213–228.

Byrne, T., Fargo, J. D., Montgomery, A. E., Munley, E., and Culhane, D. P. (2014). The relationship between community investment in permanent supportive housing and chronic homelessness. *Social Service Review*, 88(2):234–263.

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.

Chiang, A. (2013). Evaluating the performance of a filtered area weighting method in population estimation for public health studies.

Corinth, K. (2017). The impact of permanent supportive housing on homeless populations. *Journal of Housing Economics*, 35:69–84.

Council, N. R. et al. (2007). *Using the American community survey: Benefits and challenges*. National Academies Press.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Culhane, D. P., Lee, C.-M., and Wachter, S. M. (1996). Where the homeless come from: a study of the prior address distribution of families admitted to public shelters in new york city and philadelphia. *Housing Policy Debate*, 7(2):327–365.

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.

Finley, A., Banerjee, S., and Gelfand, A. (2015). spbayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software, Articles*, 63(13):1–28.

Glynn, C. and Fox, E. B. (2017). Dynamics of homelessness in urban america. *arXiv preprint arXiv:1707.09380.*

Golinelli, D., Tucker, J. S., Ryan, G. W., and Wenzel, S. L. (2015). Strategies for obtaining probability samples of homeless youth. *Field Methods*, 27(2):131–143.

Gu, C. (2013). *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media.

Helwig, N. E., Gao, Y., Wang, S., and Ma, P. (2015). Analyzing spatiotemporal trends in social media data via smoothing spline analysis of variance. *Spatial Statistics*, 14:491–504.

Helwig, N. E. and Ma, P. (2015). Fast and stable multiple smoothing parameter selection in smoothing spline analysis of variance models with large samples. *Journal of Computational and Graphical Statistics*, 24(3):715–732.

Hombs, M. E. and Snyder, M. (1986). *Homelessness in America: A forced march to nowhere.* Community for creative non-violence.

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.

Housing and Urban Development (2017). Pit and hic data since 2007.

Lee, B. A. and Price-Spratlen, T. (2004). The geography of homelessness in american communities: Concentration or dispersion? *City & Community*, 3(1):3–27.

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.

Link, B., Phelan, J., Bresnahan, M., Stueve, A., Moore, R., and Susser, E. (1995). Lifetime and five-year prevalence of homelessness in the united states: new evidence on an old debate. *American Journal of Orthopsychiatry*, 65(3):347.

Lohr, S. (2009). *Sampling: design and analysis.* Brooks/Cole: Cengage Learning, Boston, MA, 2nd edition.

Lucas, D. S. (2017). The impact of federal homelessness funding on homelessness. *Southern Economic Journal*, 84(2):548–576.

Metraux, S., Manjelievskaia, J., Treglia, D., Hoffman, R., Culhane, D. P., and Ku, B. S. (2016). Posthumously assessing a homeless population: Services use and characteristics. *Psychiatric Services*, 67(12):1334–1339.

Molloy, R., Smith, C. L., and Wozniak, A. (2011). Internal migration in the united states. *The Journal of Economic Perspectives*, 25(3):173–196.

Porter, M. (2003). The economic performance of regions. *Regional studies*, 37(6-7):549–578.

Poulin, S. R., Metraux, S., and Culhane, D. P. (2008). The history and future of homeless management information systems. *Homeless in America*, pages 171–179.

Rossi, P. H., Wright, J. D., Fisher, G. A., and Willis, G. (1987). The urban homeless: estimating composition and size. *Science*, 235(4794):1336–1341.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489.

Starsinic, M. (2005). American community survey: Improving reliability for small area estimates. In *Proceedings of the 2005 joint statistical meetings on CD-ROM*, pages 3592–3599. Citeseer.

United States Census Bureau (2013). 2007 – 2011 american community survey. Technical report, U.S. Census Bureau's American Community Survey Office.

US Census Bureau (2017). Income and poverty estimates guidance on when to use each survey. US Census Bureau Report.

U.S. Department of Housing and Urban Development (2014). Pit count methodology guide. Retrieved from https://www.hudexchange.info/resource/4036/point-in-time-count-methodology-guide/.

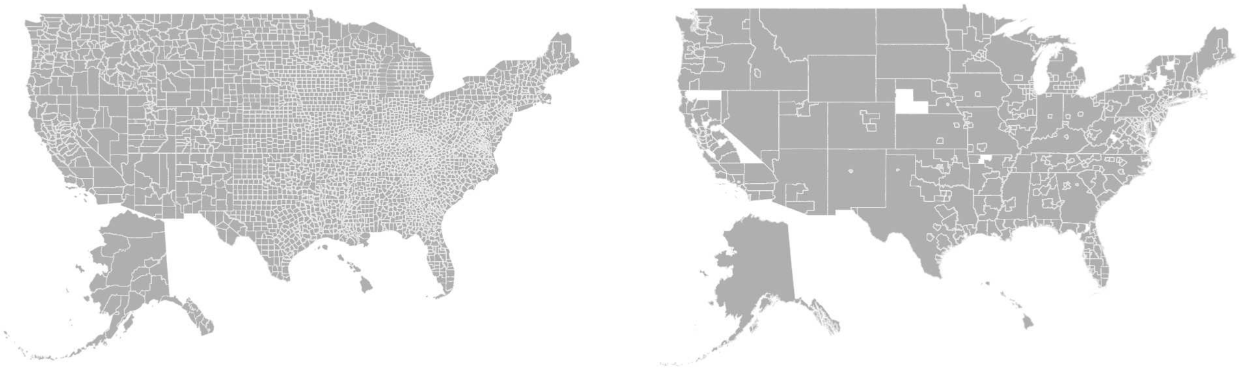
US Department of Housing and Urband Development Office of Community Planning and Development (2009). Hud's homeless assistance programs: Continuum of care 101. Technical report, US Department of Housing and Urband Development, DC.

Wahba, G. and Craven, P. (1978). Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–404.

Wickham, H. and Chang, W. (2015). devtools: Tools to make developing r code easier. *R package version*, 1(0).

Wright, J. D. and Devine, J. A. (1992). Counting the homeless: The Census Bureau's "s-night" in five us cities. *Evaluation review*, 16(4):355–364.

List of Figures



County Boundaries versus CoC Boundaries in 2010

US Census Counties, 2010

CoC, 2010

Figure 1: US Census counties and CoC boundaries in 2010.

When a CoC contains multiple counties, each of these counties points to the centroid (the red dot) of each polygon in that CoC.

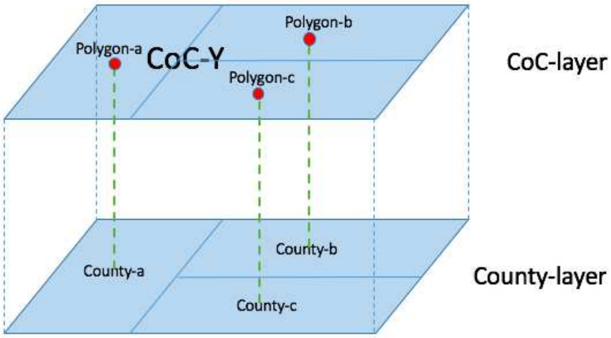


Figure 2: Example of layering CoC map over county map

When a county contains multiple CoCs, these CoCs all point to the centroid (the red dot) of that county.

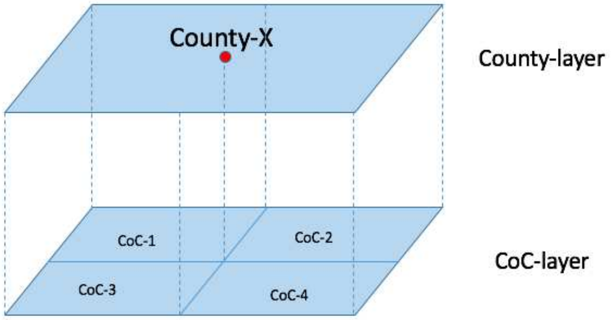
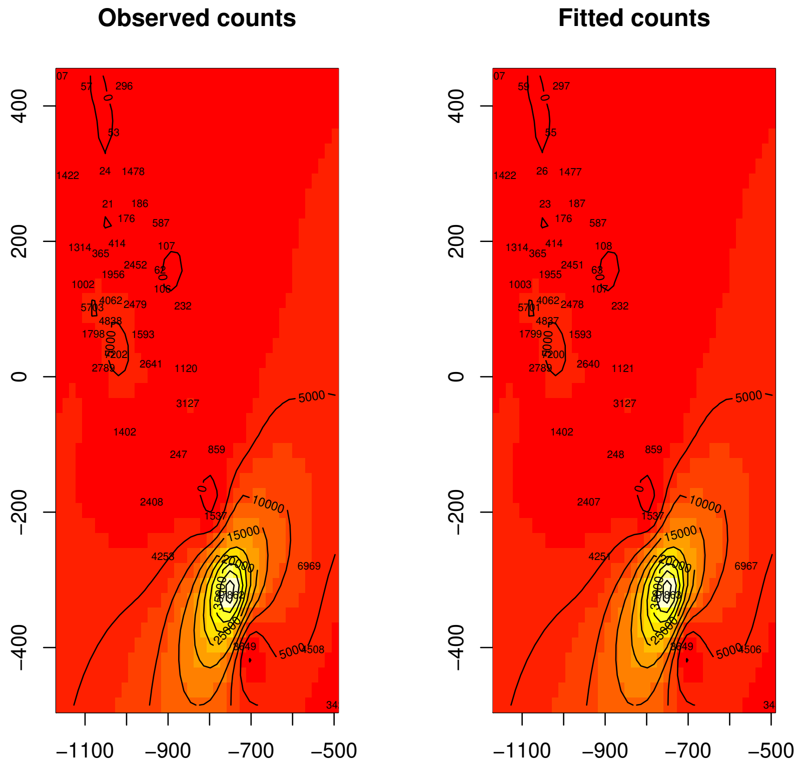


Figure 3: Example of layering county map over CoC map



Observed counts

Fitted counts

Figure 4: Spatially smoothed estimates of the fitted spatial bayesian glm compared to the predicted values for California.

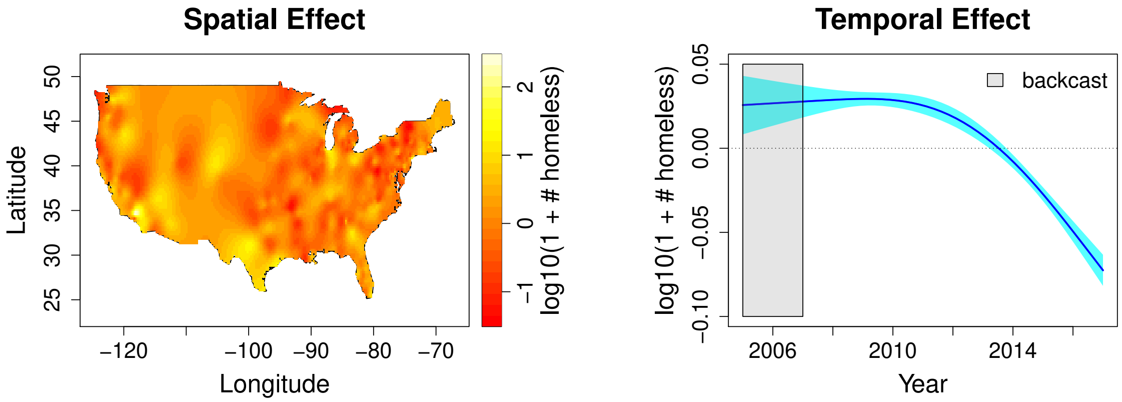


Figure 5: Estimates of η_s and η_t from the CoC data.

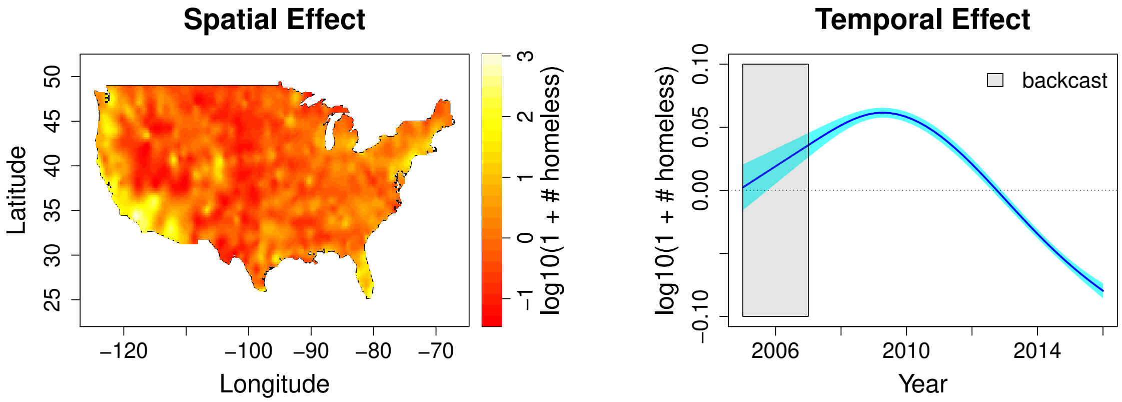


Figure 6: Estimates of η_s and η_t from the county level data.

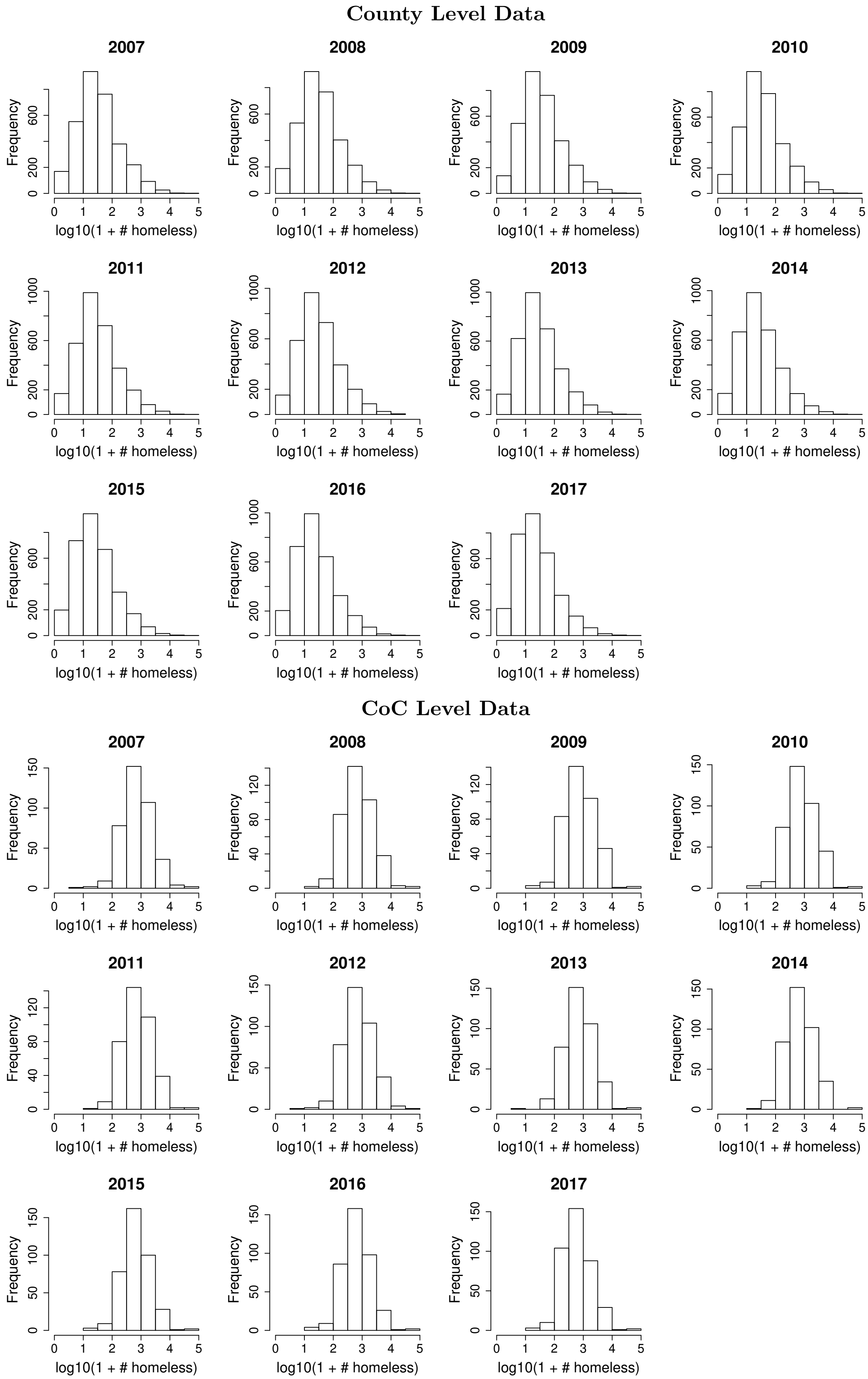


Figure 7: Histogram of the County and CoC data by year.

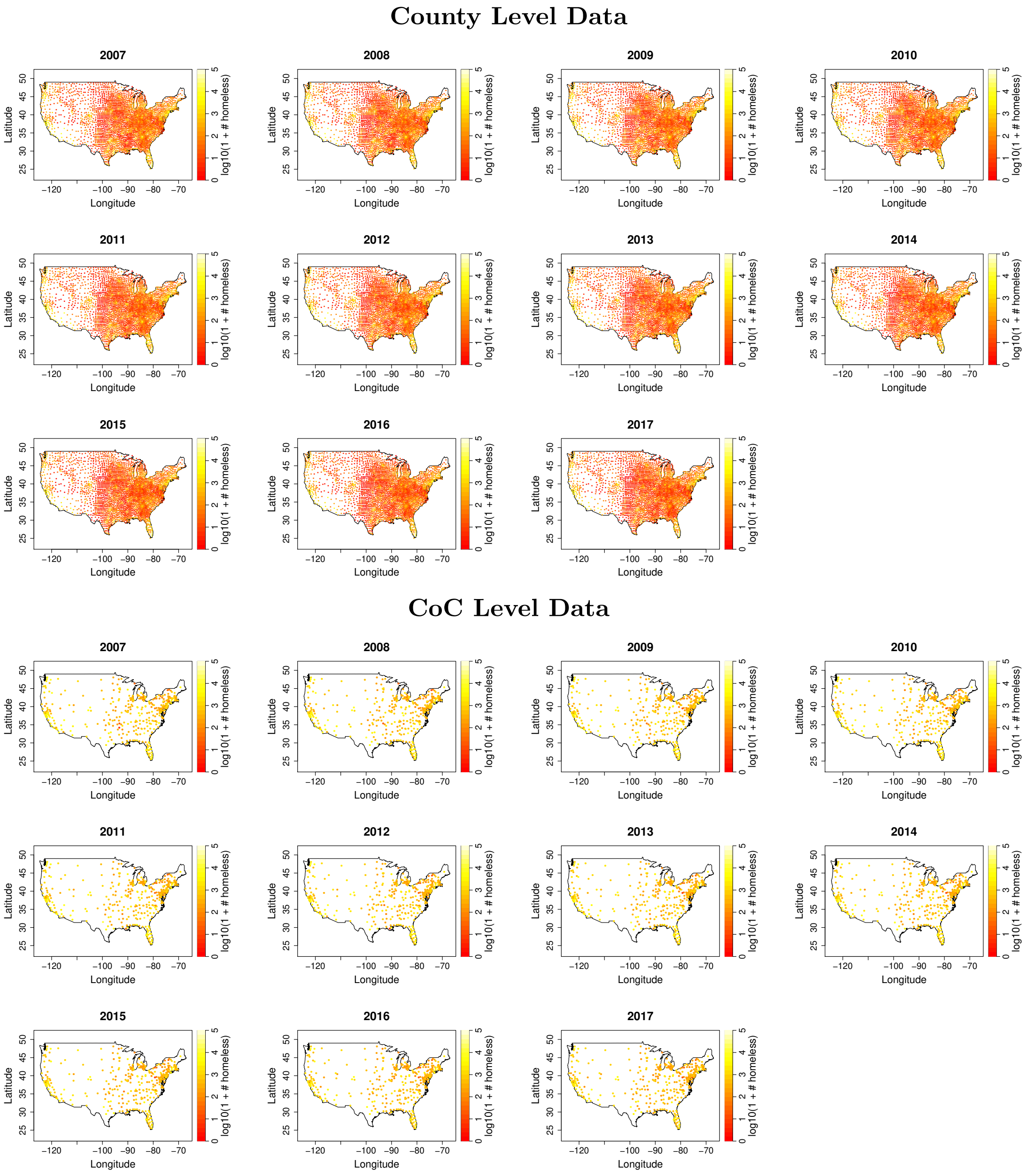


Figure 8: Heat map of the County and CoC data by year.

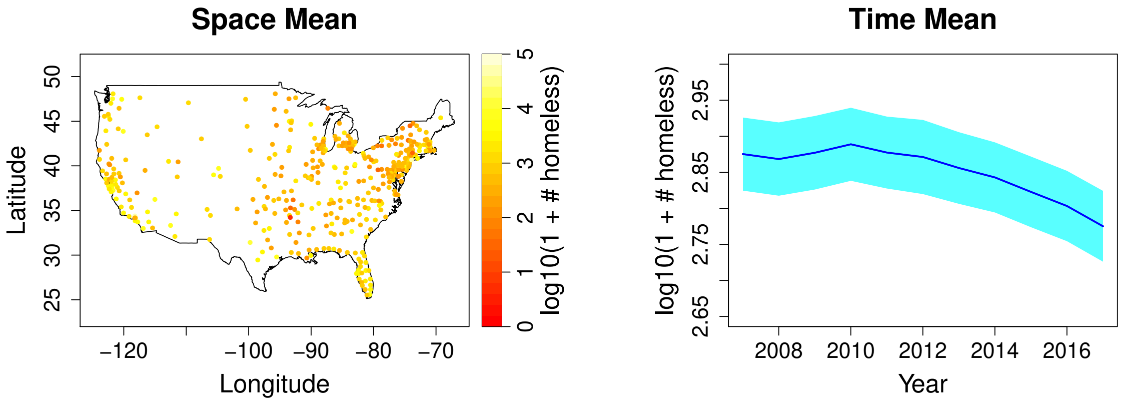


Figure 9: Mean of CoC data by space and time.

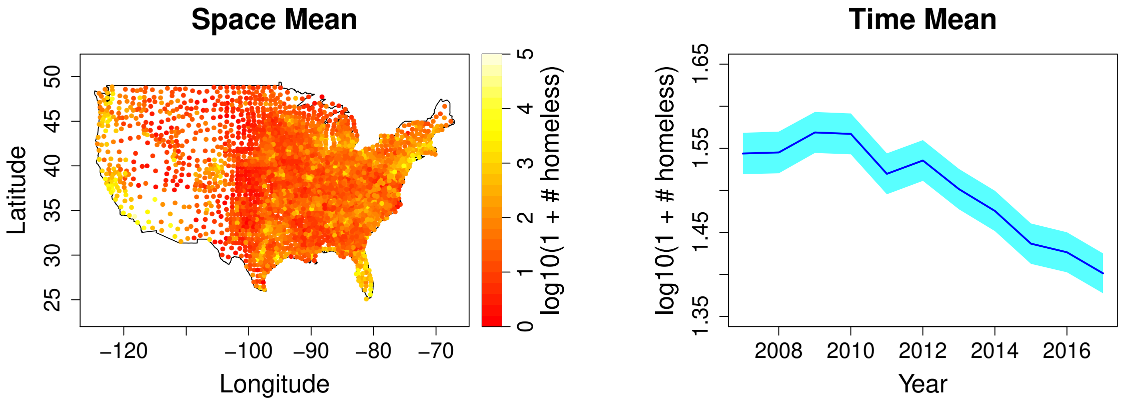


Figure 10: Mean of county level data by space and time.

List of Tables

Metropolitan Denver CoC mapping to Counties, 2007

Counties	Allocation Density	Allocation Area
Adams County	0.05	0.26
Jefferson County	0.09	0.17
Douglas County	0.04	0.19
Boulder County	0.05	0.16
Broomfield County	0.20	0.01
Denver County	0.48	0.03
Arapahoe County	0.09	0.18

Table 1: Comparison of Area allocation to Density allocation for Metropolitan Denver CoC in 2007.

Top 50 Counties, 2007-2017

	Total					Total			
Rank	Homeless	County	State	Year	Rank	Homeless	County	State	Year
1	55188	Los Angeles	California	2017	26	16136	Bronx	New York	2017
2	47862	Los Angeles	California	2007	27	15887	Bronx	New York	2015
3	47862	Los Angeles	California	2008	28	15507	Bronx	New York	2016
4	43854	Los Angeles	California	2016	29	15338	Kings	New York	2014
5	41174	Los Angeles	California	2015	30	14490	Kings	New York	2013
6	35524	Los Angeles	California	2013	31	14302	Bronx	New York	2014
7	34622	Los Angeles	California	2011	32	13512	Bronx	New York	2013
8	34393	Los Angeles	California	2014	33	12819	Kings	New York	2012
9	33243	Los Angeles	California	2009	34	12031	Kings	New York	2010
10	33243	Los Angeles	California	2010	35	11953	Bronx	New York	2012
11	31612	New York	New York	2017	36	11643	King	Washington	2017
12	31553	Los Angeles	California	2012	37	11564	Kings	New York	2011
13	31125	New York	New York	2015	38	11394	Kings	New York	2007
14	30381	New York	New York	2016	39	11369	Kings	New York	2008
15	28021	New York	New York	2014	40	11218	Bronx	New York	2010
16	26471	New York	New York	2013	41	11161	Kings	New York	2009
17	23418	New York	New York	2012	42	10783	Bronx	New York	2011
18	21978	New York	New York	2010	43	10730	King	Washington	2016
19	21125	New York	New York	2011	44	10624	Bronx	New York	2007
20	20815	New York	New York	2007	45	10601	Bronx	New York	2008
21	20769	New York	New York	2008	46	10407	Bronx	New York	2009
22	20390	New York	New York	2009	47	10122	King	Washington	2015
23	17304	Kings	New York	2017	48	10013	San Diego	California	2012
24	17038	Kings	New York	2015	49	9949	Clark	Nevada	2009
25	16631	Kings	New York	2016	50	9949	Clark	Nevada	2010

Table 2: Top fifty counties for total homeless from 2007 to 2017.