Mathematical Population Studies
An International Journal of Mathematical Demography

ISSN: 0889-8480 (Print) 1547-724X (Online) Journal homepage: https://www.tandfonline.com/loi/gmps20

Connecting Continuum of Care point-in-time homeless counts to United States Census areal units

Zack W. Almquist, Nathaniel E. Helwig & Yun You

MATHEMATICAL POPULATION STUDIES
2020, VOL. 27, NO. 1, 46–58
https://doi.org/10.1080/08898480.2019.1636574

Connecting Continuum of Care point-in-time homeless counts to United States Census areal units

Zack W. Almquist[a], Nathaniel E. Helwig[b], and Yun You[c]

[a]Department of Sociology and eScience Institute, University of Washington, Seattle, USA; [b]Department of Psychology and School of Statistics, University of Minnesota, Minneapolis, USA; [c]School of Statistics, University of Minnesota, Minneapolis, USA

ABSTRACT

In 2007, the Department of Housing and Urban Development initiated a point-in-time count of the homeless across the United States. The counts are administered by the Continuum of Care Program, which provides spatial and temporal data for the homeless population over the last decade. Unfortunately, this administrative spatial unit does not align with the more common areal units defined by the United States Census Bureau, which limits usability of these data. To unify these two areal units, spatial disaggregation, matching, and imputation allow for aligning Continuum of Care data with county data. The resulting county-level homeless counts for the years 2005 to 2017 are provided as an R package. The county-level data display more spatial precision and more temporal variation than the Continuum of Care-level data. Nonparametric regression analyses reveal that the spatiotemporal variation in the data can be well approximated by additive spatial and temporal effects at both the county and Continuum of Care level.

KEYWORDS
Homeless; homelessness; small area estimation; spatiotemporal modeling

1. Introduction

Counts of the United States homeless population are collected by the Department of Housing and Urban Development's Continuum of Care program. Continuum of Care units are defined by four core typologies: (1) a single city, (2) a city and surrounding county, (3) a region, (4) or a state (Department of Housing and Urban Development Office of Community Planning and Development, 2009). Of the four Continuum of Care definitions, only the first one ("a single city") has a precise United States Census Bureau areal unit analog, and even that analog often does not map to the United States Census Bureau produced spatial aggregates in a straightforward manner (Almquist and Butts, 2015; Almquist, 2018). The size and scope of Continuum of Care units are largely driven by the density of local organizations and care providers. Furthermore, the units tend to be shaped

CONTACT Zack W. Almquist zalmquis@uw.edu Department of Sociology and eScience Institute, University of Washington, Seattle, USA

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/gmps.

more by local politics than by United States Census Bureau administrative units (Almquist, 2010). This lack of alignment between Continuum of Care units and United States Census Bureau areal units limits the use and scope of the Housing and Urban Development point-in-time homeless counts.

Existing methods for connecting Continuum of Care data to data produced by the United States Census Bureau center around aggregating United States Census Bureau areal units up to the Continuum of Care spatial unit (the inverse of this article). Byrne et al. (2013) provided a method for attaching United States Census Bureau data, also measured at the county level, to Continuum of Care spatial aggregates. Byrne et al. (2013) matched Continuum of Care units to United States county centroids, and then allocated demographic and economic county data to the Continuum of Cares based on the following typology: (1) a Continuum of Care only matches one county, (2) a Continuum of Care matches multiple counties, (3) multiple Continuum of Care units match a single county. The authors treated (1) as a direct match (as in this article); in the case of (2), they aggregate up based on a population weighting scheme; and in (3) they aggregate the Continuum of Care units up to a single county level (as in this article).

The existing methods focus on using the Continuum of Care as the level of analysis rather than the county. Such methods are useful for studying the Continuum of Care administrative units, for example, to make resource allocation decisions. However, there are three main disadvantages of such a procedure. First, Continuum of Care units are not a standard spatial unit used for most social science data collection, so connecting homeless counts to other measures requires specialized data aggregation methods (Byrne et al., 2013). Second, the county units have reliable economic data from the American Community Survey (Starsinic, 2005), whereas the Continuum of Care units lack such information. Finally, unlike Continuum of Care units, the county is a common unit for understanding issues of internal migration (Molloy et al., 2011) and economic forecasts (Porter, 2003).

To unify the Continuum of Care and county data, we use modern statistical methods of imputation and small area estimation to disaggregate the point-in-time count data from the Continuum of Care level to the county level. We develop a county-level dataset for the homeless point-in-time count from 2005 to 2017 produced by the Department of Housing and Urban Development Office of Community Planning and Development (2009). This includes all 3,143 counties in all 50 states, as well as the District of Columbia (DC), for all 13 years (11 years of observed data and 2 years of fully imputed data). Because the Continuum of Care requirement for reporting did not start until 2007, we back impute data for the years 2005 and 2006 using a spatiotemporal model fit to the 2007 to 2017 data. The spatiotemporal modeling and imputation are conducted at both the Continuum of Care and county level, and the resulting spatiotemporal trends are compared.

2. Data and methods

2.1. *Data*

2.1.1. *Continuum of Care units*

In 1994, the United States Department of Housing and Urban Development began requiring each "community" to come together to submit a comprehensive Continuum of Care application rather than allowing applications from individual providers in a community. These local coalitions provide for stable administration for homeless services across the United States and an incentive structure to centralize homelessness planning in specified areas. Unfortunately, the Continuum of Care units do not align with United States counties. Figure 1 presents all the counties in the year 2010 alongside the Continuum of Care boundaries. Sometimes the boundaries are perfectly aligned, while often Continuum of Care units are comprised of distinct combinations of counties or overlapping segments of counties.

2.1.2. *Point-in-time counts*

The Department of Housing and Urban Development has been working with local Continuum of Care agencies to provide estimates of homeless persons in the nation since 2007. Each Continuum of Care adapts the "point-in-time count" procedure required by Housing and Urban Development to estimate the total number of homeless individuals in its jurisdiction. This point-in-time counting is conducted in such a way that interviewers go to homeless shelters and unsheltered places, such as streets or abandoned buildings, to count homeless persons. This is conducted on a single night during the last 10 days of January (US Department of Housing and Urban Development, 2014). To get the McKinney-Vento funding from Housing and Urban Development, Continuum of Care units have to conduct homeless point-in-time counts as part of their application at least biennially (ideally annually).

2.1.3. *Counting procedure*

The point-in-time 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 point-in-time count consists of sheltered homeless counts (such as the total number of homeless persons who stay at emergency shelters or transitional housing projects) and unsheltered homeless counts

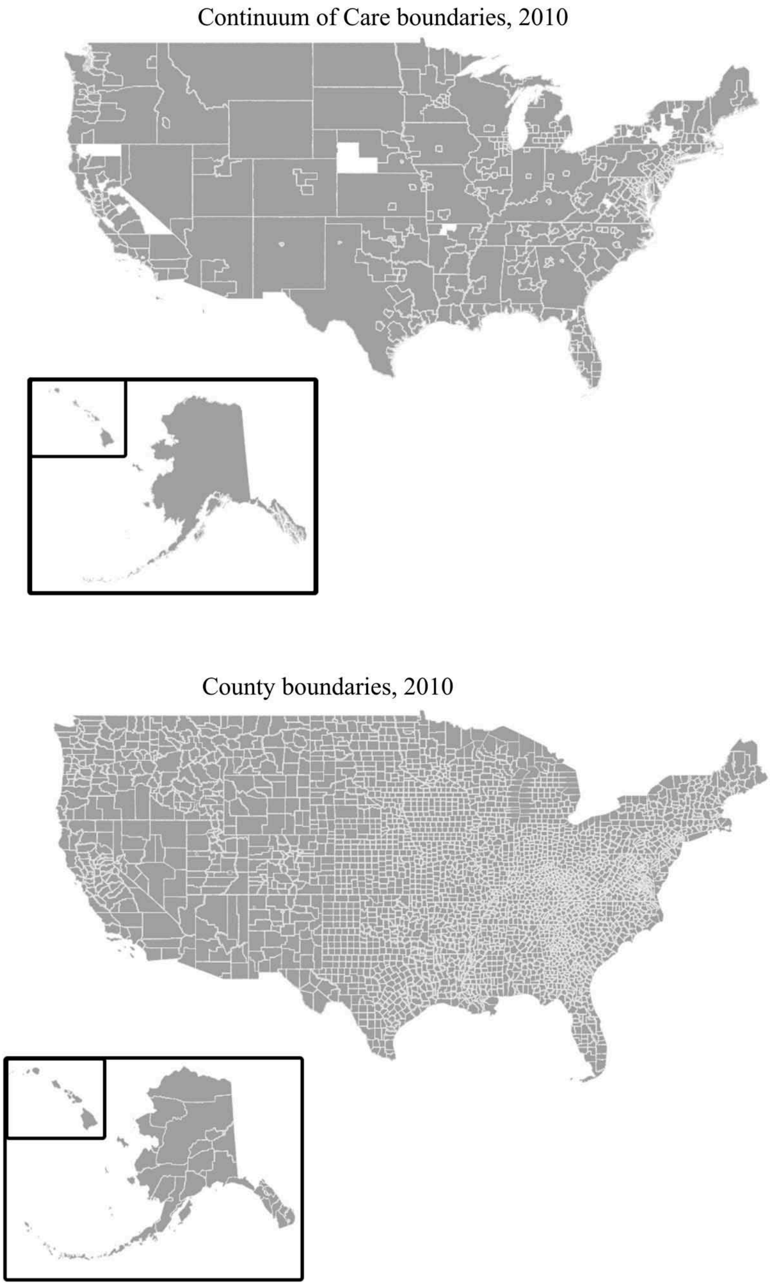

Figure 1. Continuum of Care boundaries and United States Census counties in 2010.

(such as the total number of homeless persons that sleep on streets and any other places that are not meant for human habitation). In addition to conducting homeless point-in-time counts, interviewers also collect information about sex, race, age group, veteran status, and chronic homeless status.

2.1.4. *Sampling procedures*

When a Continuum of Care covers a large geographic area, it becomes very costly or even unfeasible to interview every homeless person. In such cases, a sampling procedure is allowed to select subjects to be interviewed. The point-in-time count methodology guide (US Department of Housing and Urban Development, 2014) provides instructions for different sampling strategies for minimizing biases in the results when interviewing homeless persons. However, it does not cover youth, veterans, or women. Golinelli et al. (2015) showed 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 estimates. When sampling is needed, it usually calls for a case-by-case strategy due to human behaviors.

2.1.5. *Limitations*

Limitations of spatially based census or survey data collection often center around issues of timing (when the survey or census takes place) and location (city versus forest). For conducting the point-in-time count of unsheltered homeless people, the US Department of Housing and Urban Development (2014: 39) advises either "directly counting people in public places" or "screening those using selected services to determine whether they are homeless and without shelter." Both of these methods have limitations, which could affect the accuracy of the homeless counts.

For directly counting in public space, Hopper et al. (2008: 1438) argues that "counts of visibly homeless individuals miss unsheltered people who remain out of sight during the counts." Hopper et al. (2008) conducted a survey based on a version of capture-recapture methods, which suggests that standard point-in-time estimates could underestimate the population by as much as 15–30%. The effect should be minor on the ability of the method of the point-in-time survey to provide information on the temporal change of the homeless. A possibility is to assume that underestimation is constant in time. For the method of screening, Metraux et al. (2016) demonstrated that nonusers of social services were harder to find and more likely to reside at the periphery of metropolitan areas.

Other potential issues relate to inherent limitations of the point-in-time counting procedure. For example, it is thought that up to 30% of the homeless migrate from colder areas to warmer areas during winter months (Burt, 2001). Given that the point-in-time counts are collected at the end of January, it is plausible that these data overestimate the homeless population in warmer areas and underestimate the homeless population in colder areas. Though, this may be an optimal strategy for resource allocation. Finally, because the point-in-time counts are only collected once annually, there is no way to measure seasonal variation within a given year.

2.2. *Method*

2.2.1. *Spatial alignment*

We first performed a spatial alignment with the counties and Continuum of Care units. The spatial alignment involved either disaggregating the counties, which was the more common case, or aggregating the Continuum of Care units, which was a rare occurrence (Figure 2). We disaggregated by using the simple population density weighting method, as we have no other information than the counts of homeless for the Continuum of Care areal units. Each county within a Continuum of Care unit will be allocated a proportion of homeless individuals based on both the area and census of the population of a given county.

The estimator re-allocates the count between the counties that make up the Continuum of Care by their population density $D_i = \frac{\text{population}_i}{\text{area}_i}$. Population is benchmarked on the most accurate period, that is the 2010 census of the United States, and areas are measured in square km for each county. Then $\hat{C}_i = \frac{D_i \cdot C_k}{\sum_i D_i}$, where $\hat{C}_i$ is the estimator of the count of homeless for county i in Continuum of Care k with homeless count C_k, which is indexed on the number of counties contained within a given Continuum of Care. An alternative simple estimator is to re-allocate the count between the counties that make up the Continuum of Care by their individual area: $\hat{C}_i = \frac{\text{area}_i \cdot C_k}{\sum_i \text{area}_i}$. These methods are extended slightly by using a procedure to round a vector of real numbers to count data while preserving their sum. This is accomplished by rounding all values down, then rounding up the numbers with the highest fractional parts until the desired sum is reached.

We use the population density to allocate the homeless counts because Culhane et al. (1996) demonstrated that the homeless are more likely to reside in or near major urban areas. For example, the Metropolitan Denver Continuum of Care comprises seven counties with Denver County containing approximately 50% of the homeless population. In this case, a density-based

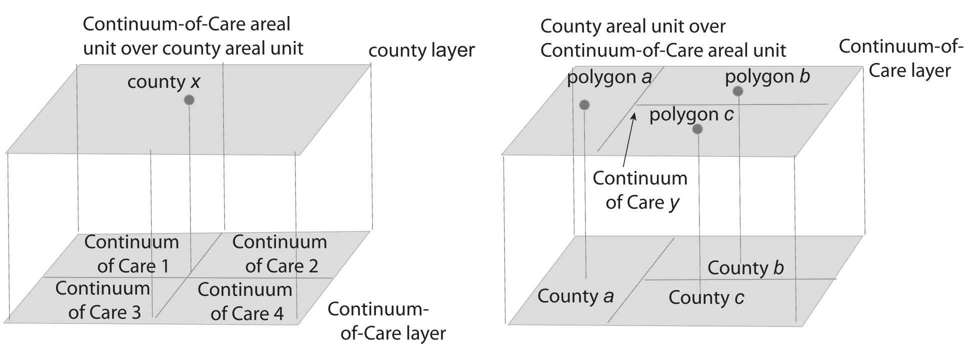

Figure 2. Examples of layering a county areal unit over multiple Continuum of Care areal units (left), and layering a Continuum of Care areal unit over multiple county areal units (right).

allocation scheme places almost 50% of the homeless in Denver County (as expected), whereas population- and area-based estimators only allocate 21% and 3% to Denver County. The population density method also satisfies the pycnophylactic or volume-preserving property, which requires the preservation of the initial data as is desired in this context (Tobler, 1979).

2.2.2. *Imputation of missing data*

For about 1% to 3% of the United States counties (on any given year), there exists no corresponding Continuum of Care. For these counties, we use multiple imputation (Rubin, 1996). Its logic is the employment of a (conditional) probability model fit to the recorded 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 multiple imputation process or used to generate a stable imputation for the missing case (Rubin, 1996). We use a Bayesian version of the 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. We sample from the posterior to make predictions on the missing counties.

Our imputation algorithm was to fit all known county data with the Bayesian spatial Poisson model. We then inspected Markov chain Monte Carlo diagnostics (Cowles and Carlin, 1996) and simple predictive checks. Convergence was assessed with Geweke and Gelmen-Rubin statistics (Finley et al., 2015), and we assessed the performance with predictive checks on the spatial distribution of counts (Figure 3). For the final estimate of counties with no mapping to a Continuum of Care, we use the mean of the posterior predictive distribution of the fitted model. Last, we followed the rounding procedure as used in the areal disaggregation case. All code is available with the accompanying R package, which is downloadable from https://github.com/SSDALab/CoCHomeless using devtools (Wickham and Chang, 2015).

2.2.3. *Spatiotemporal modeling*

The Department of Housing and Urban Development provides Continuum of Care polygon data in Esri shapefile format for the years 2005–2017. However, the Continuum of Care units were not required to start reporting point-in-time counts until 2007, and did so (at least) every two years from 2007 to 2017. To provide a complete set of estimates matching the spatial information provided by the Department of Housing and Urban Development, we apply spatiotemporal models to generate complete imputations for the 2005 and 2006 Continuum of Care data based on the spatial information provided by Housing and Urban Development. We do this for both the Continuum of Care spatial aggregates and the county aggregates.

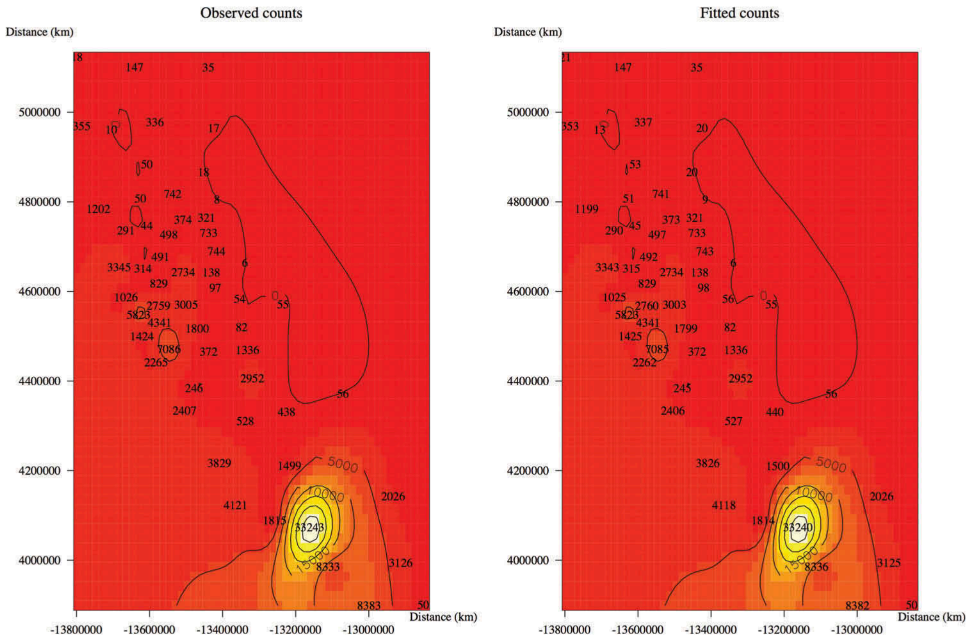

Figure 3. Spatially smoothed estimates of the fitted spatial Bayesian generalized linear model compared to the predicted values for California, 2010. Plotted in an equal area projection around the state centroid in kilometers.

To model spatiotemporal patterns in the data, we use a smoothing spline analysis of variance, which is a nonparametric regression analysis useful for discovering unknown trends in data (Wahba, 1990; Gu, 2013). Our application leverages recent computational advances that make the smoothing spline analysis of variance model feasible for analyzing large samples of spatiotemporal data (Helwig and Ma, 2015; Helwig et al., 2015; Helwig and Ma, 2016). The homeless counts are highly positively skewed, so we fit the model to the $\log_{10}$ transformed homeless counts. We assume that

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

where $y_i = \log_{10}(1 + n_i)$ with n_i denoting the total number of homeless, η_0 the intercept term, $\eta_s(\cdot,\cdot)$ the unknown spatial main effect, which describes spatial trends in the homelessness counts as a function of the input $(\text{latitude}_i, \text{longitude}_i) \in \mathbb{R}^2$, $\eta_t(\cdot)$ the unknown temporal main effect, which describes temporal trends in the homelessness counts as a function of the input $\text{year}_i \in \{2007, \ldots, 2017\}$, and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ an unknown error term. The models were fit using the bigsplines R package (Helwig, 2018).

3. Results

3.1. *Aggregation and imputation*

The county-level data provide more spatial precision compared to the Continuum of Care-level data (3,143 counties versus 403 Continuum of Care units), which allow for a more fine grain understanding of the spatial distribution of the homeless across the United States during the 2007 to 2017 period. A further advantage of the county-level results is that they can be directly linked with other publicly available county data, such as the United States Internal Revenue Service Migration data (Molloy et al., 2011) or the American Community Survey (United States Census Bureau, 2013). Figure 4 presents the resulting mean homeless counts for the Continuum of Care and County-level data. The "space mean" is calculated by averaging the data across the 11 years (2007 to 2017) separately for each location, whereas the "time mean" is calculated by averaging the data across the locations (Continuum of Care units or counties) separately for each year. Compared to the Continuum of Care-level data, the county-level data provides a more detailed spatial picture of the average homeless spatial distribution each year. The county-level data also reveal more temporal variation in the homeless population compared to the Continuum of Care-level data.

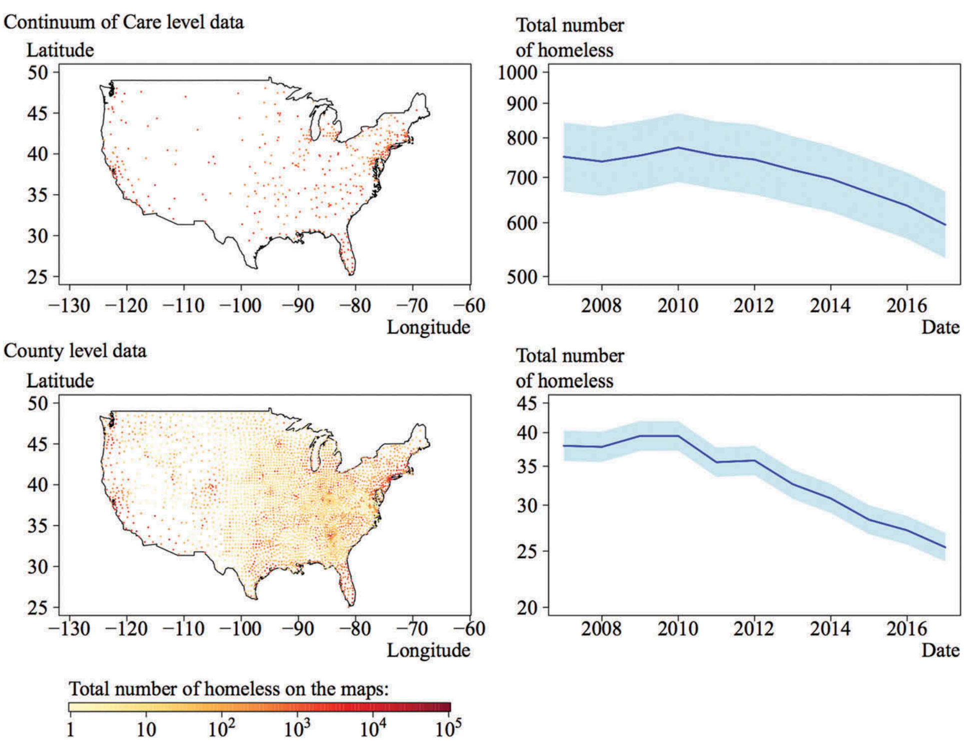

Figure 4. Mean of $\log_{10}$ data by space and time at the Continuum of Care and county.

3.2. *Space and time*

Equation (1) was fit to the Continuum of Care-level data using the locations of all 403 unique Continuum of Care units as knots. The smoothing parameters were selected by minimizing the generalized cross-validation criterion (Wahba and Craven, 1978). For the Continuum of Care data, the fit model explains 92% of the variation in the $\log_{10}$ transformed counts $(\mathrm{cor}(y_i, \hat{y}_i)^2 = 0.92)$ and 90% of the variation in the counts $(\mathrm{cor}(n_i, 10^{\hat{y}_i} - 1)^2 = 0.90)$. Eq. (1) was also fit to the county data using the locations of 774 randomly sampled counties as knots. As with the Continuum of Care model, the smoothing parameters were selected by minimizing the generalized cross-validation criterion. For the county-level data, the fit model explains 70% of the variation in the $\log_{10}$ transformed counts and 78% of the variation in the counts. Figure 5 shows the estimated spatial and temporal effect functions, as well as the backcasting results for the two missing years.

The smoothing spline analysis of variance results in Figure 5 shows that the Continuum of Care formulation provides a much smoother spatial and temporal plot. Eq. (1) works reasonably for the Continuum of Care data, but is limited because of the sparseness of the spatial information. The spatial

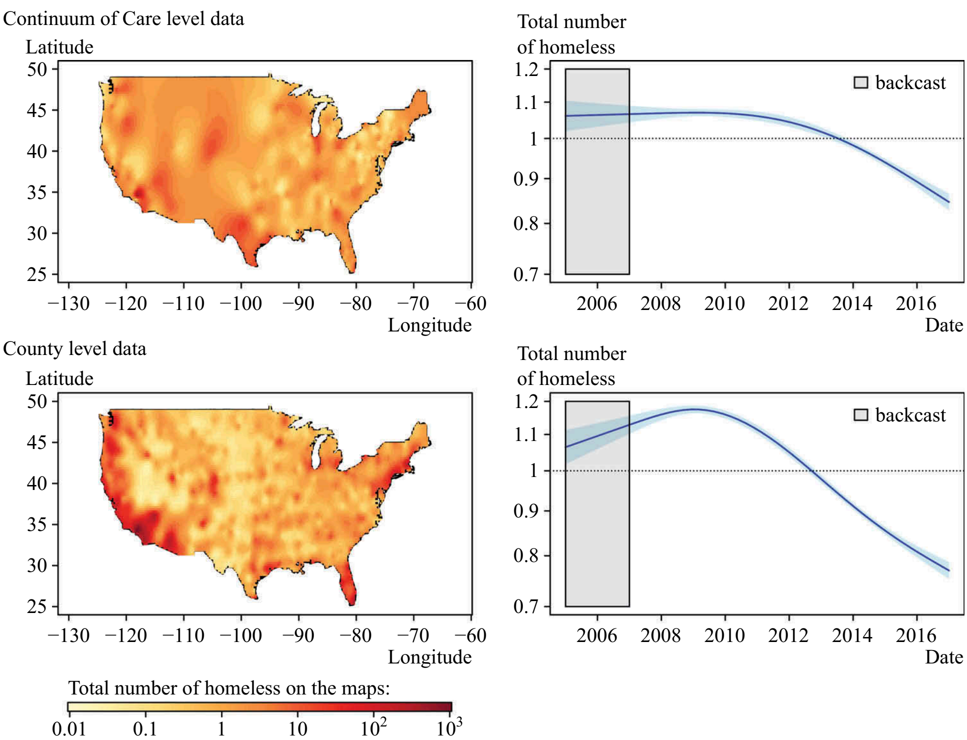

Figure 5. Estimates of spatial η_s and temporal η_t effects from the Continuum of Care and the county data. Gray shading denotes the backcast trend for the years 2005–2006.

predictions for the Continuum of Care-level data in Figure 5 are very smooth because there is little to no data to work with in many of the spatial locations (particularly in the Midwest). The smoothing spline analysis of variance model of Eq. (1) is averaging the data, and if there is simply no data to average it will interpolate, which is occurring throughout many areas in the Midwest.

In contrast, the spatiotemporal model works well for the county-level data, where there are 34,573 data points (3,143 counties by 11 years), and, thus, thousands of unique spatial locations. This produces a very detailed and less smooth picture of the spatial trend, because there is much more information to work with here (Figure 5, bottom). The county results are more useful for spatial analysis, as there is more smoothing and therefore less interpolation. Another key feature of the county-level results is that a 2-year lag from the 2008 spike in the total number of homeless is visible, as expected following the 2008 financial collapse and resulting foreclosures.

4. Conclusion

We introduced methods for connecting United States homeless point-in-time count data to United States county-level data for the years 2005–2017. We used methods of spatial disaggregation (population density weighting), multiple imputation (prediction from a spatial Bayesian generalized linear model), and nonparametric spatiotemporal modeling (smoothing spline analysis of variance) to align county and Continuum of Care data. We backcast the data for 2005 and 2006, which are not made available by the Department of Housing and Urban Development. The resulting datasets are available in R at https://github.com/SSDALab/CoCHomeless. Our work reveals that the county-level dataset provides more detailed insights about spatiotemporal trends in the homeless population, such as the increase in the homeless population following the 2008 financial collapse. The county-level dataset also has the benefit of being easily connected to United States Census Bureau data such as economic or housing information.

We have argued for the use of a density-based estimator, given the limited availability of data. However, this may not fully capture the heterogeneity of the homeless population due to issues like public policy (a city or county seat with more homeless shelters), local philanthropy, and transportation centers (Greyhound bus stops, which is the only cross United States bus line, where homeless can obtain free tickets to alternative locations). Using the Continuum of Care as the unit of analysis means that the homeless counts are being measured without error, or, more appropriately, without any additional error beyond that attributable to the method employed by local Continuum of Care units in running their point-in-time counts. Disaggregating the point-in-time counts from the Continuum of Care to the county level introduces some additional error. Aggregating from the county to the Continuum of Care level introduces error into the county-level measures from other sources, like the United States Census Bureau data. Thus, the decision about whether to aggregate from the county to Continuum of Care, or disaggregate from the Continuum of Care to the county, should be given careful consideration.

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. (2018). Large-scale spatial network models: an application to modeling information diffusion through the homeless population of San Francisco. *Environment and Planning B: Urban Analytics and City Science*, doi:10.1177/2399808318785375

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

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.

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

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.

Department of Housing and Urban Development Office of Community Planning and Development. (2009). *HUD's Homeless Assistance Programs: Continuum of Care 101*. Washington, DC: US Department of Housing and Urban Development.

Finley, A., Banerjee, S., and Gelfand, A. (2015). spbayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, *63*(13): 1–28. https://www.jstatsoft.org/v063/i13

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

Gu, C. (2013). *Smoothing Spline ANOVA Models, Vol. 297 of Springer Series in Statistics*. New York, NY: Springer.

Helwig, N. E. (2018). *Bigsplines: smoothing splines for large samples*. R package version 1.1-1. http://CRAN.R-project.org/package=bigsplines

Helwig, N. E., Gao, Y., Wang, S., et al. (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.

Helwig, N. E. and Ma, P. (2016). Smoothing spline ANOVA for super-large samples: scalable computation via rounding parameters. *Statistics and Its Interface*, *9*: 433–444.

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

Metraux, S., Manjelievskaia, J., Treglia, D., et al. (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): 549–578.

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*. Minneapolis, MN: Citeseer, 3592–3599.

Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, *74*(367): 519–530.

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

US 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/.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

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*.