Imputing Missing 2021 Unsheltered Homelessness Counts: A Two-Phase Machine Learning Error Correction Approach

Luke Maddock*

June 9, 2026

Abstract

The 2021 Point-in-Time count of unsheltered homelessness in the United States was severely disrupted by the COVID-19 pandemic, with 61.6 percent of Continuums of Care failing to conduct a complete unsheltered enumeration. This paper develops a two-phase machine learning framework to impute the missing counts. Phase 1 trains an XGBoost model on a 2015–2019 panel of community-level predictors—housing markets, economic conditions, demographics, climate, and shelter infrastructure—to estimate baseline 2021 counts under pre-pandemic conditions, validated against a locked 2020 holdout. Phase 2 models the COVID-specific deviation from baseline using the 146 CoCs that did conduct complete counts, with overlap-based propensity score weighting to address selection bias in the responding sample. The two phases combine to produce imputed counts for 230 non-counting CoCs, yielding a national unsheltered estimate of 195,191 individuals (90% prediction interval: [114,380, 255,978]). Emergency relief funding—particularly ESG-CARES allocations—dominates the COVID adjustment model, accounting for 57.6 percent of its explanatory power and suggesting that targeted shelter funding buffered unsheltered homelessness during the crisis. The imputed estimates restore continuity to a national dataset essential for longitudinal policy analysis, and the two-phase decomposition offers a generalizable template for imputing missing administrative data when exogenous shocks simultaneously disrupt data collection and alter the quantity being measured.

Keywords: homelessness, Point-in-Time count, missing data imputation, machine learning, COVID-19, propensity score weighting, unsheltered homelessness

JEL Codes: C53, C81, I38, R23

*Department of Economics, Colorado State University. I thank Anita Pena, Tim Komarek, Ray Miller, and Jesse Burkhardt for guidance throughout this project, and Levi Altringer and Sophie McKee for helpful comments. All errors are my own. Correspondence: luke.maddock@colostate.edu.

1 Introduction

Every January, the U.S. Department of Housing and Urban Development (HUD) requires each of the nation's Continuums of Care (CoCs) to conduct a Point-in-Time (PIT) count of all people experiencing homelessness: both those residing in emergency shelters and transitional housing, as well as those sleeping in unsheltered locations such as streets, parks, and encampments. These counts form the backbone of federal homelessness policy: they produce the primary national estimate of homelessness in the United States, help inform billions of dollars in annual funding allocations through HUD's Notice of Funding Opportunity (NOFO) competition, and provide the primary data source for longitudinal research on homelessness trends (O'Flaherty, 2019; Henry et al., 2022; de Sousa et al., 2023).

In January 2021, this system broke down. As the COVID-19 pandemic entered its second year, HUD encouraged CoCs to assess whether conducting unsheltered street counts posed unacceptable public health risks (Henry et al., 2022). The response was overwhelming: 61.6 percent of CoCs (236 out of 385) either conducted only a sheltered count or performed merely a partial unsheltered enumeration. The remaining 149 CoCs that did carry out complete counts were not representative of the national landscape, skewing toward smaller, less complex communities with substantially lower baseline unsheltered populations (Henry et al., 2022). The result was the largest single-year data gap in the history of the PIT count.

The consequences of this gap are far-reaching. First, no reliable national estimate of unsheltered, and therefore total, homelessness exists for 2021. The Annual Homeless Assessment Report (AHAR) to Congress, typically the definitive accounting of homelessness in the United States, was forced to restrict its 2021 findings entirely to sheltered populations (Henry et al., 2022). This means that the federal government's own assessment of homelessness during the most significant public health and economic crisis in a generation is fundamentally incomplete, as seen visually in the trend graph present in the HUD 2022 CoC Performance Profile Report[1] (Figure 1).

[1] de Sousa et al. (2023)

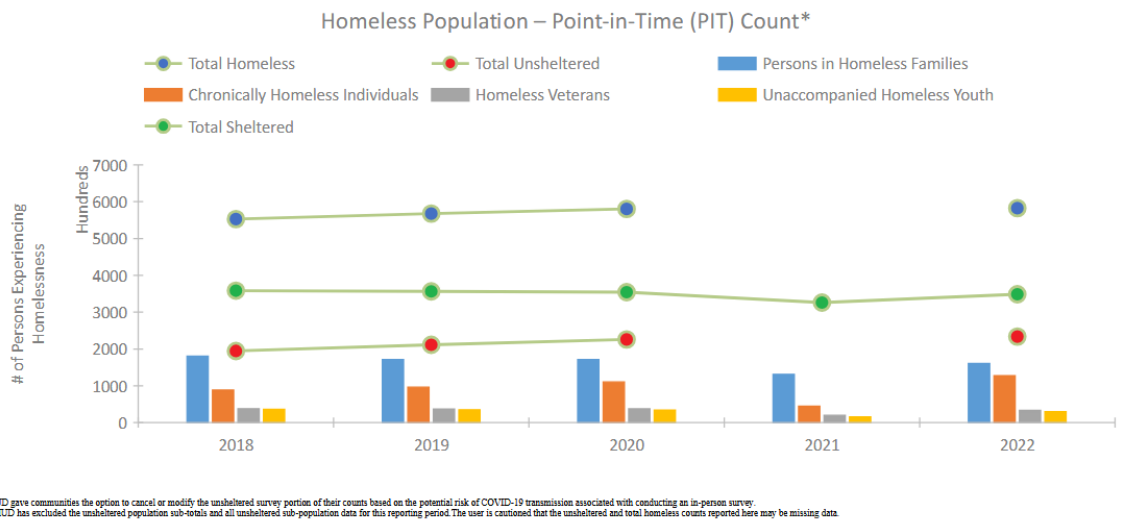

Figure 1: Missing Counts in the HUD 2022 CoC Performance Profile

Second, the missing data severely constrains researchers' ability to study how the pandemic affected homelessness. The COVID-19 period produced an unprecedented combination of economic shocks, such as mass unemployment, business closures, and housing instability. Alongside those shocks were equally unprecedented policy responses, including eviction moratoria, emergency rental assistance, CARES Act relief, and expanded shelter capacity. Whether these countervailing forces produced a net increase or decrease in unsheltered homelessness, and how those effects varied across communities, remains an open empirical question that cannot be answered with the existing data.

Third, the data gap had direct implications for federal funding. HUD's annual NOFO competition allocates approximately $2.7 billion to CoCs, with scoring criteria that historically incorporated PIT count data to assess community need and system performance (U.S. Department of Housing and Urban Development, 2021; de Sousa et al., 2023). In the absence of reliable unsheltered counts, HUD was forced to modify its evaluation framework for the FY 2021 competition, explicitly noting that "most communities could not conduct an unsheltered count in 2021 that is comparable to previous counts" and that it would evaluate only sheltered data for that year's funding decisions (U.S. Department of Housing and Urban Development, 2021). Whether this methodological shift altered the relative priority

rankings of CoCs and the resulting distribution of federal resources is an empirical question with significant policy implications.

This paper addresses the 2021 data gap by developing a two-phase imputation framework that reconstructs missing unsheltered counts while explicitly accounting for the pandemic's disruption to homelessness dynamics. The core methodological challenge is straightforward: one cannot simply predict missing 2021 counts using historical relationships between community characteristics and unsheltered homelessness. The COVID-19 pandemic fundamentally altered those relationships through simultaneous, countervailing channels: economic dislocation pushed individuals toward homelessness while emergency policy responses pulled them away from it. A model trained solely on pre-pandemic data would produce estimates reflecting a world in which COVID-19 never occurred, systematically mischaracterizing what actually happened in 2021.

The framework proceeds in two phases. In the first phase, I train multiple predictive models on a panel of CoC-level data spanning 2015 through 2019, drawing on established determinants of unsheltered homelessness, including housing market conditions, economic indicators, demographic composition, climate, and homeless services infrastructure, to generate baseline predictions of what 2021 unsheltered counts would have been under pre-pandemic conditions. Out-of-sample accuracy is assessed on a locked 2020 temporal holdout before the selected model is retrained on the full panel and used to produce 2021 baseline predictions. In the second phase, I leverage the 149 CoCs that did conduct complete unsheltered counts in 2021 to model the discrepancy between these baseline predictions and observed reality. This residual captures the net effect of COVID-19 on unsheltered homelessness, and I model it as a function of pandemic-specific factors: COVID-19 severity, emergency relief funding, and local policy responses such as eviction moratoria. Because the CoCs that counted differ systematically from those that did not, I apply overlap-based propensity score weighting to address selection bias before extrapolating the estimated COVID adjustment to non-counting CoCs. The final imputed values reflect both the long-run structural deter-

minants of unsheltered homelessness and the unique conditions of the pandemic period, with uncertainty quantified through end-to-end clustered bootstrap prediction intervals.

This paper makes two contributions. First, it produces a complete set of 2021 unsheltered homelessness estimates for all CoCs in the United States with sufficient feature variable data available, accompanied by prediction intervals that quantify the uncertainty inherent in imputed values. These estimates restore the continuity of a national dataset essential for longitudinal policy analysis and enable, for the first time, a comprehensive accounting of unsheltered homelessness during the COVID-19 pandemic. Second, the paper demonstrates a generalizable methodological framework for imputing missing administrative data when standard assumptions of random missingness are violated by exogenous shocks. The two-phase residual correction approach of establishing a counterfactual baseline and then modeling the shock-specific deviation offers a template applicable beyond homelessness to any domain where crisis events simultaneously disrupt data collection and alter the quantity being measured.

The remainder of the paper proceeds as follows. Section 2 reviews the institutional background of PIT counts, the 2021 data gap, and the relevant literature on homelessness determinants, machine learning in social policy, and missing data imputation. Section 3 describes the data sources, sample construction, and the two-phase modeling and uncertainty quantification methodology. Section 4 presents results for both phases and the final national estimates. Section 5 discusses findings, methodological contributions, limitations, and directions for future research.

2 Background

2.1 Homelessness Data in the United States

The systematic measurement of homelessness in the United States is a relatively recent undertaking. Congress first directed the Department of Housing and Urban Development to

develop an unduplicated count of the homeless population in the fiscal year 2001 appropriations act, though the first Annual Homeless Assessment Report was not produced until 2007, and the quality of its early data was admittedly limited (O'Flaherty, 2019). The AHAR has been produced annually since then under contracts with Abt Associates and the University of Pennsylvania, and it remains the most widely cited source of national homelessness statistics (Meyer et al., 2021).

The AHAR draws on three data components: a Point-in-Time count of sheltered and unsheltered homeless individuals conducted on a single night in late January, information on the characteristics of sheltered individuals over the course of the year, and a Housing Inventory Count cataloging the beds and units available to the homeless population (O'Flaherty, 2019; Meyer et al., 2021). Responsibility for collecting and reporting these data falls to the nation's approximately 385 Continuums of Care—administrative units that receive HUD funding for homeless programs and that vary widely in geographic scope, encompassing individual cities, counties, parts of states, or entire states (O'Flaherty, 2019). CoCs are required to report annually but are only mandated to count unsheltered individuals in odd-numbered years, though many conduct unsheltered counts every year.

The PIT count, while indispensable as the only national enumeration of both sheltered and unsheltered homelessness, carries well-documented limitations. The unsheltered count in particular relies on loosely supervised volunteers who canvas streets, parks, and other public spaces on a single January night (O'Flaherty, 2019). The diligence, training, and consistency of these volunteers vary considerably across CoCs, and the methodology systematically excludes individuals in locations that volunteers cannot access—restaurants, parking garages, stairwells, and other private property (O'Flaherty, 2019). People experiencing homelessness are also exceptionally difficult to enumerate for reasons that extend beyond methodology: poor mental health, substance use, lack of a fixed location, and active avoidance of being found all contribute to undercounting (Meyer et al., 2021; Glasser et al., 2014). These difficulties are particularly acute for the unsheltered population, where no institutional point of

contact—such as a shelter check-in—exists to facilitate enumeration.

The limitations of PIT data have important implications for research. As O'Flaherty (2019) notes, when PIT counts serve as dependent variables in regression analyses, two concerns arise: heteroskedasticity, if measurement error is systematically larger in some CoCs than others, and correlation between errors and policies of interest, if CoCs with particular characteristics tend to systematically overcount or undercount. Despite these shortcomings, the PIT count remains the only data source that captures unsheltered homelessness at a national scale, and it is the primary input into HUD's funding allocation decisions. Other national data sources—such as the Census Bureau's Special Report on the Emergency and Transitional Shelter Population (Smith et al., 2012) or the 1996 National Survey of Homeless Assistance Providers and Clients (Burt et al., 1999)—either exclude the unsheltered population entirely or are decades out of date. Localized studies drawing on shelter administrative records or CoC-level surveys offer richer detail but cannot be generalized nationally, given the substantial geographic heterogeneity in homelessness trends, housing markets, and shelter capacity across the country (Meyer et al., 2021).

2.2 The 2021 Data Gap

In January 2021, the COVID-19 pandemic was entering a severe winter wave. Vaccines had received emergency use authorization only weeks earlier and were not yet widely available to the general public, and communities across the country remained under varying degrees of public health restrictions. Under these conditions, conducting unsheltered PIT counts, which require mobilizing large teams of volunteers to canvas streets, parks, and encampments through the night, posed serious logistical and public health challenges. HUD recognized this reality and encouraged CoCs to evaluate whether conducting an unsheltered count would pose unacceptable risks of exacerbating COVID-19 transmission among people experiencing homelessness, homeless assistance staff, and volunteers (Henry et al., 2022).

The majority of CoCs concluded that it would. Of the nation's 385 CoCs present in 2021,

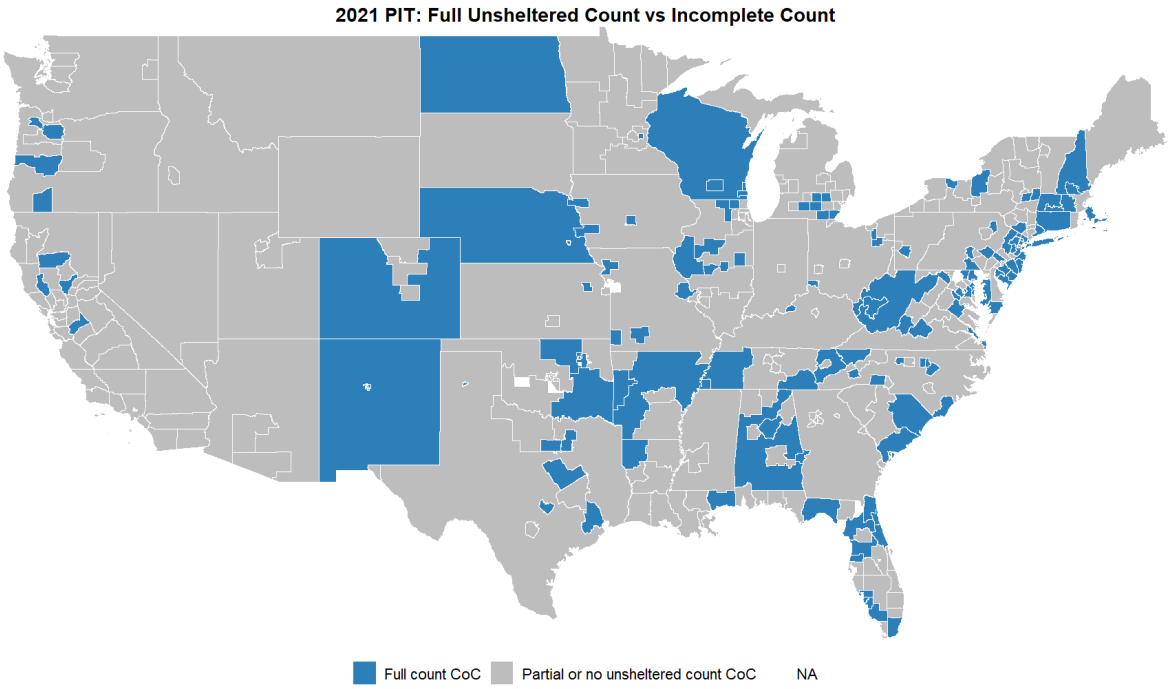

Figure 2: Map of CoCs by 2021 Unsheltered Count Status

only 149 (38.4 percent) conducted a complete sheltered and unsheltered count. Another 74 CoCs performed partial unsheltered counts (enumerating some but not all unsheltered subpopulations or geographic areas) while the remaining 162 CoCs conducted sheltered-only counts, forgoing the unsheltered component entirely. Following standard practice in the literature, I classify CoCs with partial unsheltered counts alongside those with no unsheltered count, as incomplete enumerations cannot be treated as comparable to full counts. This yields 236 CoCs (61.6 percent of the national total) with missing unsheltered count data. The geographic distribution of those CoCs can be seen in Figure 2.

The consequences were immediate and far-reaching. The 2021 AHAR to Congress was forced to exclude unsheltered populations entirely from its national estimates, noting that the communities that did conduct counts "are not representative of all communities across the United States" (Henry et al., 2022). This exclusion means that the only national assessment of homelessness during the most significant economic and public health crisis in decades captures, at best, half the picture. The 2022 AHAR subsequently reported a three percent increase in unsheltered homelessness between 2020 and 2022, alongside a seven per-

cent increase in sheltered homelessness between 2021 and 2022 that likely reflected the easing of pandemic-era shelter capacity restrictions (de Sousa et al., 2023). But without reliable 2021 unsheltered data, whether the increase observed in 2022 represents a continuation, acceleration, or reversal of trends that began during the pandemic cannot be determined.

2.3 Research on Homelessness Count Data

The empirical study of homelessness determinants has developed in tandem with improvements in national data collection. Early work relied on limited cross-sectional data: Honig and Filer (1993) and Elliott and Krivo (1991) established that local housing supply constraints, labor market conditions, and welfare generosity predicted homelessness incidence across metropolitan areas, while O'Flaherty (1995) provided the foundational theoretical framework, modeling homelessness as arising from a mismatch between the lower tail of the income distribution and the lower tail of the housing price distribution. These studies identified the broad structural forces—housing costs, poverty, and labor market weakness—that subsequent work has consistently confirmed, but were constrained by the absence of nationally standardized count data.

The introduction of HUD's Point-in-Time counts in the mid-2000s enabled a new generation of CoC-level analyses. Byrne et al. (2013) used this data to examine community-level determinants, finding housing market variables—particularly median rent and rent burden—among the strongest predictors. Corinth and Lucas (2018) documented the role of January temperatures in shaping the geographic distribution of unsheltered homelessness. Hanratty (2017) provided panel fixed-effects estimates showing that a ten percent increase in median rents was associated with a nine percent increase in homelessness rates. On the funding side, Moulton (2013) found that permanent supportive housing investments reduced chronic homelessness more effectively than general CoC awards, while Popov (2016) and Lucas (2017) addressed the endogeneity of funding allocations using instrumental variables strategies.

Recent methodological advances have improved homeless population measurement and

prediction. Glynn and Fox (2019) develop a dynamic Bayesian model distinguishing counted from true homeless populations across 25 major metros, finding rental cost effects strongest in New York, Los Angeles, and Seattle. Nisar et al. (2019) demonstrate that housing market factors—rental costs, crowding, and evictions—most consistently predict community-level homelessness rates. Meyer et al. (2023) use linked Census-HMIS microdata to estimate approximately 400,000 sheltered and 200,000 unsheltered homeless individuals nationally, with over 90 percent of those sheltered counted in the Census though often misclassified. Chien et al. (2024) show that administrative and citizen-generated data can predict unsheltered homelessness magnitude and spatial distribution in Los Angeles between annual Point-in-Time counts. These studies demonstrate both the feasibility of improving measurement through novel data linkages and the persistent challenges of enumerating a mobile population.

While this body of work has produced robust evidence on what drives cross-sectional and temporal variation in homelessness counts, it has focused almost exclusively on estimation and inference rather than prediction and imputation. No study, to my knowledge, has addressed the problem of reconstructing missing PIT count data at a national scale, particularly in a setting where missingness is driven by an exogenous shock rather than random nonresponse. This paper adapts the predictor relationships established in the determinants literature to an imputation objective, using the same variables that prior work has shown to explain variation in homelessness counts as inputs to a predictive modeling framework designed to fill the 2021 data gap.

2.4 Machine Learning in Social Policy Research

Machine learning methods have seen growing adoption in social science and policy research, particularly for prediction tasks where the goal is accurate out-of-sample forecasting rather than unbiased estimation of causal parameters. In settings with large numbers of candidate predictors and potentially nonlinear relationships, ensemble methods such as random forests

and gradient boosting machines (Breiman, 2001; Friedman, 2001; Chen and Guestrin, 2016) have consistently demonstrated superior predictive performance relative to traditional linear regression (Embaye et al., 2021; Ruhnke et al., 2022; Downing, 2025). This advantage has been documented across a range of applied contexts: Embaye et al. (2021) show that boosting, bagging, and random forest methods outperform OLS in predicting rental housing values in household surveys across multiple countries and years, while Ruhnke et al. (2022) find that random forest imputation of immigration legal status in nationally representative survey data yields greater accuracy and less bias than regression-based approaches.

The present application shares key features with this literature. The prediction task involves a moderately high-dimensional set of community-level predictors with plausible nonlinearities and interactions—precisely the conditions under which tree-based ensemble methods tend to outperform linear models. At the same time, the imputation context demands more than point prediction accuracy: downstream users of the imputed data need reliable uncertainty quantification, and the predictions must generalize to observations (non-counting CoCs) that may differ systematically from the training sample. These requirements motivate the multi-model comparison strategy adopted in this paper, in which linear and machine learning approaches are evaluated side by side, and the selection bias inherent in the Phase 2 training sample is addressed explicitly.

2.5 Imputation Methods and Error Correction Approaches

The standard statistical framework for handling missing data is multiple imputation, formalized by Rubin (1987) and operationalized through methods such as multivariate imputation by chained equations. These approaches assume that data are missing at random—that is, conditional on observed covariates, the probability of missingness is unrelated to the missing values themselves. Recent work has extended this framework by incorporating machine learning into the imputation step: Chen and Xu (2025) provide a unified treatment of ML-based imputation, inverse propensity score weighting, and doubly robust methods

for high-dimensional settings, demonstrating that methods such as XGBoost and deep neural networks can handle the nonlinear relationships that challenge traditional imputation in large-scale datasets. Dang et al. (2026) show that even basic imputation models with modest predictor sets can produce reliable poverty estimates when survey data are missing, though accuracy improves with richer covariates.

The 2021 PIT count data gap, however, violates the missing-at-random assumption in a fundamental way. Missingness was driven by an exogenous shock (the COVID-19 pandemic) that simultaneously altered the outcome of interest. Communities did not simply fail to report their unsheltered counts; the pandemic changed what those counts would have been through economic dislocation, emergency policy responses, and shifts in shelter capacity. A standard imputation model trained on pre-pandemic data would recover the counterfactual (what counts would have been without COVID) rather than the actual 2021 values, while a model incorporating 2021 covariates would conflate the stable structural determinants of homelessness with transient pandemic effects. This paper's two-phase approach addresses the problem by separating these components: the first phase estimates the counterfactual using pre-pandemic relationships, and the second phase models the COVID-specific deviation as a function of pandemic-related variables observed for the subset of CoCs that did conduct counts. The logic parallels the selection correction tradition in econometrics (Heckman, 1976), where modeling the selection process explicitly allows for unbiased estimation in the presence of non-random sample composition, though the objective here is prediction rather than causal identification.

3 Data and Method

3.1 Methodological Overview

The imputation framework developed in this paper proceeds in two phases, each addressing a distinct component of the prediction problem. The central insight motivating this structure

is that predicting missing 2021 unsheltered counts requires solving two problems simultaneously: estimating what counts would have been under normal conditions, and estimating how the COVID-19 pandemic altered those conditions. Collapsing these into a single model would conflate the stable, long-run determinants of unsheltered homelessness with the transient shock of the pandemic, making it difficult to disentangle structural relationships from crisis-specific effects. Separating them allows each phase to be estimated on the data best suited to identify it.

A critical feature of the temporal structure warrants emphasis: Point-in-Time counts conducted in January of year t reflect conditions from the preceding calendar year $t-1$. People experiencing unsheltered homelessness in January 2020, for example, arrived at that state through processes unfolding in 2019: job losses, evictions, relationship dissolutions, or reductions in shelter capacity that occurred months earlier. Throughout this paper, predictors are therefore temporally aligned: the January 2020 PIT count is predicted using 2019 community characteristics, the January 2021 count from 2020 characteristics, and so forth. This one-year lag ensures that predictions capture the causal priority of community conditions over homelessness outcomes and avoids the mechanical relationship that would arise from using contemporaneous measures.

In the first phase, I estimate a baseline prediction model using a panel of CoC-level observations from 2015 through 2019. This model captures the historical relationship between community characteristics, such as housing markets, economic conditions, demographics, climate, shelter infrastructure, and unsheltered homelessness counts. The outcome is modeled on the log scale to address heteroskedasticity in count data, with predictions back-transformed to the count scale for interpretability. Multiple predictive algorithms are trained and compared, ranging from linear regression to ensemble machine learning methods, with out-of-sample predictive accuracy assessed on a locked 2020 holdout set. The selected model is then retrained on the full 2015–2019 panel and used to generate 2021 predictions for every CoC. These predictions represent a counterfactual: the unsheltered counts one would expect

if pre-pandemic relationships had continued unchanged into 2021.

The second phase models the deviation between this counterfactual and observed reality. For the 149 CoCs that conducted complete unsheltered counts in 2021, I compute the residual—the difference between the actual count and the Phase 1 prediction. This residual isolates the COVID-specific component: the portion of the 2021 count that cannot be explained by historical patterns alone. I then model these residuals as a function of pandemic-related variables, including COVID-19 severity, emergency relief programs, changes in shelter capacity, and local policy responses. Because the CoCs that conducted complete counts in 2021 differ systematically from those that did not, this phase incorporates overlap-based propensity score weighting to address the resulting selection bias and ensure that the estimated COVID adjustment generalizes to the non-counting CoCs.

The final imputed value for each non-counting CoC is the sum of its Phase 1 baseline prediction and its Phase 2 predicted COVID adjustment. All imputed values are accompanied by prediction intervals constructed via end-to-end clustered bootstrap, propagating uncertainty from both phases and providing downstream users with a transparent measure of estimate precision. Importantly, this framework does not claim to recover the true unsheltered count for any CoC; rather, it produces principled, uncertainty-quantified estimates that are preferable to the available alternatives of dropping 2021 from longitudinal analyses entirely, analyzing only the non-representative subset of CoCs that did count, or naively extrapolating from pre-pandemic trends.

3.2 Data Sources

The analysis draws on ten data sources spanning community characteristics, health conditions, climate, shelter infrastructure, federal funding, and pandemic-specific conditions. Tables 1 and 2 enumerate the full set of Phase 1 predictors along with their units and sources; Table 3 lists the Phase 2 predictors.

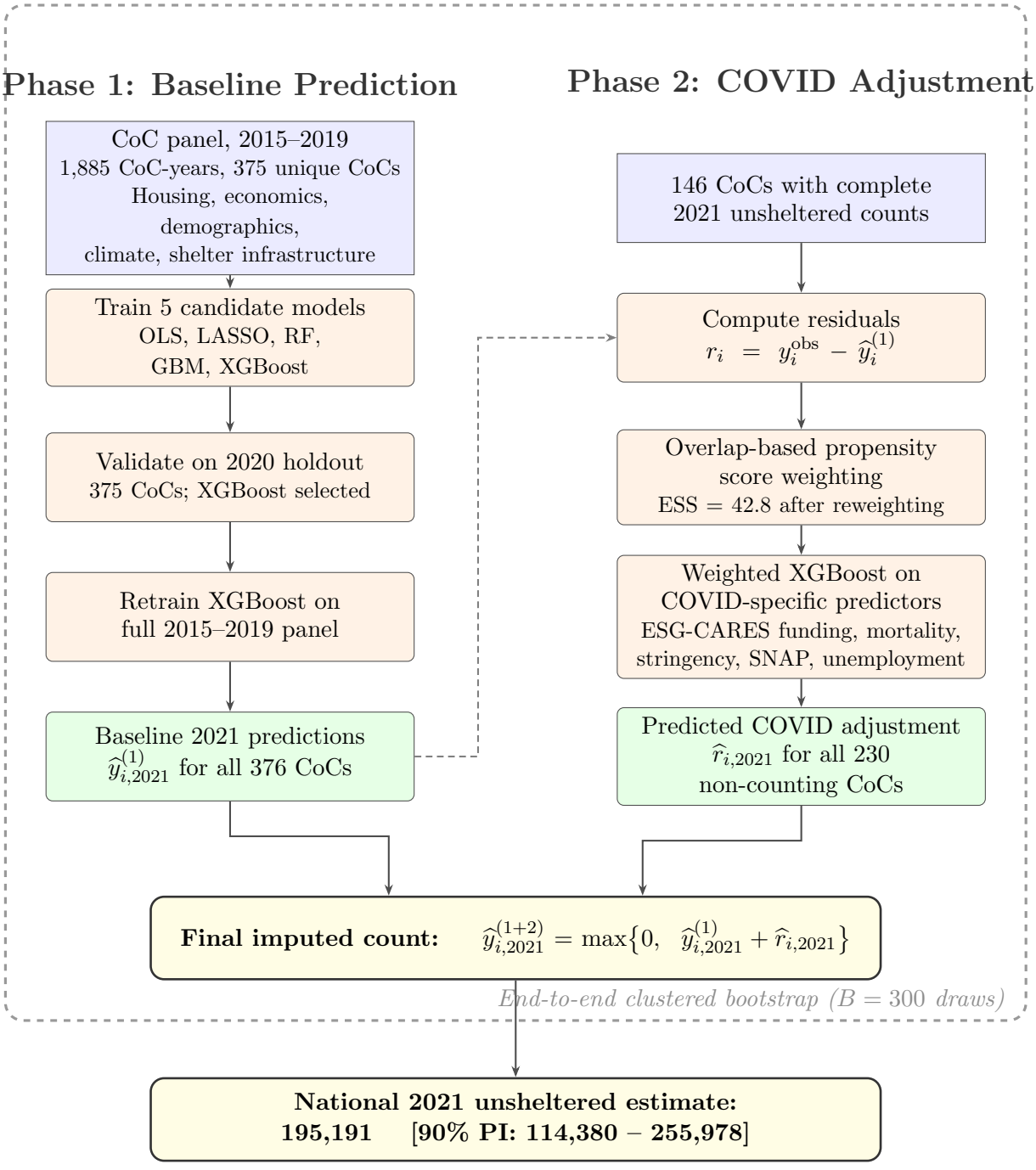

Figure 3: Methodological Flowchart: Two-Phase Imputation Framework

Table 1: Phase 1 Predictor Variables: Descriptions and Data Sources (Part I)

Variable	Unit	Data Source
Homeless Services Infrastructure		
Emergency Shelter Beds	Count	Housing and Urban Development Housing Inventory Count
Transitional Housing Beds	Count	Housing and Urban Development Housing Inventory Count
Permanent Supportive Housing Beds	Count	Housing and Urban Development Housing Inventory Count
Seasonal Shelter Beds	Count	Housing and Urban Development Housing Inventory Count
Overflow Shelter Beds	Count	Housing and Urban Development Housing Inventory Count
Housing Market		
Median Gross Rent	USD	American Community Survey 5-Year Estimates
Rent-Burdened Households	% households	American Community Survey 5-Year Estimates
Rental Vacancy Rate	%	American Community Survey 5-Year Estimates
Economic Conditions		
Unemployment Rate	%	American Community Survey 5-Year Estimates
Poverty Rate	%	American Community Survey 5-Year Estimates
Median Household Income	USD	American Community Survey 5-Year Estimates
Demographics and Mobility		
% Black Population	% population	American Community Survey 5-Year Estimates
% Hispanic Population	% population	American Community Survey 5-Year Estimates
Population Age 65+	% population	American Community Survey 5-Year Estimates
Population Age 18–24	% population	American Community Survey 5-Year Estimates
Bachelor's Degree or Higher	% pop. 25+	American Community Survey 5-Year Estimates
One-Person Households	% households	American Community Survey 5-Year Estimates
Divorced Population	% adults	American Community Survey 5-Year Estimates
Moved from Different State	% population	American Community Survey 5-Year Estimates
Moved from Abroad	% population	American Community Survey 5-Year Estimates

Note: All variables are measured at the Continuum of Care (CoC) level via spatial aggregation to CoC boundaries using annual HUD CoC shapefiles. American Community Survey variables reflect 5-year estimates aggregated from county-level data using a population-weighted crosswalk. All variables in the Phase 1 estimation panel span the years 2015–2019. All predictors from year t are aligned with the January year $t + 1$. t+1 PIT count outcome.

Table 2: Phase 1 Predictor Variables: Descriptions and Data Sources (Part II)

Variable	Unit	Data Source
Health and Social Conditions		
Mental Health Providers	Count	Centers for Medicare & Medicaid Services
Adult Obesity Rate	% adults	Behavioral Risk Factor Surveillance System
Teen Birth Rate	Per 1,000 females 15–19	National Center for Health Statistics Natality Files
Uninsured Adults	% adults	Small Area Health Insurance Estimates
Diabetes Prevalence	% adults	Behavioral Risk Factor Surveillance System
Households Receiving SNAP	% households	American Community Survey 5-Year Estimates
Population with Disability	% population	American Community Survey 5-Year Estimates
Veteran Population	% adults	American Community Survey 5-Year Estimates
Single-Parent Households	% households	American Community Survey 5-Year Estimates
Households with No Vehicle	% households	American Community Survey 5-Year Estimates
Very-Low-Income Renter Households	% renter hh.	American Community Survey 5-Year Estimates
Climate (January)		
Average Temperature	Degrees Celsius	National Oceanic and Atmospheric Administration
Total Precipitation	Millimeters	National Oceanic and Atmospheric Administration
CoC Characteristics and Federal Funding		
CoC Urbanicity Classification	Categorical	Housing and Urban Development Point-in-Time Count
Tenant Protection Policies	Count	National Low Income Housing Coalition Tenant Protection Database

Note: All Health and Social Condition variables are aggregated to the CoC level using an area-weighted spatial join to CoC boundaries. Climate variables are derived from county weather station data spatially interpolated to CoC centroids. All predictors from year t are aligned with the January year $t + 1$.

Table 3: Phase 2 Predictor Variables: COVID-19 Adjustment Model

Variable	Unit	Data Source
COVID-19 Severity and Policy		
COVID-19 Deaths	Count	Centers for Disease Control and Prevention
Any Mask Mandate	Binary (0/1)	CDC State, Territory, and County Public Mask Mandates
Lockdown Policy Stringency	Score (0-100)	Oxford Coronavirus Government Response Tracker
Economic Conditions		
Unemployment Rate	%	American Community Survey 5-Year Estimates
Households Receiving SNAP	% households	American Community Survey 5-Year Estimates
Emergency Relief and Housing Policy		
ESG-CARES Funding	USD (total)	Housing and Urban Development Allocations and Awards
Emergency Rental Assistance	Binary (0/1)	National Low Income Housing Coalition Tenant Protection Database
Any Eviction Moratorium	Binary (0/1)	National Low Income Housing Coalition Tenant Protection Database

Note: Phase 2 is estimated on the cross-section of CoCs that conducted complete unsheltered counts in 2021. The dependent variable is the residual between the Phase 1 naive prediction and the observed 2021 unsheltered count, which was conducted in January 2021. All variables reflect 2020 annual values. The mask mandate variable is coded as 1 if any county within the CoC had an active mask mandate policy; the underlying CDC dataset records mandate status at the county-day level. Emergency rental assistance and eviction moratorium indicators are coded as 1 if any jurisdiction within the CoC had an active policy as of January 2021.

The primary outcome variable, the annual unsheltered Point-in-Time count, comes from HUD's Point-in-Time Count dataset, which reports CoC-level counts of sheltered and unsheltered people on a single night each January. I use counts from 2015 through 2022. All predictors are temporally aligned to reflect the one-year lag between community conditions and homelessness outcomes: the January 2020 PIT count is predicted using 2019 community characteristics, and so forth. Phase 1 models the outcome on the log scale to address heteroskedasticity, with predictions back-transformed to counts for interpretation. The Housing Inventory Count, also administered by HUD, supplies annual CoC-level data on shelter bed capacity, including emergency shelter, transitional housing, permanent supportive housing, seasonal, and overflow beds. Both datasets are available from HUD's public data portal and are reported at the CoC level, requiring no spatial aggregation.

The bulk of the Phase 1 predictor set comes from the American Community Survey (ACS) five-year estimates, accessed via the Census Bureau's API. Health indicators, including adult obesity rates, teen birth rates, uninsured rates, and the count of mental health providers, come from sources including the Behavioral Risk Factor Surveillance System, the National Center for Health Statistics, and the Centers for Medicare and Medicaid Services National Provider Identification file. These data are also available at the county level.

Average January temperature and precipitation, two predictors with well-documented associations with unsheltered homelessness, come from the NOAA Global Historical Climatology Network Daily dataset. CoC Program funding allocations, used to construct the federal grant funding per capita predictor, come from HUD's annual awards data. Tenant protection policies are sourced from the National Low Income Housing Coalition's Tenant Protection Database, which documents the universe of active renter protections at the state and local level, including just-cause eviction requirements, rent stabilization ordinances, and right-to-counsel provisions among others.

The Phase 2 predictors introduce pandemic-specific data sources. COVID-19 death counts at the county level come from the Centers for Disease Control and Prevention.

County-level mask mandate records come from the CDC's State and Territorial Public Mask Mandates dataset, which records mandate status at the county-day level from April 2020 through January 2021. I aggregate this to a CoC-level binary indicator equal to one if any county within the CoC had an active mandate. Emergency Solutions Grants-CARES Act funding allocations come from HUD's awards data. Eviction moratorium and Emergency Rental Assistance policy indicators are constructed from the National Low Income Housing Coalition's Tenant Protection Database, restricted to policies enacted between March 2020 and January 2021.

3.3 Data Construction and Sample Inclusion Criteria

A central challenge in constructing the analysis dataset is that CoC boundaries do not correspond to any standard Census geography. CoC boundaries are administrative units defined by HUD that may encompass individual cities, single counties, multi-county regions, or entire states, and they frequently cross county and metropolitan area lines. Because demographic, economic, and housing market variables are collected at the Census tract or county level, and health indicators and policy data are reported at the county or jurisdiction level, a systematic procedure is required to aggregate sub-CoC data up to the CoC level.

I adopt a centroid-based spatial assignment approach. For each year in the panel, I obtain the corresponding CoC boundary shapefiles from HUD and county or jurisdiction geometries from the Census Bureau. I compute the geographic centroid of each sub-CoC unit—county, Census place, or other jurisdiction—and assign it to the CoC whose boundaries contain that centroid. Variables are then aggregated to the CoC level using summation for count variables (e.g., total COVID-19 cases, total ESG-CARES funding) and population-weighted averaging for rate and proportion variables (e.g., poverty rate, unemployment rate, percent rent-burdened). Matching jurisdiction-level policy data, such as tenant protection laws and COVID-era eviction moratoria, to CoC boundaries requires an additional step since policy databases report jurisdictions by name rather than by standardized geographic identifier. I

match jurisdiction names to Census place and county geographies using exact string matching where possible and Levenshtein-distance fuzzy matching for the remainder, with manual corrections applied to consolidated city-counties and other non-standard jurisdictions that resist automated matching.

The final consolidated dataset spans 2015 through 2019 and merges data from HUD's Point-in-Time Counts and Housing Inventory Counts, the American Community Survey, the Behavioral Risk Factor Surveillance System, the Small Area Health Insurance Estimates, the Center for Disease Control (CDC), NOAA climate records, CoC Program funding allocations, tenant protection policy databases, CDC COVID-19 case and mortality data, county-level mask mandate records, state and local lockdown stringency measures, ESG-CARES emergency funding allocations, and Emergency Rental Assistance and eviction moratorium policy records. Both modeling phases impose complete-case requirements—observations with missing outcome or predictor data are excluded—but differ in their treatment of panel structure and temporal coverage.

Phase 1 estimates baseline relationships using a panel of CoC-year observations from 2015 through 2019. The sample construction proceeds in three steps. First, the temporally shifted panel is restricted to non-territory CoCs, as U.S. territories exhibit fundamentally different housing markets, federal funding structures, and data reporting patterns that make pooled estimation inappropriate. Second, CoC-years with missing outcome data (unsheltered PIT count) are dropped, removing 107 of 2,214 observations. Third, CoC-years with missing values on any Phase 1 predictor are excluded via complete-case analysis, removing an additional 622 observations and yielding a final training sample of 1,485 CoC-years spanning 375 unique CoCs. This is done as the unsheltered count outcomes are transformed to the log scale (specifically, $\log(\text{count} + 1)$) to address heteroskedasticity in count data.

Critically, Phase 1 allows CoCs to enter and exit the panel across years. Of the 375 unique CoCs in the training sample, 367 (98 percent) appear in all five years (2015–2019), six appear in four years, one appears in three years, and one appears in fewer than three years. This

unbalanced panel structure reflects real-world administrative dynamics: CoC boundaries are periodically redrawn, new CoCs are created when regions split or reorganize their homeless assistance systems, and data quality improves over time as reporting infrastructure matures. Restricting the sample to a balanced panel would exclude CoCs experiencing boundary changes and bias the training data toward stable, administratively mature jurisdictions. The machine learning methods employed in Phase 1 handle unbalanced panels naturally by treating each CoC-year as an independent observation, avoiding the fixed-effects logic that requires within-CoC variation (Efron and Hastie, 2016). The 2020 validation set contains 371 CoCs, nearly matching the 375 unique CoCs in training, and the 2021 baseline prediction set contains 386 CoCs, representing effectively complete national coverage of non-territory jurisdictions.

Phase 2 training imposes stricter requirements due to the cross-sectional structure and the need for comparability with Phase 1 predictions. While there are 385 CoCs that reported a PIT count of any kind in 2021, 5 come from US territories (Guam, Puerto Rico, the Virgin Islands, and the Mariana Islands) that lack sufficient feature variable data to be included. Furthermore, 4 CoCs (AR-505, MO-604, MD-514, and OK-504) lacked crucial COVID-related feature variable data that necessitated inclusion as well. Because of this, 376 CoCs in total are considered for the Phase 2 model process, with 230 CoCs listed as having partial or no unsheltered counts (the CoCs to impute for) and 146 CoCs as having completed the full unsheltered count in 2021.

The complete-case approach adopted here is conservative but appropriate for the prediction objective. Multiple imputation of missing predictors would introduce additional uncertainty into already-uncertain imputations and complicate the interpretation of prediction intervals. Moreover, the variables with the most missingness (health indicators) contribute modestly to Phase 1 predictive accuracy, as evidenced by feature importance rankings that place them outside the top ten predictors. The near-complete coverage of the 2021 baseline prediction set (376 of approximately 380 non-territory CoCs) confirms that predictor miss-

ingness is concentrated in the training period rather than the application year, mitigating concerns about coverage bias in the final imputed estimates.

3.4 Phase 1: Baseline Prediction Model

Phase 1 constructs a pre-pandemic baseline mapping from CoC characteristics to subsequent unsheltered PIT counts. The goal is not causal identification, but a counterfactual prediction: what each CoC's January 2021 unsheltered count would have been under the pre-2020 relationship between local conditions and unsheltered homelessness. These baseline predictions serve two purposes. First, they provide a "no-pandemic" benchmark against which observed 2021 counts can be compared. Second, for CoCs that conducted complete PIT counts in 2021, the baseline prediction errors (residuals) isolate the component of the 2021 deviation that is not explained by pre-pandemic covariates and therefore becomes the target for the Phase 2 COVID-era adjustment model.

I compare five prediction algorithms spanning a spectrum from interpretable linear models to flexible nonparametric ensembles. The linear benchmark is ordinary least squares (OLS), which provides a transparent baseline and a reference point for incremental gains from more flexible methods. The second model is LASSO regression, which augments the OLS objective with an L_1 penalty and therefore performs shrinkage and implicit variable selection when predictors are numerous and potentially collinear; the penalty parameter is selected by cross-validation. The remaining models are regression tree-based ensembles that accommodate nonlinearities and interactions without requiring them to be specified ex ante. Random Forest averages many trees grown on bootstrap samples to reduce variance and typically performs well in settings with complex predictor interactions (Breiman, 2001). Gradient Boosting Machines (GBM) build trees sequentially to reduce bias, fitting each new tree to the residual structure of the existing ensemble (Friedman, 2001). Finally, Extreme Gradient Boosting (XGBoost) is an efficient boosting implementation that combines tree boosting with regularization and subsampling features designed to improve generalization

in high-dimensional settings (Chen and Guestrin, 2016). Collectively, these models provide a disciplined way to evaluate whether the data support a strong nonlinear or interaction structure beyond what linear specifications capture.

The predictor set comprises community characteristics from the prior year, but notably excludes the lagged unsheltered count as a predictor. This design choice departs from much of the homelessness forecasting literature, which routinely includes lagged outcomes to capture the strong empirical persistence of homelessness levels (Byrne et al., 2013; O'Flaherty, 2019; Hanratty, 2017; Corinth and Lucas, 2018). Three considerations motivate the exclusion.

First, methodological portability: a framework reliant on lagged counts cannot be applied when outcome data are missing for multiple consecutive years, precisely the scenario that arises during prolonged crises. Second, structural break concerns: flexible machine learning methods may overweight lagged outcomes and approximate autoregressive rules, which becomes problematic when pandemic disruptions decouple current homelessness from prior-year levels. Third, the cross-sectional imputation context creates a recursive dependency: including a 2020 lagged count would require first imputing that 2020 count, compounding uncertainty. Omitting the lag forces the model to extract signal from community characteristics alone and maintains a clear separation between the baseline counterfactual (Phase 1) and the pandemic-specific adjustment (Phase 2). All models are estimated with the unsheltered count outcome transformed to $\log(\text{count} + 1)$ to address heteroskedasticity in count data, with predictions back-transformed to the count scale for evaluation. Summary statistics for the full Phase 1 feature variable set is seen in Table 4.

For each algorithm, I adopt a common training and validation workflow. To evaluate which candidate model performs the best, I train all models on 2015 to 2017 data with hyperparameters being selected by grid search using grouped cross-validation on the 2018 set, with the objective of minimizing out-of-sample prediction error. For Random Forest, the grid spans the number of trees, the number of predictors sampled at each split, and minimum node size. For GBM and XGBoost, the grid spans tree depth, learning rate, subsampling

Table 4: Phase 1 Modeling Sample (2015–2019) Summary Statistics

Variable	N	Mean	SD	Min	Median	Max
Homelessness outcome						
Total unsheltered homeless	1,885	521.8	2,124.4	0	113	42,471
Homeless services infrastructure						
Emergency shelter beds	1,885	739	3,853	0	266	75,245
Transitional housing beds	1,885	303	507	0	147	6,760
Permanent supportive housing beds	1,885	942	2,167	0	380	32,150
Seasonal shelter beds	1,885	56	119	0	15	1,791
Overflow shelter beds	1,885	45	89	0	12	1,101
Housing market						
Median rent (USD)	1,885	1,064	321	597	986	2,439
Rent-burdened households (%)	1,885	50.1	4.9	36.4	49.5	66.5
Rental vacancy rate (%)	1,885	6.4	2.8	1.7	6.0	24.7
Tenant protection policies (count)	1,885	1.0	1.8	0	0	10
Economic conditions						
Unemployment rate (%)	1,885	6.4	2.0	1.9	6.0	22.2
Poverty rate (%)	1,885	14.3	4.6	3.4	14.2	39.5
Median household income (USD)	1,885	64,489	17,989	27,901	60,003	147,536
Demographics and mobility						
% Black population	1,885	12.2	12.6	0.4	8.2	79.6
% Hispanic population	1,885	13.4	13.3	1.0	8.7	84.2
Population age 65+ (%)	1,885	15.7	3.7	7.4	15.5	39.3
Population age 18–24 (%)	1,885	10.0	2.7	5.3	9.3	28.3
Bachelor's degree or higher (%)	1,885	31.2	10.6	12.6	29.7	79.0
One-person households (%)	1,885	28.3	4.4	12.7	28.5	48.2
Divorced (%)	1,885	11.2	1.9	5.9	11.2	17.4
Moved from different state (%)	1,885	2.5	1.3	0.5	2.2	8.4
Moved from abroad (%)	1,885	0.6	0.5	0.02	0.5	4.9
Health and social conditions						
Mental health providers (count)	1,885	2,071	3,577	60	1,067	31,986
Obesity rate (%)	1,885	28.5	4.7	14.8	28.8	39.8
Teen birth rate (per 1,000)	1,885	29.1	12.9	3.6	27.8	90.0
Uninsured rate (%)	1,885	11.9	5.2	2.1	11.3	33.1
Diabetes rate (%)	1,885	10.2	2.0	4.2	10.1	16.6
Households with SNAP (%)	1,885	12.3	4.8	2.5	12.1	42.4
Disability rate (%)	1,885	13.1	3.0	5.4	13.1	22.5
Veteran population (%)	1,885	8.2	2.7	1.8	8.2	20.5
Single-parent households (%)	1,885	13.9	3.4	6.5	13.6	32.3
Households without vehicle (%)	1,885	7.9	5.3	2.3	6.7	54.8
Very low income renters (%)	1,885	28.3	7.9	7.3	28.6	54.1
Climate (January)						
Average temperature (°C)	1,885	2.3	7.0	−16.7	1.2	23.2
Total precipitation (mm)	1,885	90.5	71.1	0.0	78.5	651.7

Note: Summary statistics for the Phase 1 complete-case modeling sample (N = 1,885 CoC-years from 375 unique CoCs, 2015–2019). This sample excludes CoC-years with missing outcomes or predictors. All variables are measured at the Continuum of Care level via spatial aggregation. The outcome variable (total unsheltered homeless) is modeled on the log scale; summary statistics shown here are in levels (raw counts).

fractions, and regularization settings, with the number of boosting rounds chosen using early stopping. Using a unified holdout and consistent tuning protocol ensures that comparisons reflect predictive performance rather than differences in evaluation design. Once the optimal hyperparameter values are determined, each candidate model is evaluated on its performance in predicting the held-out test sample of 2019 data. Once the most accurate model is determined, the final model used for the Phase 1 predictions is re-trained on all 2015 to 2019 data to make the 2021 baseline predictions. Figure 4 displays this process.

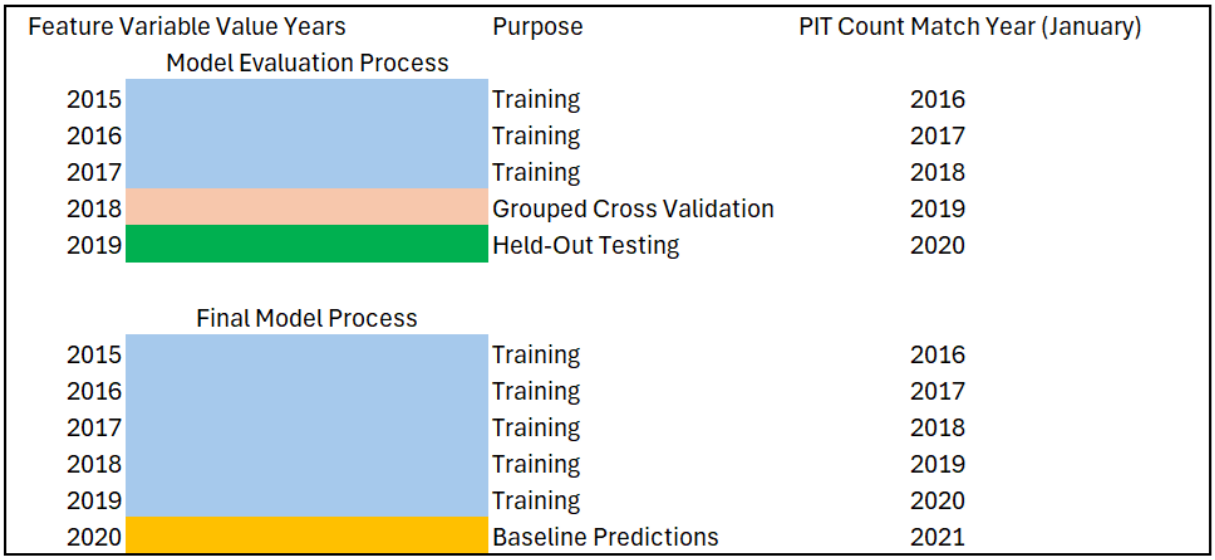

Figure 4: Visual Diagram of Phase 1 Training Process

Model performance is summarized using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2) on the held-out test data. Because CoCs vary substantially in size and urbanicity, I additionally examine holdout performance by CoC category and by size strata to assess whether prediction error is systematically concentrated in particular parts of the CoC distribution (e.g., large urban CoCs versus smaller Balance-of-State CoCs).

I select the Phase 1 model as the specification (algorithm $\times$ predictor set) that yields the best holdout performance under the primary accuracy criterion (RMSE, with MAE used to assess robustness to outliers). After this selection, I then generate baseline predictions for

each CoC using its predictor-year 2020 covariates (aligned to the January 2021 PIT outcome) to maximize information used to estimate the baseline mapping, producing $\widehat{U}_{c,2021}^{\text{baseline}}$, the predicted unsheltered count absent pandemic-era disruptions.

3.5 Phase 2: COVID Adjustment Model

Phase 1 produces a baseline prediction for each CoC's 2021 unsheltered count using pre-pandemic relationships learned from the Phase 1 panel. However, 2021 is not a typical missing-data year: the pandemic plausibly induced systematic deviations from historic patterns in ways that differ across CoCs. Phase 2 is designed to explicitly model this COVID-era deviation and use it to adjust the Phase 1 baseline for CoCs that did not conduct a complete unsheltered count in 2021.

This is the primary concern in Phase 2: that CoCs that completed full unsheltered counts in 2021 may differ systematically from those that did not, which can be seen in Table 5. If so, a residual model trained on full-count CoCs could extrapolate poorly to the non-counting group. To diagnose and mitigate this covariate shift, I estimate a propensity score model for the probability of completing a full 2021 count using a broader set of contextual covariates (population size, income, rents and vacancy rates, poverty, age composition, insurance and health measures, and shelter capacity measures), along with COVID-era variables and CoC category. Specifically, I fit a logistic regression

$$\Pr(T_i = 1 \mid X_i) = \frac{1}{1 + \exp(-X_i'\beta)},$$

where $T_i = 1$ indicates a full 2021 count and X_i' is the covariate set shown in Table 5.

From this model I compute a propensity score $\widehat{p}_i$ and inspect overlap between full-count and non-full-count CoCs by comparing the distributions of $\widehat{p}_i$ across groups. To improve transportability of the residual model to non-counting CoCs, I then construct an inverse-propensity reweighting scheme that reweights full-count CoCs to resemble the covariate

distribution of non-full-count CoCs. Concretely, I weight each full-count CoC by

$$w_i \;=\; \frac{1-\widehat{p}_i}{\widehat{p}_i} \quad \text{for } T_i = 1,$$

which reweights the treated (counting) sample to approximate the covariate distribution of the non-counting CoCs, ensuring that the residual model learns patterns representative of the population it will be applied to. To stabilize the procedure, propensity scores are truncated away from 0 and 1, and extreme treated weights are capped at the 99th percentile. Finally, I compute standardized mean differences (SMDs) before and after weighting as a balance diagnostic.

Phase 2 is trained only on CoCs that conducted a full 2021 count (sheltered and unsheltered), because these CoCs provide the only observations where the true 2021 unsheltered count is known. For each of these CoCs, I define the dependent variable as the level residual

$$r_{i,2021} \;=\; y^{\text{obs}}_{i,2021} \;-\; \widehat{y}^{(1)}_{i,2021},$$

where $y^{\text{obs}}_{i,2021}$ is the observed 2021 unsheltered PIT count and $\widehat{y}^{(1)}_{i,2021}$ is the Phase 1 baseline prediction. A key modeling choice is that residuals are kept entirely in levels (no transformations). This is necessary because true deviations may be negative, and the residual model should be allowed to predict negative adjustments when supported by the data.

The Phase 2 feature set is intentionally COVID-specific, capturing channels through which 2021 unsheltered counts may have departed from pre-pandemic expectations. The final predictor set includes: COVID mortality burden (total deaths in 2021), policy environment (average stringency index and indicator variables for eviction-related policies such as moratoria and mandates), emergency resources (total ESG-CARES funding), and household economic stress proxies (e.g., SNAP participation and unemployment), along with a CoC category factor to allow systematic differences by urbanicity. These predictors are assembled into a one-row-per-CoC modeling frame for 2021, with consistent indicator coding and factor

Table 5: Phase 2: Covariate Balance by 2021 Counting Status

	Mean		
Variable	Full Count	No Count	SMD
ESG-CARES funding (USD thousands)	58,337	99,149	−0.437
Average policy stringency score	4.7	5.2	−0.328
COVID-19 deaths (2021)	1,451	2,203	−0.247
Total population	706,147	959,486	−0.214
Has eviction moratorium (indicator)	0.02	0.09	−0.195
Has mask mandate (indicator)	0.90	0.96	−0.186
Overflow shelter beds	54	87	−0.177
Unemployment rate (%)	5.4	5.6	−0.122
Obesity rate (%)	30.8	30.2	0.112
Population age 65+ (%)	17.0	16.7	0.088
Poverty rate (%)	12.6	12.9	−0.086
Has Emergency Rental Assistance (indicator)	0.11	0.17	−0.067
Seasonal shelter beds	48	60	−0.065
Emergency shelter beds	964	707	0.062
Median rent (USD)	1,165	1,188	−0.061
Median household income (USD)	75,834	74,662	0.057
Rental vacancy rate (%)	5.97	5.91	0.047
Uninsured rate (%)	9.41	9.62	−0.041
Diabetes rate (%)	10.9	10.8	0.041

Note: Comparison of observable characteristics between CoCs that conducted complete unsheltered counts in 2021 (Full Count, N = 146) and those that did not (No Count, N = 230). SMD = standardized mean difference, calculated as $(m_1 - m_0)/\sqrt{(s_1^2 + s_0^2)/2}$, where m and s denote means and standard deviations. Values sorted by absolute SMD. CoCs that did not count are systematically larger, received more emergency funding, experienced higher COVID mortality, and faced stricter policy environments. CoC category distribution: Full Count (48% largely suburban, 20% other urban, 20% largely rural, 12% major city); No Count (40% largely suburban, 34% largely rural, 13% other urban, 13% major city).

handling prior to model fitting. Table 6 displays summary statistics for the final predictor set of all CoCs in the 2021 data.

Table 6: Phase 2 (2021) Summary Statistics: COVID Adjustment Predictors

Variable	N	Mean	SD	Min	Median	Max
COVID-19 deaths (2021)	376	1,918	3,295	0	1,042	37,139
Average policy stringency score	376	35.01	31.53	1.00	25.36	76.77
ESG-CARES funding (USD thousands)	376	83,803	101,153	4,390	48,352	616,388
Has Emergency Rental Assistance	376	0.13	0.34	0	0	1
Has eviction moratorium	376	0.02	0.15	0	0	1
Has mask mandate	376	0.90	0.29	0	1	1
Households with SNAP (%)	376	11.4	4.6	2.3	11.0	36.9
Unemployment rate (%)	376	5.5	1.5	2.4	5.3	15.3

Note: Summary statistics for Phase 2 COVID-specific predictors measured at the CoC level in 2021. Sample includes all non-territory CoCs with complete Phase 2 predictor data (N = 376). ESG-CARES funding scaled to thousands for readability. Binary indicators (ERA, moratorium, mandate) take values 0 or 1.

Table 7: CoC Urbanicity Composition by 2021 Count Completion Status

CoC category	Share (Full count)	Share (Not full)	Diff.
Largely Rural CoC	19.7	33.6	-13.9
Largely Suburban CoC	48.3	40.1	8.2
Other Largely Urban CoC	20.4	13.4	7.0
Major City CoC	11.6	12.9	-1.4

Notes: Shares are computed within each group (Full count vs. Not full) for the 2021 cross-section of non-territory CoCs. "Diff." reports the difference in shares (Full count − Not full). The urbanicity classification is the HUD CoC category used throughout the analysis to stratify model performance and to assess selection into conducting a full unsheltered count in 2021.

Using the full-count CoCs, the Phase 2 residual model is estimated as a supervised learning problem with outcome $r_{i,2021}$ and predictors described above. I fit the same five candidate model families used in Phase 1: weighted OLS, weighted LASSO, weighted random forest, weighted gradient boosting (GBM), and weighted XGBoost. All models are trained with the overlap-based weights w_i applied to the treated (full-count) sample.

To ensure consistent feature construction across methods, I implement a single preprocessing pipeline using a recipe that removes zero-variance predictors and one-hot encodes factor variables. This produces a stable design matrix for the linear models and for the

tree-based learners. For the ensemble methods (RF, GBM, and XGB), I evaluate a small hyperparameter grid and select the best-performing specification using cross-validated error on the treated sample, subject to the constraint that the sample size (146 full-count CoCs) limits how aggressive tuning can be without overfitting.

For every in-scope CoC, the chosen Phase 2 model generates a predicted residual $\widehat{r}_{i,2021}$. The final imputed 2021 unsheltered count is then formed as

$$\widehat{y}_{i,2021}^{(1+2)} \;=\; \widehat{y}_{i,2021}^{(1)} + \widehat{r}_{i,2021}.$$

Because population counts cannot be negative, I enforce a hard floor at zero:

$$\widehat{y}_{i,2021}^{(1+2,\text{clamp})} \;=\; \max\{0,\; \widehat{y}_{i,2021}^{(1)} + \widehat{r}_{i,2021}\}.$$

Importantly, I also retain an indicator for whether the unclamped prediction was negative, which serves as a diagnostic flag for CoCs where the model implies an unusually large negative adjustment relative to the Phase 1 baseline.

To contextualize model performance against a widely used ad hoc alternative, I compare Phase 1 and Phase 1+2 predictions to midpoint interpolation based on adjacent PIT counts. For CoCs with observed 2021 unsheltered counts, and where both 2020 and 2022 counts are available, the midpoint benchmark is defined as

$$y_{i,2021}^{\text{mid}} \;=\; \frac{y_{i,2020}^{\text{obs}} + y_{i,2022}^{\text{obs}}}{2}.$$

I evaluate prediction accuracy using RMSE and MAE for: (i) midpoint interpolation, (ii) Phase 1 baseline, (iii) Phase 1+2 adjusted predictions, and (iv) the clamped Phase 1+2 predictions. This comparison is reported overall and by CoC category, providing a transparent check on whether the two-phase procedure improves upon a simple, commonly used interpolation rule.

3.6 Uncertainty Quantification

To construct uncertainty intervals that correctly propagate error from both stages of the imputation pipeline, I implement an end-to-end clustered bootstrap at the CoC level. The bootstrap resamples CoCs with replacement from the Phase 1 training panel (clustered by CoC to preserve within-CoC serial dependence), refits both the Phase 1 preprocessing recipe and the log-scale baseline model on each resample, back-transforms predictions to the count scale, and generates a bootstrap draw of $\widehat{y}^{(1)}_{i,2021}$ for all CoCs. Conditional on that draw, I then rebuild the Phase 2 training residuals $r_{i,2021}$ for full-count CoCs, refit the Phase 2 residual model using the overlap-based weights computed on the full sample, and generate bootstrap residual predictions $\widehat{r}_{i,2021}$ for all CoCs. Each bootstrap draw therefore yields an end-to-end imputed prediction

$$\widehat{y}^{(1+2)}_{i,2021,b} \;=\; \widehat{y}^{(1)}_{i,2021,b} + \widehat{r}_{i,2021,b},$$

along with the clamped counterpart $\max\{0, \widehat{y}^{(1+2)}_{i,2021,b}\}$. Repeating this procedure across $B = 300$ bootstrap draws produces an empirical distribution of predictions for each CoC, from which I report percentile-based intervals (e.g., 5th and 95th percentiles for 90% intervals) and the bootstrap frequency of negative *unclamped* predictions as an additional diagnostic.

This end-to-end bootstrap is intentionally conservative: it treats the two-stage prediction pipeline as a single composite estimator and captures uncertainty from both the baseline model and the COVID adjustment model, rather than conditioning on Phase 1 predictions as fixed.

4 Results

4.1 Phase 1: Baseline Model Performance

The Phase 1 baseline model predicts 2021 unsheltered counts using only pre-pandemic structural determinants—housing markets, economic conditions, demographics, climate, shelter

infrastructure, and federal funding—without relying on lagged outcome data. This design choice prioritizes methodological portability: the framework can be applied to settings where outcome data are missing for multiple consecutive years, a common feature of administrative data gaps during prolonged crises.

I compare five candidate algorithms on a locked 2020 holdout set: ordinary least squares, LASSO, random forest, gradient boosting (GBM), and XGBoost. Table 8 reports test-set performance. XGBoost substantially outperforms all alternatives, achieving an RMSE of 905 and an R^2 of 0.972, compared to the next-best model (random forest: RMSE = 1,457, R^2 = 0.934). The linear models perform poorly by comparison—OLS and LASSO produce RMSEs exceeding 9,000—indicating that the relationship between community characteristics and unsheltered homelessness exhibits strong nonlinearities that tree-based ensembles capture but linear specifications miss. XGBoost is selected as the final Phase 1 model and retrained on the full 2015–2019 panel before generating 2021 baseline predictions.

Table 8: Phase 1: Model Comparison on 2020 Holdout Set

Model	N	RMSE	MAE	R^2
XGBoost	375	905	164	0.972
Random Forest	375	1,457	250	0.934
GBM	375	1,834	302	0.771
LASSO	375	9,233	796	0.475
OLS	375	9,886	831	0.474

Note: Out-of-sample performance on the 2020 validation set (year 2019 predictors → January 2020 PIT outcome). All models trained on 2016–2018 data (using 2015-2017 predictors) for cross-validation and hyperparameter tuning, with validation on 2020 holdout (using 2019 predictors). Selected model retrained on full 2015-2019 panel for final predictions. RMSE = root mean squared error; MAE = mean absolute error; R^2 = coefficient of determination. XGBoost selected as Phase 1 model based on superior RMSE.

Prediction accuracy varies systematically by CoC type. Table 9 stratifies holdout per-

formance by urbanicity. The model performs best in largely suburban CoCs (RMSE = 18, MAE = 8) and other largely urban CoCs (RMSE = 12, MAE = 7), where unsheltered counts are moderate and relatively stable. Performance degrades slightly in largely rural CoCs (RMSE = 27, MAE = 11), likely reflecting greater heterogeneity in local conditions and smaller sample sizes for model training. The largest errors occur in major city CoCs (RMSE = 409, MAE = 112), where unsheltered populations are both large and volatile. Importantly, however, even in major cities the model captures the broad magnitude of unsheltered homelessness—mean predicted and actual counts differ by less than 3 percent (2,338 versus 2,410)—suggesting that the imputation procedure will produce defensible estimates across the full CoC distribution, albeit with wider uncertainty intervals for large urban jurisdictions.

Table 9: Phase 1: Prediction Accuracy by CoC Urbanicity (2020 Holdout)

CoC Category	N	RMSE	MAE	Mean Actual	Mean Predicted
Largely Suburban CoC	162	18	8	302	299
Largely Rural CoC	102	27	11	394	388
Other Largely Urban CoC	60	12	7	215	214
Major City CoC	47	409	112	2,410	2,338

Note: XGBoost model performance stratified by CoC urbanicity category on the 2020 validation set. RMSE and MAE reported in levels (unsheltered count). Mean actual and predicted counts show the model captures central tendency well across all categories, with proportionally larger absolute errors in major cities due to scale.

Figure 5 plots predicted versus actual 2020 unsheltered counts on a log scale. The tight clustering around the 45-degree line confirms that the model predicts well across the full range of CoC sizes, with no systematic over- or underprediction at either tail of the distribution. Table 10 reports summary statistics of the baseline predictions.

What drives the baseline predictions? Feature importance rankings (Table 11) reveal that January average temperature dominates the model, accounting for 23.3 percent of total predictive gain. This aligns with prior research documenting that mild winter climates enable year-round outdoor habitation and attract unsheltered populations through migration (Lucas, 2017; Corinth and Lucas, 2018). The next most important predictors are mental

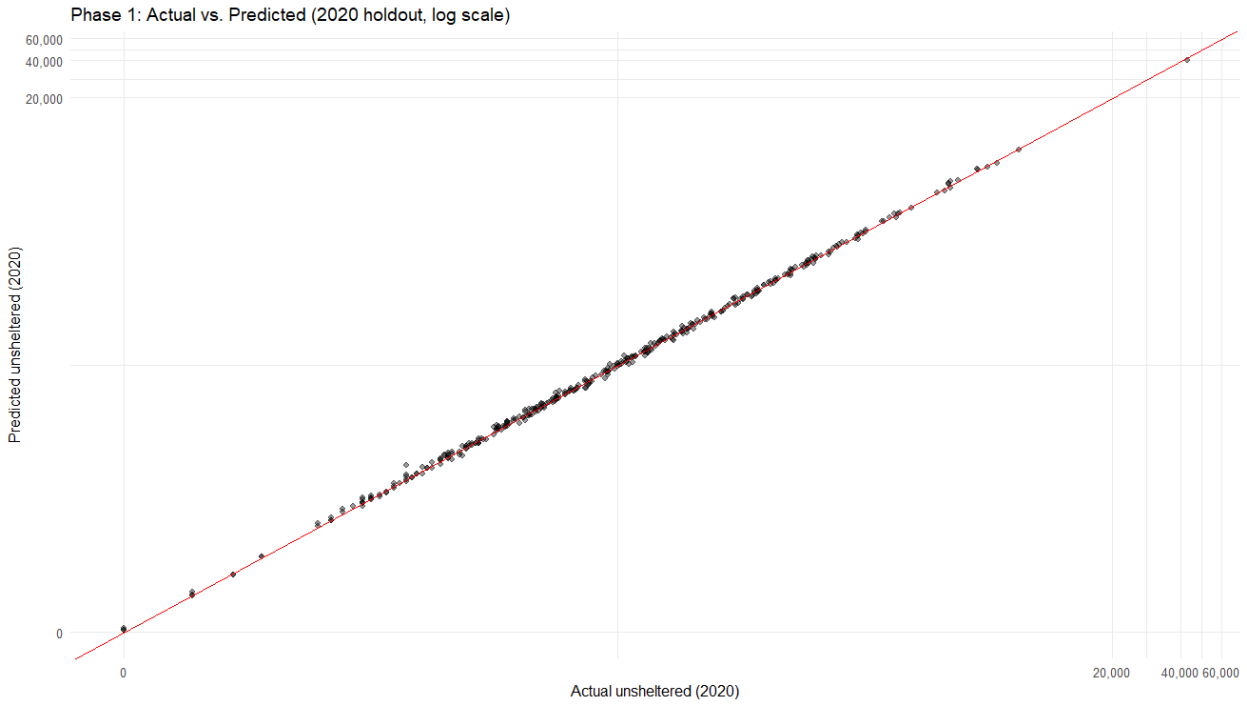

Figure 5: Phase 1 Actual-Predicted Scatterplot, Log Scale

Table 10: Phase 1 Baseline Predictions: Summary Statistics (2021)

	N	Mean	SD	Min	Median	Max
Overall	376	468	1,545	0.6	136	26,830
By CoC Category						
Major City CoC	47	1,916	3,977	106	902	26,830
Largely Rural CoC	105	307	579	0.6	141	4,706
Largely Suburban CoC	163	277	487	6	105	3,412
Other Largely Urban CoC	61	196	296	10	94	1,418

Note: Summary statistics for Phase 1 baseline predictions of 2021 unsheltered counts, generated using 2020 community characteristics and the log-scale XGBoost model retrained on 2015–2019 data. Predictions generated for all 376 CoCs in the 2020 predictor set. Major City CoCs have substantially higher predicted counts on average (mean = 1,916) compared to other categories, reflecting both larger populations and structural conditions conducive to unsheltered homelessness. The overall distribution is right-skewed (median = 136, mean = 468), consistent with the observed distribution of unsheltered counts.

health provider availability (12.6 percent), emergency shelter bed capacity (10.4 percent), transitional housing beds (7.5 percent), and permanent supportive housing beds (6.8 percent). Together, these five variables account for over 60 percent of the model's predictive power, underscoring the centrality of climate and service infrastructure in shaping unsheltered homelessness patterns.

Table 11: Phase 1: XGBoost Feature Importance (Top 20 Predictors)

Rank	Feature	Gain
1	January average temperature (°C)	0.233
2	Mental health providers (count)	0.126
3	Emergency shelter beds	0.104
4	Transitional housing beds	0.075
5	Permanent supportive housing beds	0.068
6	% Hispanic population	0.051
7	% Black population	0.040
8	% Divorced	0.025
9	Uninsured rate (%)	0.021
10	Median rent (USD)	0.019
11	January precipitation (mm)	0.017
12	% Veteran population	0.016
13	% Households without vehicle	0.015
14	% Moved from different state	0.014
15	Rental vacancy rate (%)	0.012
16	Population age 18–24 (%)	0.012
17	% Rent-burdened households	0.011
18	% One-person households	0.011
19	Population age 65+ (%)	0.011
20	Disability rate (%)	0.011

Note: Feature importance ranked by total gain contribution in the Phase 1 XGBoost model. Gain measures the improvement in prediction accuracy from splits on each variable, summed across all trees. Top 20 of 38 total features shown. January temperature dominates with 23.3% of total gain.

Figure 8 displays the top 20 features by gain, and Figure 6 presents SHAP values for the full feature set. SHAP values decompose each prediction into feature-specific contributions, revealing both the direction and magnitude of each variable's effect on the model's output for individual observations. The SHAP beeswarm plot reveals directional effects consistent with

the homelessness literature: higher January temperatures, more emergency shelter beds, and larger shares of Hispanic populations are associated with higher predicted unsheltered counts, while higher mental health provider density, more permanent supportive housing, and higher poverty rates are associated with lower counts. The negative association with poverty is initially counterintuitive but likely reflects the model learning that CoCs with extreme poverty concentrate in rural areas with small unsheltered populations, while large unsheltered populations concentrate in high-cost urban areas where poverty rates are moderate. Similarly, the positive association between emergency shelter beds and unsheltered counts likely reflects reverse causality in the training data: jurisdictions build more shelter capacity in response to large unsheltered populations, not the other way around.

Table 12 illustrates Phase 1 prediction accuracy across the CoC distribution with a random sample of five CoCs from each urbanicity category. The model captures variation reasonably well, with errors ranging from near-perfect predictions (Miami-Dade: 892 actual vs. 981 predicted) to larger misses in smaller jurisdictions (New Mexico Balance of State: 362 actual vs. 753 predicted). These examples confirm the pattern in Table 9: proportional errors are larger in smaller CoCs, while absolute errors are larger in major cities.

The final Phase 1 model, retrained on 2015–2019 data, generates baseline 2021 predictions for all 376 CoCs. These predictions range from 1 to 26,830 (median = 136, mean = 468), closely matching the distribution of observed unsheltered counts in pre-pandemic years. For the 146 CoCs that conducted complete 2021 counts, the residual (the difference between the observed count and the Phase 1 baseline prediction) isolates the COVID-specific component of the 2021 outcome. This residual becomes the dependent variable in Phase 2, to which I now turn.

4.2 Phase 2: COVID Adjustment Model

The Phase 2 residual model addresses the central challenge of the 2021 data gap: the 146 CoCs that conducted complete unsheltered counts are not representative of the national

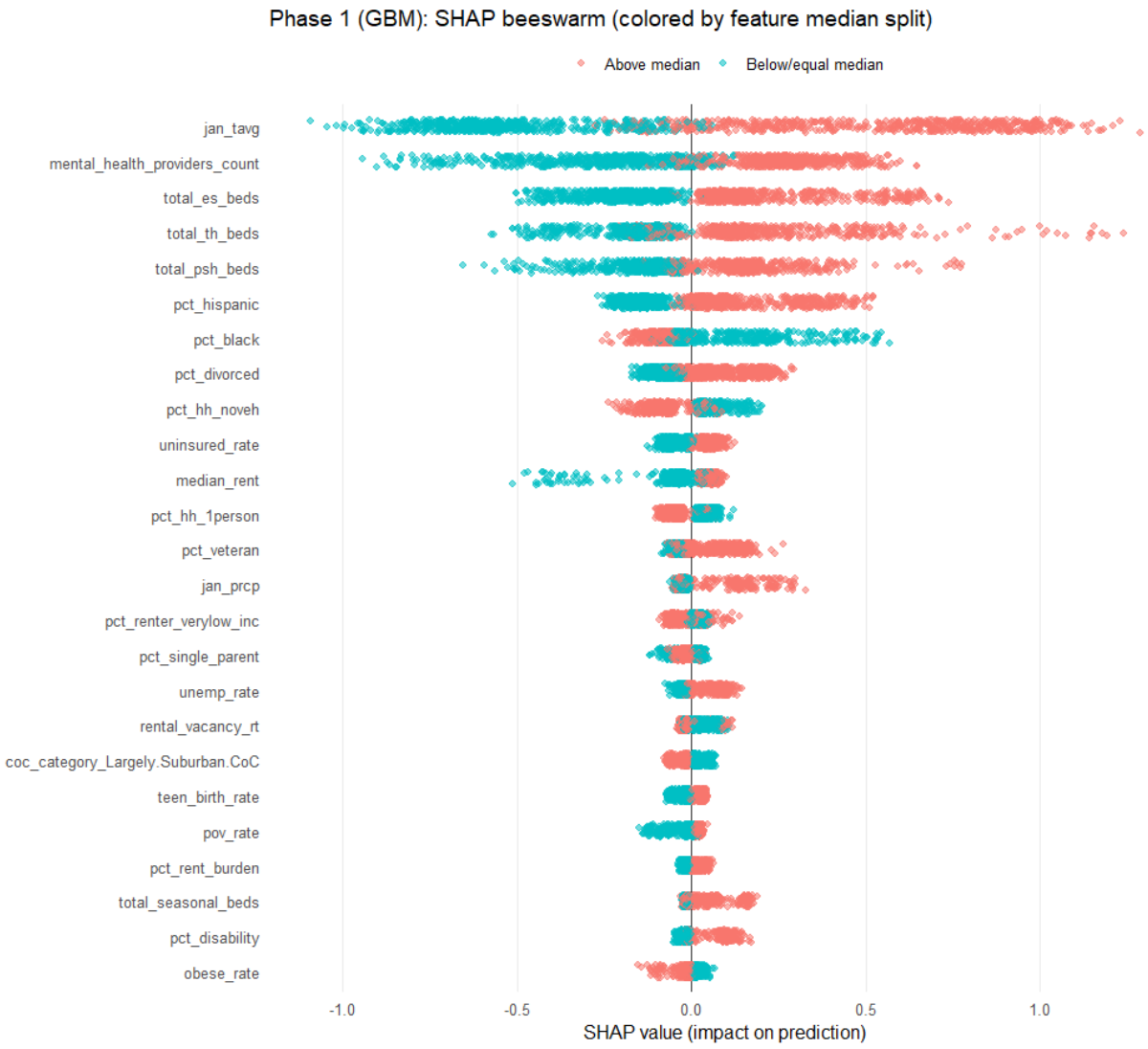

Figure 6: Phase 1 SHAP Values

Table 12: Phase 1 Baseline Predictions: Random Sample by CoC Category

CoC ID	CoC Name	Actual 2021	Predicted 2021
Largely Rural CoC			
AL-507	Alabama Balance of State	98	168
CO-500	Colorado Balance of State	528	359
MI-511	Lenawee County	0	16
NM-501	New Mexico Balance of State	362	753
TX-604	Waco/McLennan County	41	119
Largely Suburban CoC			
MA-506	Worcester City & County	247	148
MD-506	Carroll County	18	21
NJ-504	Newark/Essex County	91	452
NY-500	Rochester, Irondequoit, Greece/Monroe County	33	55
NY-505	Syracuse, Auburn/Onondaga, Oswego, Cayuga Counties	12	25
Major City CoC			
FL-510	Jacksonville-Duval, Clay Counties	230	476
FL-600	Miami-Dade County	892	981
KY-501	Louisville-Jefferson County	257	229
OK-501	Tulsa City & County	287	375
TX-601	Fort Worth/Arlington/Tarrant County	479	902
Other Largely Urban CoC			
GA-505	Columbus-Muscogee	27	66
IA-500	Sioux City/Dakota, Woodbury Counties	16	10
IL-516	Decatur/Macon County	31	28
VA-505	Newport News/Hampton/Virginia Peninsula	30	90
VA-600	Arlington County	27	33

Note: Random sample of 5 CoCs per urbanicity category from the 146 CoCs that conducted complete unsheltered counts in 2021. Actual 2021 = observed January 2021 PIT unsheltered count. Predicted 2021 = Phase 1 baseline prediction using 2020 community characteristics and the log-scale XGBoost model. Sample selected via random sampling within category.

landscape. CoCs that counted differ systematically from those that did not, creating a covariate shift problem that would bias naive extrapolation of the residual pattern. I address this through overlap-based propensity score weighting, reweighting the 146 full-count CoCs to resemble the covariate distribution of the 230 non-counting CoCs. After weighting, standardized mean differences on included covariates fall below 0.1 for all but 6 predictors, indicating drastically improved balance (see Appendix Figure 9). The effective sample size after weighting is 42.8 (Table 13), reflecting moderate efficiency loss from reweighting but preserving sufficient information for residual modeling.

Table 13: Phase 2: Overlap Weights Distribution (Treated CoCs)

N	Min	5th %	Median	95th %	Max	ESS
146	0.034	0.333	1.02	4.32	17.8	42.8

Note: Distribution of overlap-based propensity score weights $w_i = (1 - \widehat{p}_i)/\widehat{p}_i$ for the 146 treated CoCs (complete 2021 unsheltered counts) used in Phase 2 training. ESS = effective sample size, calculated as $(\sum w_i)^2 / \sum w_i^2$. The ESS of 42.8 reflects moderate efficiency loss from reweighting but preserves sufficient information for residual modeling.

I fit five weighted candidate models to predict the residual as a function of COVID-specific variables: mortality burden, policy stringency, emergency relief funding, eviction policy, and economic stress. Each model is trained on 110 randomly sampled CoCs and their performance is evaluated on the remaining 36 held-out CoCs. Table 14 reports residual prediction performance on the held-out sample. XGBoost performs well again, achieving an RMSE of 181 and explaining 47.8 percent of residual variation. The next-best model (gradient boosting: RMSE = 252, $R^2 = 0.369$) performs substantially worse, and the linear models fail entirely: weighted LASSO and OLS produce negative R^2 values, indicating they predict worse than a horizontal line through the residual mean. The baseline-only benchmark (predicting zero adjustment for all CoCs) yields an RMSE of 318, confirming that the Phase 2 model meaningfully improves over naive baseline predictions. The sample size constraint (N = 146, ESS = 42.8 after weighting) limits the complexity of models that can be reliably

estimated, but XGBoost's strong performance suggests the COVID adjustment pattern is sufficiently systematic to be learned even with a small effective sample. The summary statistics of the residual predictions can be seen in Table 15.

Table 14: Phase 2: COVID Adjustment Model Comparison

Model	RMSE	MAE	R^2
Phase 2 XGBoost	181	54	0.478
Phase 2 GBM	252	82	0.369
Phase 2 Random Forest	276	74	0.245
Phase 2 LASSO	339	192	−0.136
Phase 2 OLS	374	182	−0.386

Note: Each model is trained on 75% (110) randomly sampled CoCs, and the performance metrics above are computed from evaluation on the remaining 25% (36) CoCs. All models use overlap-based propensity score weights from the model fitted on Table 5. Dependent variable is $r_{i,2021} = y^{\text{obs}}_{i,2021} - \widehat{y}^{(1)}_{i,2021}$ (actual minus Phase 1 baseline). Negative R^2 indicates the model performs worse than predicting the mean. XGBoost selected as Phase 2 model.

Table 15: Phase 2 Predicted COVID Adjustments: Summary Statistics

	N	Mean	SD	Min	Median	Max
All Imputed CoCs	230	88	332	−1,068	−8	1,288
By CoC Category						
Major City CoC	30	146	621	−1,068	−66	1,288
Other Largely Urban CoC	31	123	227	−279	30	762
Largely Suburban CoC	92	109	322	−393	−16	1,191
Largely Rural CoC	77	26	185	−766	−1	682

Note: Summary statistics for Phase 2 predicted residuals (COVID adjustments) for the 230 CoCs with missing 2021 unsheltered counts. Residuals represent the predicted deviation from Phase 1 baseline due to COVID-19 effects. Negative values indicate predicted reductions in unsheltered homelessness relative to pre-pandemic baselines. Of the 230 imputed residuals, 122 (53%) are negative. Major City CoCs show the largest average positive adjustments (mean = 146), while Largely Rural CoCs show the smallest (mean = 26). The near-zero overall median (−8) indicates COVID-era forces pushed unsheltered counts both above and below baseline expectations with roughly equal frequency.

Table 16: Phase 2 COVID Adjustment Predictions: Random Sample by CoC Category

CoC ID	CoC Name	Baseline	Actual 2021	Actual Resid.	Pred. Resid.
Largely Rural CoC					
AL-507	Alabama Balance of State	168	98	−70	−98
CO-500	Colorado Balance of State	359	528	169	68
MI-511	Lenawee County	16	0	−16	−37
NM-501	New Mexico Balance of State	753	362	−391	−364
TX-604	Waco/McLennan County	119	41	−78	−68
Largely Suburban CoC					
MA-506	Worcester City & County	148	247	99	77
MD-506	Carroll County	21	18	−3	8
NJ-504	Newark/Essex County	452	91	−361	−456
NY-500	Rochester/Monroe County	56	33	−22	2
NY-505	Syracuse/Onondaga Counties	25	12	−13	−2
Major City CoC					
FL-510	Jacksonville-Duval Counties	476	230	−246	−207
FL-600	Miami-Dade County	981	892	−89	−122
KY-501	Louisville-Jefferson County	229	257	28	20
OK-501	Tulsa City & County	375	287	−88	−103
TX-601	Fort Worth/Tarrant County	902	479	−423	−474
Other Largely Urban CoC					
GA-505	Columbus-Muscogee	66	27	−39	−51
IA-500	Sioux City/Woodbury Counties	10	16	6	19
IL-516	VA-505 Newport News/Virginia Peninsula	90	30	−60	−44
VA-600	Arlington County	33	27	−6	−30

Note: Random sample of 5 CoCs per urbanicity category from the 146 CoCs that conducted complete 2021 unsheltered counts. Baseline = Phase 1 prediction using pre-pandemic relationships. Actual 2021 = observed January 2021 PIT unsheltered count. Actual Resid. = Actual 2021 − Baseline (true COVID adjustment). Pred. Resid. = Phase 2 XGBoost predicted COVID adjustment. The relationship Actual 2021 = Baseline + Actual Resid. holds by construction. The Phase 2 model achieves solid residual predictions in this sample (overall RMSE = 181, $R^2 = 0.478$). Sample selected via random sampling within category.

What explains the COVID-specific deviations from baseline? Table 17 reports feature importance rankings for the Phase 2 XGBoost model. Emergency relief funding dominates: total ESG-CARES allocations account for 57.6 percent of predictive gain, more than four times the contribution of the next most important variable. COVID mortality burden ranks second (12.4 percent of gain), followed by SNAP participation (9.5 percent), policy stringency (8.3 percent), and unemployment (7.6 percent). Local eviction moratoria and Emergency Rental Assistance programs contribute minimally to the model (under 1 percent each), likely because these policies were widespread and exhibited limited cross-sectional variation in 2021. Figure 10 displays the full feature importance ranking.

Table 17: Phase 2: XGBoost Feature Importance (COVID Adjustment Model)

Rank	Feature	Gain
1	ESG-CARES funding (USD)	0.576
2	COVID-19 deaths (2021)	0.124
3	Households with SNAP (%)	0.095
4	Average policy stringency score	0.083
5	Unemployment rate (%)	0.076
6	Has Emergency Rental Assistance	0.009
7	Has eviction moratorium	0.009
8	CoC category: Major City	0.009
9	CoC category: Largely Suburban	0.009
10	CoC category: Largely Rural	0.006
11	CoC category: Other Largely Urban	0.002
12	Has mask mandate	0.001

Note: Feature importance ranked by total gain contribution in the Phase 2 XGBoost residual model. ESG-CARES emergency shelter funding dominates with 57.6% of total gain, more than four times the contribution of the next most important variable (COVID-19 deaths). Local eviction moratoria and Emergency Rental Assistance contribute minimally, likely due to limited cross-sectional variation in 2021.

The Phase 2 model generates residual predictions for all 230 CoCs that did not conduct complete unsheltered counts. The predicted adjustments range from $-1,068$ to 1,288 (median $= -8$, mean $= 88$), closely matching the residual distribution in the training sample. However, 122 of the 230 imputed residuals are negative, and for 26 CoCs the predicted neg-

ative adjustment exceeds the Phase 1 baseline prediction, yielding a negative final count. These predictions are clamped at zero, as unsheltered populations cannot be negative, but the prevalence of large negative adjustments is a limitation worth noting. The model appears to learn that CoCs with very low baseline unsheltered populations and high emergency relief funding experienced sharp reductions in 2021, but in cases where the baseline was already near zero, the model extrapolates implausibly. This affects primarily small rural CoCs (e.g., NY-514, IL-512) and a handful of suburban jurisdictions (e.g., NJ-514, CA-612).

4.3 Final Imputed Unsheltered Counts and National Estimates

The two-phase imputation framework produces a national 2021 unsheltered estimate of 195,191 individuals (90% prediction interval: [114,380, 255,978]), combining 146 observed complete counts with imputed estimates for 230 CoCs that did not conduct complete unsheltered counts. This represents the first comprehensive accounting of unsheltered homelessness during the COVID-19 pandemic, filling a critical gap in the administrative data record and enabling longitudinal analysis of homelessness trends through the crisis period.

Table 18 presents the 2015–2022 unsheltered time series. To ensure comparability across years, the observed totals for all non-2020 years are restricted to the consistent sample of 376 non-territory CoCs included in the 2021 analysis. The 2021 imputed estimate falls 13.7 percent below the 2020 observed count of 226,080, suggesting that emergency relief measures continued to suppress unsheltered homelessness through early 2021 even as their coverage began to erode, such as the expanded shelter capacity funded by ESG-CARES, eviction moratoria, and Emergency Rental Assistance. This decline reverses sharply in 2022, when the consistent-sample count surges 25.2 percent to 244,320 as emergency protections fully expired. The 2022 count lies within the upper portion of the 2021 prediction interval, confirming that the two-phase procedure captures the broad trajectory: emergency relief temporarily suppressed unsheltered homelessness below pre-pandemic baselines, but the effect proved transient once protections lapsed.

Table 18: National Unsheltered Homelessness Estimates, 2015–2022

Year	Total Unsheltered	Interval	YoY Change	YoY % Change
2015	166,322		—	—
2016	176,900		10,578	6.4
2017	187,835		10,935	6.2
2018	197,215		9,380	5.0
2019	215,682		18,467	9.4
2020	226,080		10,398	4.8
2021	195,191	[114,380, 255,978]	−30,889	−13.7
2022	244,320		49,129	25.2

Note: National unsheltered PIT counts restricted to a consistent sample of 376 non-territory CoCs present in the 2021 analysis to ensure comparability across years, with the exception of 2020, which uses official HUD national totals. 2021 estimate combines observed counts from 146 CoCs with imputed estimates for 230 CoCs; 90% prediction interval shown in brackets. YoY = year-over-year.

Figure 7 displays the time series visually. The 2021 imputed estimate (shown with 90% prediction interval) falls substantially below 2020, consistent with emergency relief measures suppressing unsheltered homelessness through early 2021. The subsequent 2022 surge places the observed count within the upper range of the 2021 prediction interval, validating the imputed estimate as a plausible midpoint in the post-pandemic trajectory and reinforcing the interpretation that pandemic-era protections produced a temporary, not permanent, reduction in unsheltered homelessness.

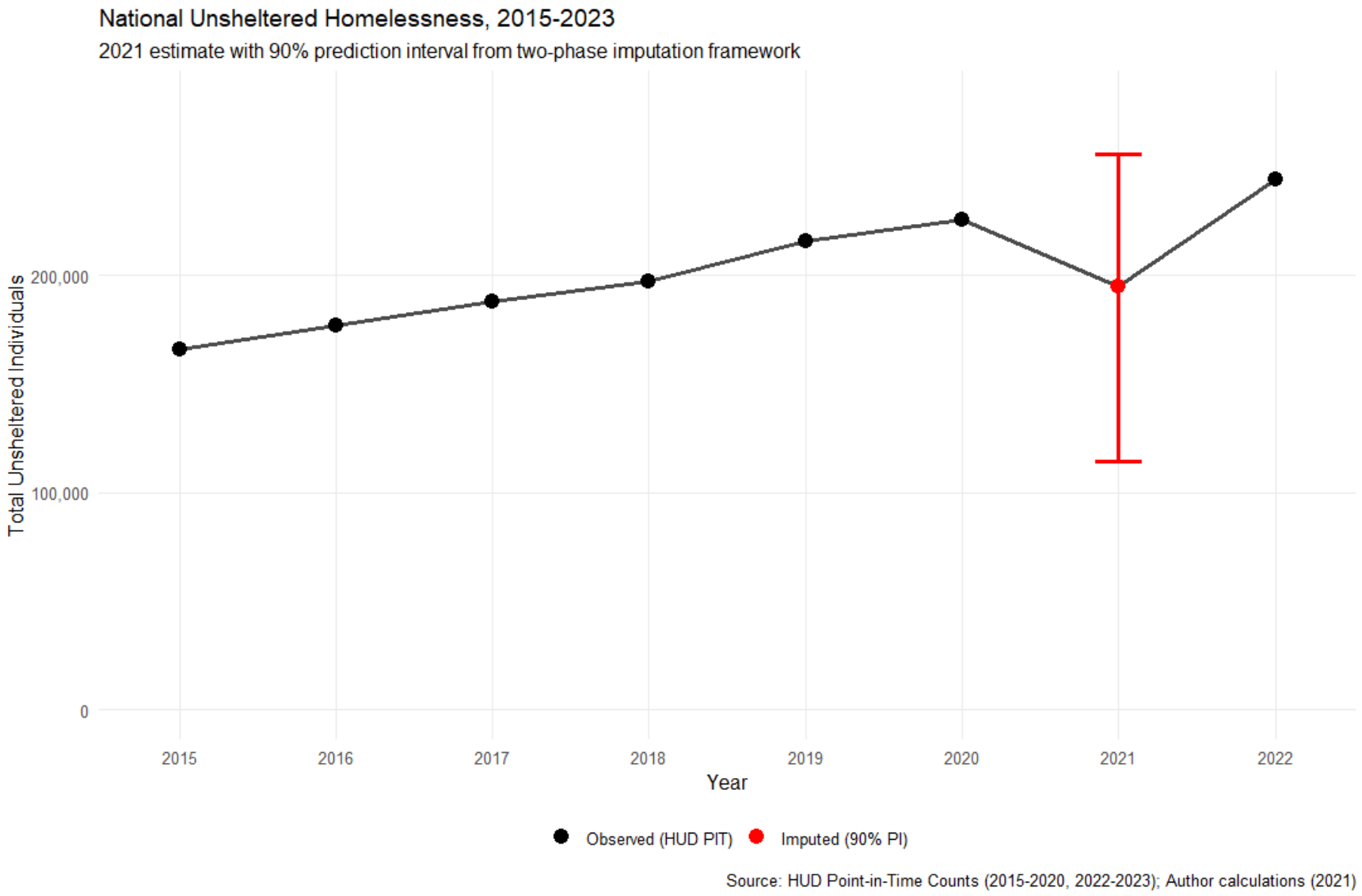

Figure 7: National Unsheltered Homelessness Time Series, 2015–2022

The imputed unsheltered population concentrates heavily in California and major cities. Los Angeles (CA-600) alone accounts for an estimated 28,038 unsheltered individuals—more than the combined total of the bottom 150 CoCs—with a 90% prediction interval spanning [4,466, 35,180]. The interval width of 30,714 reflects extreme uncertainty in both the baseline prediction and the COVID adjustment for the nation's largest unsheltered population, where volatile housing markets and heterogeneous policy responses resist precise prediction. Beyond Los Angeles, California dominates the top of the distribution: nine of the ten largest imputed unsheltered populations are in California, and fourteen of the thirty largest overall (Table 19). This concentration underscores the geographic inequality of unsheltered homelessness in the United States, where a handful of jurisdictions bear a disproportionate share of the crisis.

Prediction intervals quantify the uncertainty inherent in the imputed estimates and vary dramatically across the CoC distribution. The median 90% prediction interval width is 259, indicating moderate uncertainty for typical CoCs, but the distribution is severely right-skewed: the mean width is 671, and the maximum reaches 30,714 (Table 20). Uncertainty varies systematically by CoC type. Major City CoCs exhibit the widest intervals (median width = 813), reflecting both larger population scales and greater volatility in unsheltered counts, where housing markets are tighter and policy responses more heterogeneous. Largely Rural and Largely Suburban CoCs have narrower intervals (median widths of 221 and 222, respectively), consistent with smaller baseline populations and more stable COVID effects.

The widest prediction intervals concentrate overwhelmingly in California and the West (Table 21). Los Angeles dominates with an interval width of 30,714, an order of magnitude larger than any other CoC. The remaining wide intervals cluster in California major cities (San Jose, Oakland, San Francisco, Sacramento, San Diego) and western urban centers (Seattle, Las Vegas), where large baseline unsheltered populations combine with substantial COVID adjustments to produce extreme uncertainty. The narrowest intervals, conversely, concentrate in small suburban and rural CoCs in the Northeast and Midwest (Table 22),

Table 19: Top 30 CoCs by Imputed 2021 Unsheltered Count

Rank	CoC ID	Category	Median Estimate
1	CA-600	Major City	28,038
2	CA-502	Major City	5,316
3	CA-601	Major City	5,046
4	CA-503	Major City	4,392
5	CA-500	Major City	4,417
6	CA-602	Largely Suburban	4,386
7	CA-501	Major City	4,279
8	TX-607	Largely Rural	3,859
9	CA-608	Largely Suburban	3,166
10	WA-500	Major City	3,601
11	CA-514	Major City	2,877
12	NY-600	Major City	2,383
13	OR-501	Major City	2,376
14	WA-501	Largely Rural	2,202
15	NV-500	Major City	2,142
16	HI-501	Largely Suburban	1,842
17	TX-503	Major City	2,238
18	CA-510	Largely Suburban	2,201
19	CA-504	Other Urban	2,180
20	CA-506	Largely Rural	2,167
21	CA-604	Major City	1,995
22	OR-500	Other Urban	1,990
23	GA-501	Largely Suburban	1,897
24	CA-611	Other Urban	1,830
25	CA-508	Largely Suburban	1,694
26	CA-603	Largely Suburban	1,628
27	AZ-502	Major City	2,823
28	CA-511	Largely Suburban	2,384
29	CA-505	Largely Suburban	2,350
30	CA-609	Largely Suburban	3,391

Note: Thirty largest imputed 2021 unsheltered populations (bootstrap median). Los Angeles accounts for 28,038 individuals. Twenty-three of thirty are in California or the West. Category labels abbreviated. Mix of observed and imputed counts; table focuses on imputed CoCs to illustrate geographic concentration.

Table 20: Prediction Interval Widths by CoC Category

CoC Category	N	Median Width	Mean Width	Max Width
Major City CoC	30	813	2,335	30,714
Other Largely Urban CoC	31	245	326	725
Largely Suburban CoC	92	222	367	1,913
Largely Rural CoC	77	221	359	4,046
Overall	230	259	671	30,714

Note: Distribution of 90% prediction interval widths for the 230 imputed CoCs, stratified by urbanicity category. Width = $p_{95} - p_{05}$ from end-to-end clustered bootstrap (B = 300). Major City CoCs exhibit substantially wider intervals (median = 813) than other categories, reflecting both larger population scales and greater volatility in COVID effects.

where baseline predictions are low and COVID adjustments near zero. One CoC (NY-523: Glens Falls, Saratoga Springs) has a degenerate interval—width of zero—because the bootstrap consistently predicts zero unsheltered individuals in every draw, reflecting a small baseline and large predicted negative COVID adjustment that the zero-clamping enforces uniformly.

Table 21: CoCs with Widest 90% Prediction Intervals

CoC ID	CoC Name	Category	p_{50}	p_{95}	Width
CA-600	Los Angeles City & County	Major City	28,038	35,180	30,714
CA-500	San Jose/Santa Clara	Major City	4,417	6,906	4,707
CA-502	Oakland, Berkeley/Alameda	Major City	5,316	6,734	4,110
TX-607	Texas Balance of State	Largely Rural	3,859	5,107	4,046
WA-500	Seattle/King County	Major City	3,601	4,973	4,013
CA-501	San Francisco	Major City	4,279	5,717	3,517
NV-500	Las Vegas/Clark County	Major City	2,142	3,027	2,364
CA-503	Sacramento City & County	Major City	4,392	5,175	2,247
CA-601	San Diego City and County	Major City	5,046	6,008	1,949
CA-602	Orange County	Largely Suburban	4,386	5,438	1,913

Note: Ten CoCs with widest 90% prediction intervals. Los Angeles dominates with width of 30,714. Nine of ten are in California or the West. Width = $p_{95} - p_{05}$. CoC names abbreviated for space.

An important caveat warrants emphasis: these are prediction intervals, not confidence intervals. Prediction intervals quantify uncertainty in individual CoC-level imputations and include both parameter estimation error and irreducible forecast error. They cannot be

Table 22: CoCs with Narrowest 90% Prediction Intervals

CoC ID	CoC Name	Category	p_{50}	p_{95}	Width
NY-523	Glens Falls, Saratoga Springs	Largely Suburban	0	0	0
MN-508	Moorhead/West Central MN	Largely Rural	0	6	6
MI-523	Eaton County	Largely Suburban	0	17	17
PA-603	Beaver County	Largely Suburban	0	23	23
MN-511	Southwest Minnesota	Largely Rural	0	28	28
IL-508	East St. Louis, Belleville	Largely Suburban	0	32	32
IL-502	Waukegan, North Chicago	Largely Suburban	0	33	33
NY-520	Franklin, Essex Counties	Largely Rural	0	35	35
NY-525	New York Balance of State	Largely Rural	0	36	36
MD-504	Howard County	Largely Suburban	11	37	37

Note: Ten CoCs with narrowest 90% prediction intervals. NY-523 has degenerate interval (width = 0) because bootstrap consistently predicts zero. Narrow intervals concentrate in small suburban and rural CoCs in Northeast and Midwest. CoC names abbreviated.

aggregated across CoCs without additional assumptions. The national-level interval reported above (90% PI: [114,380, 255,978]) is constructed by summing the 5th and 95th percentiles of the bootstrap distribution across all imputed CoCs, which treats CoC-level prediction errors as independent. This is conservative but likely overstates national uncertainty, since common shocks—federal relief policy, nationwide policy stringency—induce positive correlation in prediction errors across CoCs, partially offsetting in aggregate. Nonetheless, even under conservative assumptions, the 2021 imputed national unsheltered count is precise enough to support policy analysis: the interval spans 141,598 individuals, or 73 percent of the point estimate, comparable to the uncertainty in large-scale household surveys and substantially narrower than the alternative of dropping 2021 from longitudinal analyses entirely.

5 Discussion

The 2021 Point-in-Time count, disrupted by COVID-19, left 61.6 percent of Continuums of Care without complete unsheltered enumeration data. This paper develops a two-phase machine learning framework to fill that gap, producing the first comprehensive national estimate of unsheltered homelessness during the pandemic. The imputed national unshel-

tered count of 195,191 individuals (90% prediction interval: [114,380, 255,978]), combined with the observed sheltered count of 326,126, yields a reconstructed total of 521,317 people experiencing homelessness in January 2021. Beyond filling a critical data gap, the methodology demonstrates how to handle missing administrative data when both the missingness mechanism and the outcome process are altered by crisis events.

5.1 Key Findings

Three substantive findings emerge from the analysis. First, emergency relief funding overwhelmingly explains how the pandemic reshaped local unsheltered homelessness. ESG-CARES allocations account for 57.6 percent of the Phase 2 model's predictive gain, more than four times the next most important variable. Meanwhile, COVID mortality, policy stringency, and economic stress play comparatively minor roles. The implication is that direct shelter funding, not the severity of the health crisis itself, determined whether a community's unsheltered population grew or shrank relative to pre-pandemic levels. This extends evidence from the Great Recession (Popov, 2016) into a fundamentally different crisis setting: whereas conventional recessions displace people through job loss and foreclosure, the pandemic simultaneously destroyed livelihoods and deployed unprecedented countervailing resources. That targeted relief appears to have buffered unsheltered homelessness even under these conditions suggests emergency shelter investment operates through channels robust to the specific mechanism of economic disruption.

Second, the pandemic amplified the geographic concentration of unsheltered homelessness rather than dispersing it. California jurisdictions account for nine of the ten largest imputed unsheltered populations, and Los Angeles alone exceeds the combined total of the bottom 150 CoCs. This pattern is not simply an artifact of California's pre-existing crisis. ESG-CARES allocations followed existing CoC grant formulas that weight population and poverty but do not scale with baseline unsheltered severity, meaning that jurisdictions already bearing outsized burdens received relief proportional to their general need rather than their crisis

magnitude. The result is a federal funding structure that, during an emergency, mechanically widens the gap between high-burden and low-burden communities.

Third, the imputed 2021 estimate sits plausibly within the trajectory defined by adjacent observed counts. The national unsheltered total declines 13.7 percent from 2020, consistent with emergency protections still operating in early 2021, then rebounds sharply by 2022 as those protections expired. This V-shaped pattern is difficult to reconcile with alternative explanations such as measurement artifact or model overfitting, since it aligns with the known timeline of federal relief expiration and independently corroborates the Phase 2 finding that emergency funding drove short-run reductions. The 2022 observed count falls within the upper portion of the 2021 prediction interval, providing external validation that the two-phase framework captures the correct order of magnitude and directional trend even if point estimates carry substantial uncertainty.

5.2 Methodological Contributions

The central methodological contribution is a framework for imputing missing administrative data when the standard missing-at-random assumption fails in two simultaneous ways: the shock that causes missingness also changes the quantity being measured, and the units that do report are not representative of those that do not. Neither problem alone is unusual: selection models address non-random missingness, and structural break methods handle regime changes. However, their combination creates a setting where no single-stage approach suffices. A model trained on pre-crisis data recovers the counterfactual, not the actual outcome. A model trained on crisis-year responders extrapolates from a biased sample. The two-phase decomposition addresses both problems by assigning each to the estimation stage where it can be most credibly handled.

This decomposition carries a broader lesson for applied work with disrupted administrative data. The logic does not depend on the specific context of homelessness counts or COVID-19. Any setting where a crisis event simultaneously interrupts routine data col-

lection and alters the data-generating process shares the same structure: vital statistics registration after natural disasters, census enumeration during civil conflict, and labor force surveys suspended by pandemics. In each case, pre-disruption panel data can anchor the counterfactual, and partial observations from the disruption period can identify the deviation. The two-phase template is agnostic to the domain; what matters is that the researcher can credibly separate stable structural relationships from transient shock effects and has at least some crisis-year observations to learn from.

Three technical choices strengthen the framework's credibility in practice. The overlap-based propensity score weighting in Phase 2 directly confronts the selection problem: CoCs that conducted full counts in 2021 skew smaller, less urban, and less affected by COVID restrictions than those that did not, and naive residual extrapolation would systematically mischaracterize the adjustment for non-counting CoCs. Reweighting cannot eliminate bias from unobservables, but balance diagnostics confirm that it substantially reduces observable covariate shift. The end-to-end clustered bootstrap treats the entire two-stage pipeline as a single composite estimator, propagating uncertainty from Phase 1 into Phase 2 rather than conditioning on baseline predictions as though they were known. This is conservative by design, as it produces wider intervals than stage-wise approaches, but it avoids the well-documented problem of understating imputation uncertainty when estimation error in early stages is ignored (Rubin, 1987). Finally, the systematic model comparison across linear and tree-based methods disciplines the functional form choice. XGBoost dominates in both phases, but the margin of improvement over linear models differs starkly: modest in Phase 2 (where the small effective sample constrains complexity) and dramatic in Phase 1 (where nonlinearities in climate, shelter capacity, and demographics are strong). Reporting the full model comparison makes this choice transparent rather than stipulated.

5.3 Limitations

Five limitations warrant emphasis. First, selection bias in Phase 2 is mitigated but not eliminated. Propensity score reweighting reduces observable covariate imbalance between counting and non-counting CoCs, but the effective sample size of 42.8 after weighting signals that only a fraction of the 146 full-count CoCs carry meaningful weight in the residual model. More fundamentally, the decision to conduct an unsheltered count during a pandemic likely reflects unobserved institutional characteristics that plausibly correlate with how effectively a community deployed emergency resources, such as administrative capacity, political will, volunteer infrastructure, organizational culture around data collection. If CoCs that counted also happened to manage pandemic response more competently, the residual model would learn an adjustment pattern systematically different from what non-counting CoCs actually experienced, and no amount of reweighting on observables can correct this.

Second, the imputed estimates are modeled quantities, not observed data, and should be treated accordingly. The national prediction interval spans roughly 73 percent of the point estimate, reflecting genuine uncertainty that cannot be wished away through methodological sophistication. Researchers incorporating these estimates into downstream analyses—panel regressions, trend decompositions, funding simulations—should propagate this uncertainty rather than treating imputed values as equivalent to observed counts. Practical approaches include conducting sensitivity analysis at the prediction interval bounds, using multiple draws from the bootstrap distribution as parallel datasets, or restricting analyses to CoCs where interval widths are narrow enough to support the inferential claims being made.

Third, the zero-clamping of negative predictions introduces asymmetric bias at the bottom of the distribution. Nineteen CoCs receive imputed counts of zero because the Phase 2 model predicts a negative adjustment larger than the Phase 1 baseline. While zero unsheltered counts do occur in PIT data (15 CoCs reported exactly zero in 2019), the clamping mechanically prevents the model from distinguishing genuine near-zero populations from cases where the additive structure of the two-phase framework overshoots. This affects

small rural and suburban CoCs where baselines are already low and ESG-CARES allocations are high relative to population, and it biases the national total modestly downward. The alternative of allowing negative counts would be incoherent, making the floor at zero a defensible, if imperfect, compromise.

Fourth, the imputation inherits every limitation of the underlying PIT count methodology. Volunteer-driven street enumeration on a single January night systematically undercounts unsheltered individuals, exhibits high measurement error, and varies in quality across jurisdictions (O'Flaherty, 2019; Meyer et al., 2023). The framework predicts what a PIT count would have found, not the true size of the unsheltered population: a distinction that matters for interpreting both the point estimates and the prediction intervals. If a CoC's pre-pandemic counts systematically undercounted by 40 percent, the imputed 2021 value will reproduce that undercount, not correct it.

Fifth, the analysis is predictive, not causal. The dominance of ESG-CARES funding in the Phase 2 model is an associational finding: communities that received more emergency shelter funding saw smaller deviations from baseline, but this correlation reflects both the direct effect of funding and the confounded selection process through which funding was allocated. CoCs receiving large ESG-CARES awards tend to have stronger existing shelter infrastructure, more experienced grant administrators, and greater political engagement with HUD: all factors that could independently suppress unsheltered homelessness during a crisis. Credible causal identification would require exploiting exogenous variation in funding, such as discontinuities in formula-based allocation rules or quasi-random timing of grant disbursement, which lies beyond the present scope.

5.4 Future Directions

The most immediate application of these estimates is substantive rather than methodological. The 2021 data gap has forced researchers studying pandemic-era homelessness to either drop the year entirely, restrict analysis to the non-representative subset of CoCs that counted, or

treat sheltered counts as a proxy for total homelessness. The imputed estimates remove this constraint. With a complete 2021 cross-section in hand, researchers can trace unsheltered homelessness trajectories through the full pandemic arc, decompose the 2022 surge into a rebound component and a new-inflow component, and test whether communities that received more emergency funding experienced durably lower homelessness or merely delayed an inevitable increase. The prediction intervals make it possible to distinguish findings that are robust to imputation uncertainty from those that depend on taking the point estimates at face value.

The framework itself invites extension in at least two directions. First, applying the same two-phase procedure to other partially disrupted PIT count years, such as sporadic missingness in non-pandemic years, would produce a more complete national panel and allow researchers to study whether the structural relationships estimated in Phase 1 remain stable over time or drift in ways that affect imputation quality. Second, the current framework treats each CoC independently in both phases. Spatial models that allow prediction errors to correlate across neighboring CoCs could tighten national uncertainty bounds, since common regional shocks induce dependence that the current bootstrap treats as independent noise. Hierarchical Bayesian approaches offer a related avenue: partial pooling across CoCs would borrow strength from data-rich jurisdictions to improve estimates for data-poor ones, a particularly attractive property when the effective sample size after propensity score weighting is small.

Beyond homelessness, the methodological template applies wherever administrative data collection is interrupted by the same event that changes the quantity being measured. The specific technical choices are modular and can be adapted to different data structures and sample sizes. What is less clear, and worth formalizing in future work, is when the two-phase decomposition actually improves over simpler alternatives. If the crisis-year deviation is small relative to baseline prediction error, a single-stage model may suffice. If the responding sample is so selected that no reweighting scheme achieves adequate balance, the Phase 2

adjustment may introduce more bias than it corrects. Developing diagnostic criteria for these boundary conditions (when to decompose and when not to) would make the framework more practically useful for applied researchers confronting the next disrupted dataset.

5.5 Conclusion

The 2021 PIT count disruption left a hole in the national homelessness record at precisely the moment that record mattered most. This paper fills it, imperfectly but transparently. The two-phase framework produces a national unsheltered estimate of 195,191 individuals (90% prediction interval: [114,380, 255,978]) and, more importantly, makes the uncertainty in that estimate explicit rather than leaving it as an unquantified absence in the data. The substantive picture that emerges is consistent with the known timeline of federal intervention and independently corroborated by adjacent observed counts: emergency relief temporarily suppressing unsheltered homelessness before a sharp post-pandemic rebound.

The deeper contribution is methodological. Missing administrative data during crises is not a one-time inconvenience; it is a recurring structural problem. Climate disasters disrupt vital statistics systems. Armed conflicts prevent census enumeration. Future pandemics will again force trade-offs between data collection and public safety. In each case, the missingness is entangled with the very phenomenon being measured, violating the assumptions on which standard imputation methods rest. The two-phase decomposition, anchoring a counterfactual in pre-crisis data and modeling the crisis-specific deviation from partial observations, offers one principled response. It will not be the only response needed, but it demonstrates that the choice between accepting a permanent data gap and pretending the gap does not exist is a false one. With appropriate uncertainty quantification and honest acknowledgment of limitations, defensible reconstruction is possible.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Burt, M. R., Aron, L. Y., Douglas, T., Valente, J., Lee, E., and Iwen, B. (1999). Homelessness: Programs and the people they serve. Findings of the National Survey of Homeless Assistance Providers and Clients.

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.

Chen, S. and Xu, C. (2025). On the use of machine learning methods for missing data problems. In *Handbook of Statistics*, volume 53, pages 161–174. Elsevier.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Chien, J., Henwood, B. F., St. Clair, P., Kwack, S., and Kuhn, R. (2024). Predicting hotspots of unsheltered homelessness using geospatial administrative data and volunteered geographic information. *Health & Place*, 88:103267.

Corinth, K. and Lucas, D. S. (2018). When warm and cold don't mix: The implications of climate for the determinants of homelessness. *Journal of Housing Economics*, 41(C):45–56.

Dang, H.-A., Kilic, T., Hlasny, V., Abanokova, K., and Carletto, C. (2026). Using survey-to-survey imputation to fill poverty data gaps at a low cost: Evidence from a randomized survey experiment. *The World Bank Economic Review*. lhaf037.

de Sousa, T., Andrichik, A., Cuellar, M., Marson, J., Prestera, E., and Rush, K. (2023). The 2022 annual homeless assessment report (AHAR) to congress. Technical report, U.S.

Department of Housing and Urban Development, Office of Community Planning and Development. Prepared by Abt Associates.

Downing, N. J. (2025). Missing value imputation in environmental, social, and governance data: An impact on emissions scores. *Finance Research Letters*, 85:107818.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, Cambridge.

Elliott, M. and Krivo, L. J. (1991). Structural determinants of homelessness in the united states. *Social Problems*, 38(1):113–131.

Embaye, W. T., Zereyesus, Y. A., and Chen, B. (2021). Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches. *PLOS ONE*, 16(2):e0244953.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Glasser, I., Hirsch, E., and Chan, A. (2014). Reaching and enumerating homeless populations: Geospatial methods and promising practices. *Cityscape*, 16(3):167–188.

Glynn, C. and Fox, E. B. (2019). Dynamics of homelessness in urban America. *The Annals of Applied Statistics*, 13(1):573–605.

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*, 27(4):1–16.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4):475–492.

Henry, M., de Sousa, T., Tano, C., Dick, N., Hull, R., Shea, M., Morris, T., and Morris, S. (2022). The 2021 annual homeless assessment report (AHAR) to congress. Technical report, U.S. Department of Housing and Urban Development, Office of Community Planning and Development. Prepared by Abt Associates.

Honig, M. and Filer, R. K. (1993). Causes of intercity variation in homelessness. *American Economic Review*, 83(1):248–255.

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

Meyer, B. D., Wyse, A., and Corinth, K. (2023). The size and census coverage of the u.s. homeless population. *Journal of Urban Economics*, 136:103559.

Meyer, B. D., Wyse, A., Grunwaldt, A., Medalia, C., and Wu, D. (2021). Learning about homelessness using linked survey and administrative data. Working Paper 28861, National Bureau of Economic Research.

Moulton, S. (2013). Does increased funding for homeless programs reduce chronic homelessness? *Southern Economic Journal*, 79(3):600–620.

Nisar, H., Vachon, M., Horseman, C., and Murdoch, J. (2019). Market predictors of homelessness: How housing and community factors shape homelessness rates within continuums of care. Technical report, U.S. Department of Housing and Urban Development.

O'Flaherty, B. (1995). An economic theory of homelessness and housing. *Journal of Housing Economics*, 4:13–49.

O'Flaherty, B. (2019). Homelessness research: A guide for economists (and friends). *Journal of Housing Economics*, 44:1–25.

Popov, I. (2016). Homeless programs and social insurance. Working paper, Stanford Institute for Economic Policy Research.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Ruhnke, S. A., Wilson, F. A., and Stimpson, J. P. (2022). Using machine learning to impute legal status of immigrants in the National Health Interview Survey. *MethodsX*, 9:101848.

Smith, A. C., Holmberg, C., and Jones-Puthoff, M. (2012). Emergency and transitional shelter population: 2010.

U.S. Department of Housing and Urban Development (2021). Notice of funding opportunity (NOFO) for fiscal year (FY) 2021 continuum of care competition and noncompetitive award of youth homeless demonstration program renewal and replacement grants. Technical Report FR-6500-N-25, U.S. Department of Housing and Urban Development, Community Planning and Development.

6 Appendix

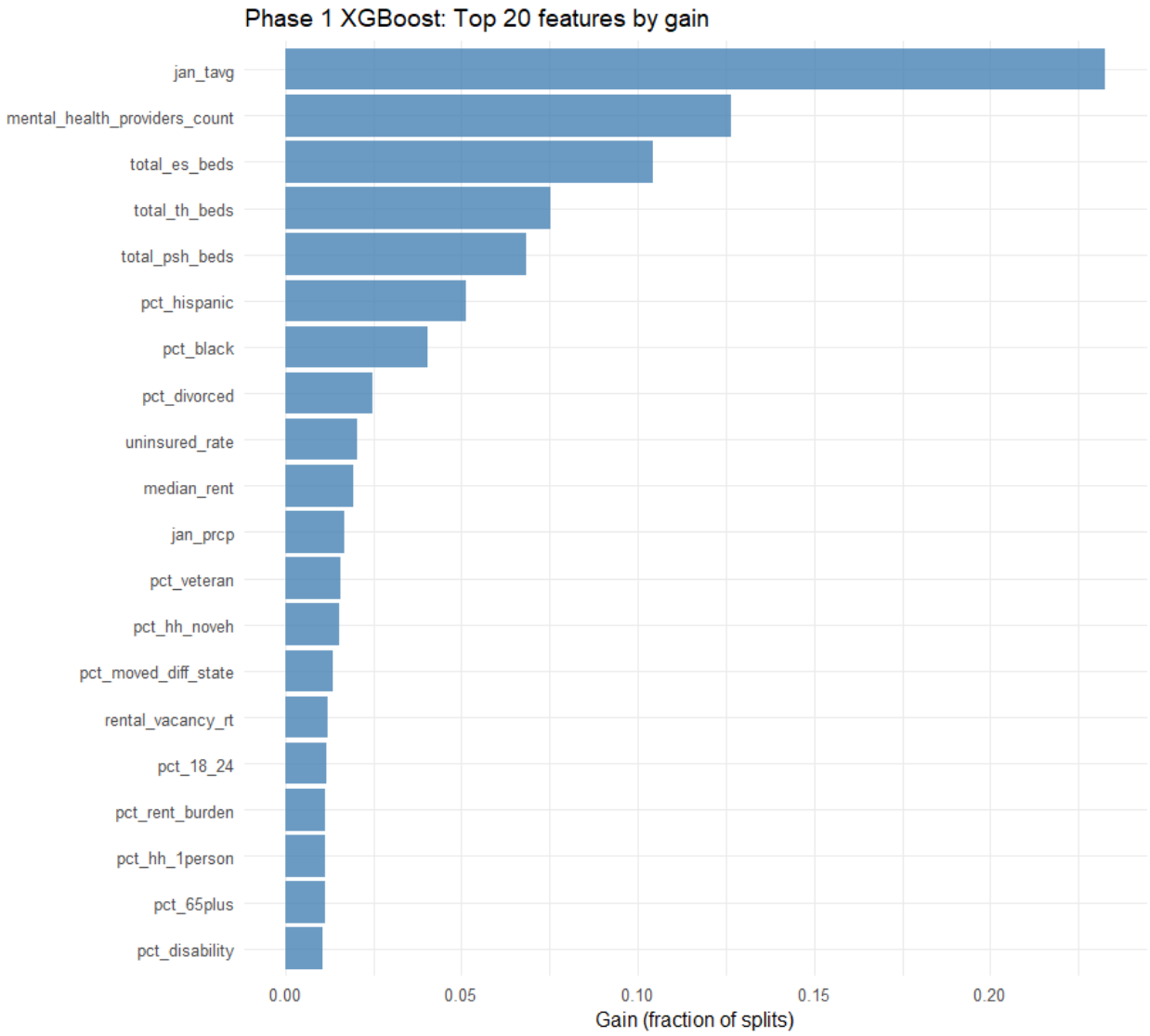

Figure 8: Phase 1 Feature Importance

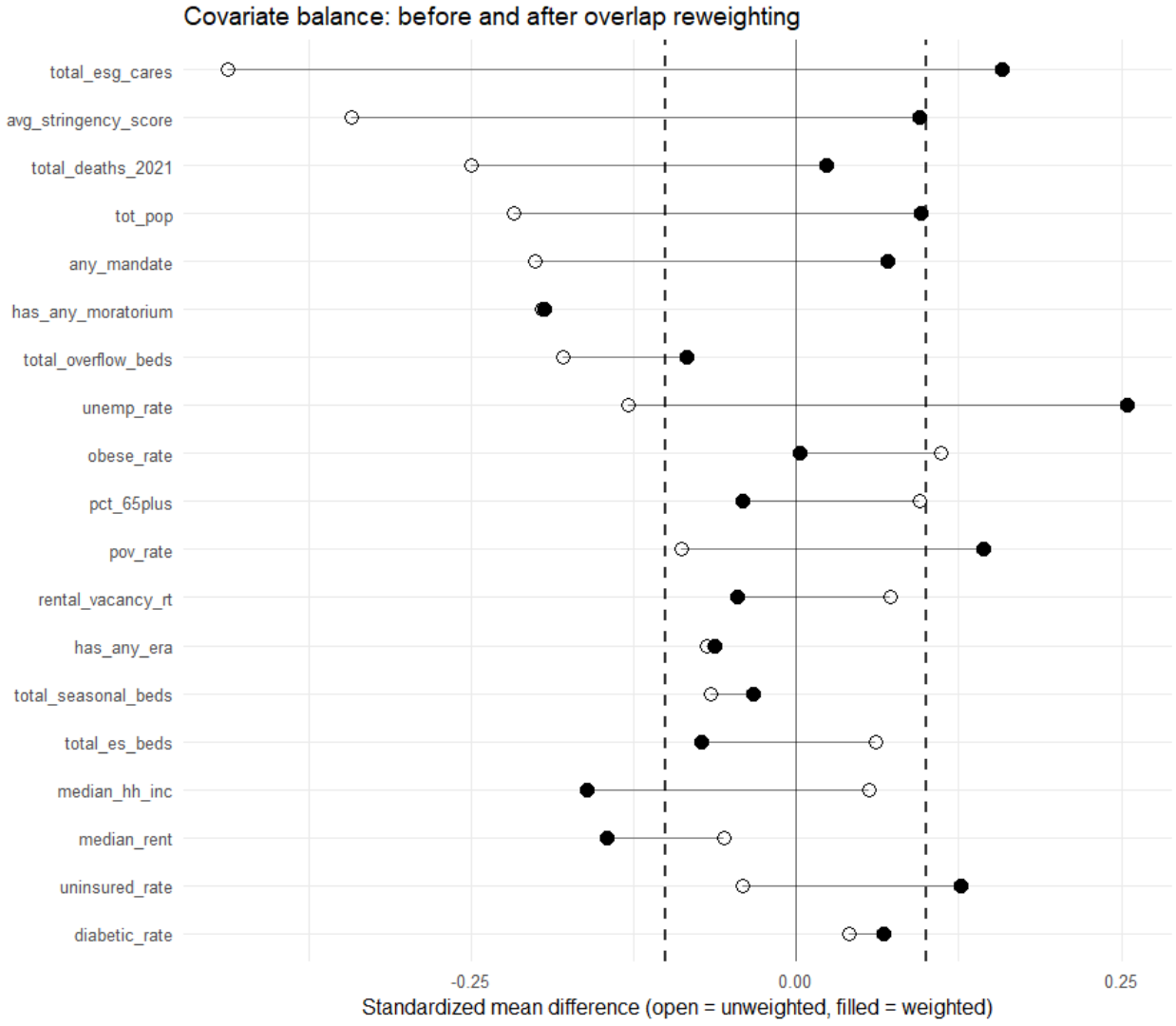

Figure 9: Love Plot of Standardized Mean Difference Before vs. After Reweighting, Phase 2

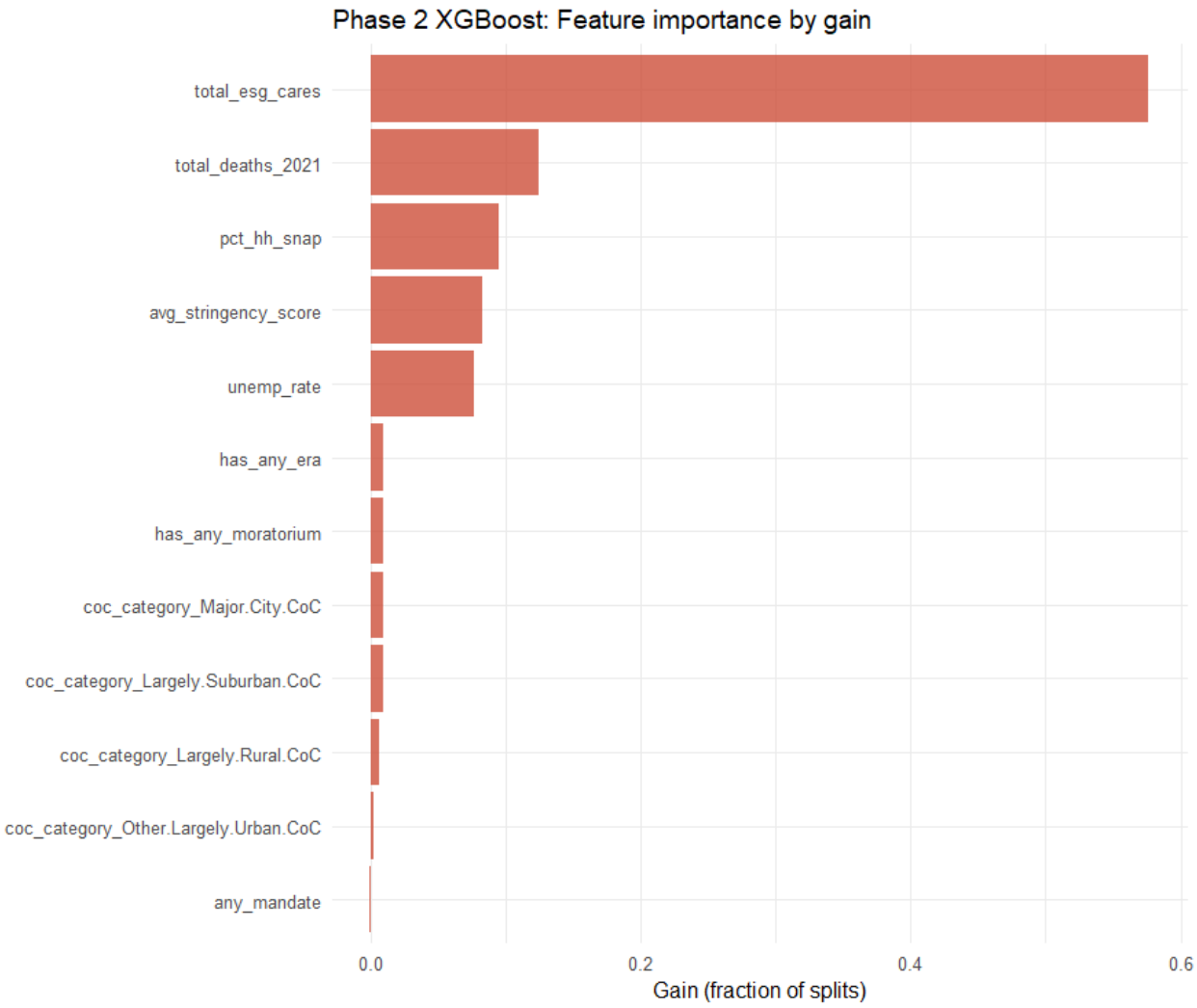

Figure 10: Phase 2 Feature Importance

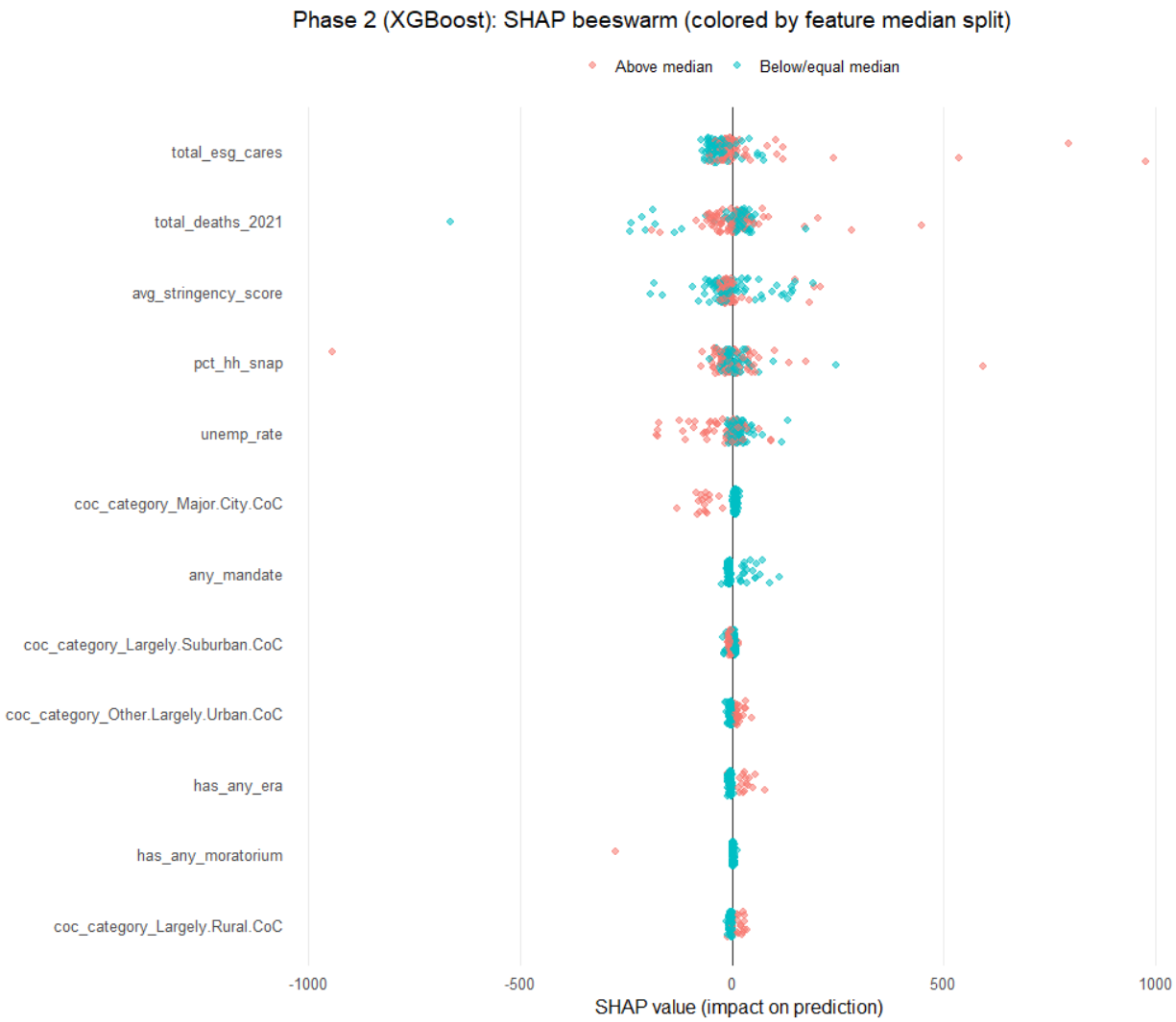

Figure 11: Phase 2 SHAP Values

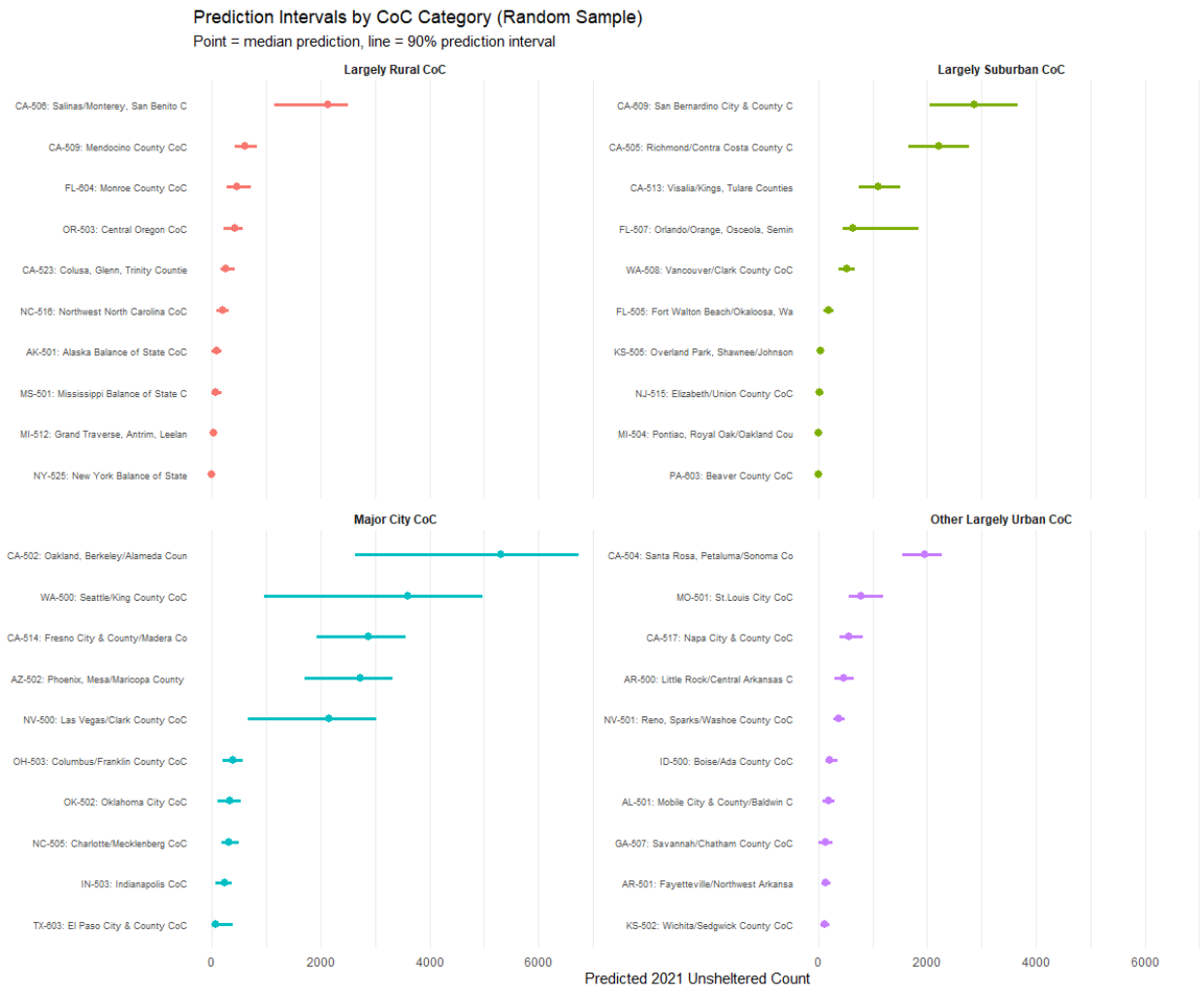

Figure 12: Cleveland Plot: Prediction Intervals of 30 CoCs by Category

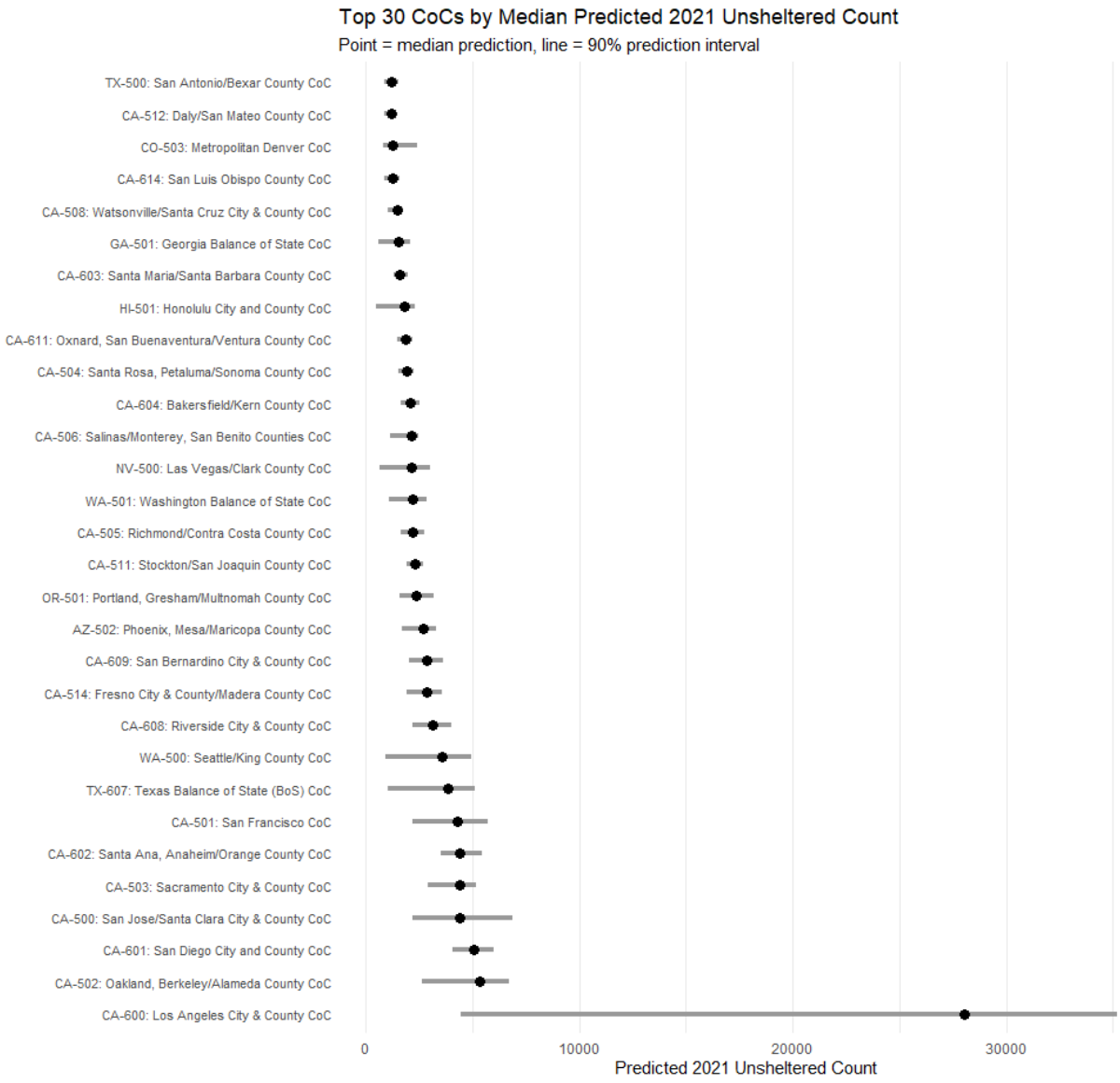

Figure 13: Cleveland Plot: Prediction Intervals of Top 30 CoCs

Table 23: Imputed 2021 Unsheltered Counts: Part 1 of 5 (AK–CA-530)

CoC ID	CoC Name	Median	p_{05}	p_{95}
AK-501	Alaska Balance of State	86	0	181
AL-501	Mobile City & County/Baldwin County	197	86	298
AL-502	Florence/Northwest Alabama	87	0	188
AL-503	Huntsville/North Alabama	194	129	286
AL-504	Montgomery City & County	158	56	274
AL-505	Gadsden/Northeast Alabama	248	149	373
AL-506	Tuscaloosa City & County	118	48	187
AR-500	Little Rock/Central Arkansas	474	303	646
AR-501	Fayetteville/Northwest Arkansas	129	59	230
AZ-500	Arizona Balance of State	746	361	957
AZ-501	Tucson/Pima County	587	326	953
AZ-502	Phoenix, Mesa/Maricopa County	2,720	1,704	3,321
CA-500	San Jose/Santa Clara City & County	4,417	2,198	6,906
CA-501	San Francisco	4,279	2,200	5,717
CA-502	Oakland, Berkeley/Alameda County	5,316	2,625	6,734
CA-503	Sacramento City & County	4,392	2,928	5,175
CA-504	Santa Rosa, Petaluma/Sonoma County	1,954	1,553	2,272
CA-505	Richmond/Contra Costa County	2,217	1,648	2,774
CA-506	Salinas/Monterey, San Benito Counties	2,135	1,152	2,503
CA-507	Marin County	980	701	1,170
CA-508	Watsonville/Santa Cruz City & County	1,490	1,043	1,762
CA-509	Mendocino County	602	415	836
CA-511	Stockton/San Joaquin County	2,292	1,954	2,674
CA-512	Daly/San Mateo County	1,199	908	1,443
CA-513	Visalia/Kings, Tulare Counties	1,096	744	1,500
CA-514	Fresno City & County/Madera County	2,877	1,926	3,569
CA-515	Rocklin/Roseville/Placer County	943	745	1,156
CA-516	Redding/Shasta, Siskiyou Counties	896	639	1,166
CA-517	Napa City & County	566	396	814
CA-518	Vallejo/Solano County	1,003	712	1,252
CA-519	Chico, Paradise/Butte County	1,110	912	1,357
CA-520	Merced City & County	806	538	1,154
CA-521	Davis, Woodland/Yolo County	662	501	832
CA-522	Humboldt County	700	213	959
CA-523	Colusa, Glenn, Trinity Counties	249	157	410
CA-525	El Dorado County	628	442	790
CA-526	Amador, Calaveras, Mariposa Counties	646	456	881
CA-530	Alpine, Inyo, Mono Counties	279	147	509

Note: Part 1 of 5. Complete list of 230 imputed 2021 unsheltered counts. Median = bootstrap median prediction; p_{05} and p_{95} = 5th and 95th percentiles (90% prediction interval). All estimates from end-to-end clustered bootstrap (B = 300 draws). CoC names abbreviated for space.

Table 24: Imputed 2021 Unsheltered Counts: Part 2 of 5 (CA-531–GA-508)

CoC ID	CoC Name	Median	p_{05}	p_{95}
CA-531	Nevada County	527	326	760
CA-600	Los Angeles City & County	28,038	4,466	35,180
CA-601	San Diego City and County	5,046	4,059	6,008
CA-602	Santa Ana, Anaheim/Orange County	4,386	3,525	5,438
CA-603	Santa Maria/Santa Barbara County	1,623	1,302	1,978
CA-604	Bakersfield/Kern County	2,086	1,673	2,549
CA-606	Long Beach	1,157	597	1,580
CA-607	Pasadena	650	413	981
CA-608	Riverside City & County	3,166	2,220	4,008
CA-609	San Bernardino City & County	2,874	2,052	3,660
CA-611	Oxnard/Ventura County	1,859	1,491	2,212
CA-612	Glendale	139	0	725
CA-613	Imperial County	1,095	529	1,578
CA-614	San Luis Obispo County	1,282	861	1,595
CO-503	Metropolitan Denver	1,268	826	2,417
CO-504	Colorado Springs/El Paso County	208	2	365
CO-505	Fort Collins/Larimer, Weld Counties	245	143	388
DE-500	Delaware Statewide	133	28	284
FL-500	Sarasota/Manatee, Sarasota Counties	324	205	462
FL-501	Tampa/Hillsborough County	572	379	780
FL-502	St. Petersburg/Pinellas County	822	522	1,000
FL-503	Lakeland, Winterhaven/Polk County	198	60	357
FL-505	Fort Walton Beach/Okaloosa Counties	196	100	274
FL-507	Orlando/Orange, Osceola Counties	643	440	1,848
FL-509	Fort Pierce/St. Lucie Counties	827	306	1,035
FL-511	Pensacola/Escambia, Santa Rosa	214	82	345
FL-513	Palm Bay, Melbourne/Brevard County	348	234	455
FL-514	Ocala/Marion County	197	33	374
FL-515	Panama City/Bay, Jackson Counties	229	86	328
FL-517	Hendry, Hardee, Highlands Counties	376	182	519
FL-518	Columbia, Hamilton Counties	637	332	997
FL-601	Ft Lauderdale/Broward County	865	517	1,127
FL-604	Monroe County	455	274	721
FL-605	West Palm Beach/Palm Beach County	867	391	1,050
GA-500	Atlanta	538	141	775
GA-501	Georgia Balance of State	1,565	580	2,069
GA-502	Fulton County	261	177	375
GA-503	Athens-Clarke County	88	33	153
GA-504	Augusta-Richmond County	600	387	1,009
GA-506	Marietta/Cobb County	409	250	590
GA-507	Savannah/Chatham County	140	0	259
GA-508	DeKalb County	248	153	374

Note: Part 2 of 5. See Table 23 notes for details.

Table 25: Imputed 2021 Unsheltered Counts: Part 3 of 5 (HI-500–MN-511)

CoC ID	CoC Name	Median	p_{05}	p_{95}
HI-500	Hawaii Balance of State	1,048	362	1,267
HI-501	Honolulu City and County	1,842	489	2,304
IA-501	Iowa Balance of State	109	0	212
ID-500	Boise/Ada County	204	137	354
ID-501	Idaho Balance of State	497	110	663
IL-502	Waukegan/Lake County	0	0	32
IL-504	Madison County	7	0	50
IL-506	Joliet/Will County	377	247	510
IL-508	East St. Louis/St. Clair County	0	0	32
IL-510	Chicago	1,125	350	1,627
IL-511	Cook County	0	0	121
IL-512	Bloomington/Central Illinois	85	3	167
IL-514	DuPage County	0	0	61
IL-515	South Central Illinois	13	0	98
IL-517	Aurora, Elgin/Kane County	108	64	182
IL-518	Rock Island/Northwestern Illinois	36	0	111
IL-520	Southern Illinois	67	0	139
IN-502	Indiana Balance of State	637	282	912
IN-503	Indianapolis	235	62	367
KS-502	Wichita/Sedgwick County	107	38	206
KS-505	Overland Park/Johnson County	47	19	88
KS-507	Kansas Balance of State	172	78	258
KY-500	Kentucky Balance of State	842	554	1,270
KY-502	Lexington-Fayette County	59	0	126
LA-500	Lafayette/Acadiana	165	69	284
LA-503	New Orleans/Jefferson Parish	528	264	716
LA-505	Monroe/Northeast Louisiana	231	0	393
LA-506	Slidell/Southeast Louisiana	80	0	222
LA-507	Alexandria/Central Louisiana	241	72	436
LA-509	Louisiana Balance of State	153	45	266
MA-502	Lynn	147	0	312
MA-516	Massachusetts Balance of State	150	0	353
MD-501	Baltimore	317	0	556
MD-504	Howard County	11	0	37
MD-601	Montgomery County	118	54	202
ME-500	Maine Statewide	202	73	617
MI-500	Michigan Balance of State	217	102	335
MI-501	Detroit	23	0	265
MI-502	Dearborn/Wayne County	0	0	51
MI-503	Warren/Macomb County	1	0	62
MI-504	Pontiac/Oakland County	0	0	57
MI-505	Flint/Genesee County	193	73	410
MI-506	Grand Rapids/Kent County	76	18	149
MI-507	Kalamazoo City & County	62	24	101
MI-509	Washtenaw County	42	1	98
MI-510	Saginaw City & County	565	400	876
MI-512	Grand Traverse Counties	31	0	84
MI-516	Muskegon City & County	19	0	89
MI-519	Holland/Ottawa County	44	2	107
MI-523	Eaton County	0	0	17
MN-500	Minneapolis/Hennepin County	526	275	730
MN-502	Rochester/Southeast Minnesota	58	15	94
MN-503	Dakota, Anoka Counties	55	0	123
MN-504	Northeast Minnesota	2	0	43
MN-505	St. Cloud/Central Minnesota	83	0	188
MN-506	Northwest Minnesota	15	0	60
MN-508	Moorhead/West Central Minnesota	0	0	6
MN-509	Duluth/St.Louis County	113	0	208
MN-511	Southwest Minnesota	0	0	28

Note: Part 3 of 5. See Table 23 notes for details.

Table 26: Imputed 2021 Unsheltered Counts: Part 4 of 5 (MO-500–PA-603)

CoC ID	CoC Name	Median	p_{05}	p_{95}
MO-500	St. Louis County	20	0	97
MO-501	St.Louis City	788	555	1,199
MO-606	Missouri Balance of State	411	222	561
MS-500	Jackson/Rankin, Madison Counties	26	0	146
MS-501	Mississippi Balance of State	64	0	179
MT-500	Montana Statewide	273	129	407
NC-503	North Carolina Balance of State	1,168	698	1,626
NC-504	Greensboro, High Point	154	111	195
NC-505	Charlotte/Mecklenberg	312	183	497
NC-511	Fayetteville/Cumberland County	280	77	415
NC-516	Northwest North Carolina	186	81	302
NJ-514	Trenton/Mercer County	61	0	121
NJ-515	Elizabeth/Union County	20	0	95
NV-500	Las Vegas/Clark County	2,142	663	3,027
NV-501	Reno, Sparks/Washoe County	379	289	491
NV-502	Nevada Balance of State	148	51	231
NY-501	Elmira/Steuben Counties	0	0	50
NY-507	Schenectady City & County	23	0	111
NY-508	Buffalo/Erie, Niagara Counties	63	0	167
NY-510	Ithaca/Tompkins County	0	0	45
NY-511	Binghamton/Broome Counties	53	0	157
NY-513	Wayne, Ontario Counties	0	0	53
NY-514	Jamestown/Chautauqua County	110	0	226
NY-518	Utica, Rome/Oneida Counties	0	0	55
NY-519	Columbia, Greene Counties	0	0	58
NY-520	Franklin, Essex Counties	0	0	35
NY-522	Jefferson, Lewis Counties	80	23	146
NY-523	Glens Falls/Saratoga Counties	0	0	0
NY-525	New York Balance of State	0	0	36
NY-601	Poughkeepsie/Dutchess County	2	0	53
NY-604	Yonkers/Westchester County	202	59	704
NY-606	Rockland County	8	0	48
NY-608	Kingston/Ulster County	93	40	143
OH-501	Toledo/Lucas County	305	158	471
OH-502	Cleveland/Cuyahoga County	83	0	227
OH-503	Columbus/Franklin County	380	199	574
OH-504	Youngstown/Mahoning County	0	0	48
OH-505	Dayton/Montgomery County	183	120	281
OH-507	Ohio Balance of State	568	184	859
OK-502	Oklahoma City	332	101	531
OK-503	Oklahoma Balance of State	53	0	142
OK-504	Norman/Cleveland County	109	25	179
OK-505	Northeast Oklahoma	324	200	504
OK-506	Southwest Oklahoma Regional	32	0	156
OR-501	Portland/Multnomah County	2,376	1,601	3,221
OR-503	Central Oregon	412	220	567
OR-504	Salem/Marion, Polk Counties	935	721	1,239
OR-505	Oregon Balance of State	1,129	743	1,623
PA-500	Philadelphia	850	429	1,177
PA-502	Delaware County	11	0	113
PA-503	Wilkes-Barre/Luzerne County	118	9	267
PA-505	Chester County	25	0	83
PA-509	Eastern Pennsylvania	362	212	596
PA-510	Lancaster City & County	69	1	140
PA-511	Bristol/Bucks County	461	321	586
PA-601	Western Pennsylvania	138	28	276
PA-603	Beaver County	0	0	23

Note: Part 4 of 5. See Table 23 notes for details.

Table 27: Imputed 2021 Unsheltered Counts: Part 5 of 5 (PA-605–WY-500)

CoC ID	CoC Name	Median	p_{05}	p_{95}
PA-605	Erie City & County	225	132	440
RI-500	Rhode Island Statewide	141	68	268
SC-501	Greenville/Anderson Upstate	295	66	466
SC-502	Columbia/Midlands	375	260	516
SD-500	South Dakota Statewide	171	83	274
TN-502	Knoxville/Knox County	211	136	301
TN-503	Central Tennessee	472	335	630
TN-504	Nashville/Davidson County	388	217	563
TN-506	Oak Ridge/Upper Cumberland	501	410	610
TN-510	Murfreesboro/Rutherford County	271	170	400
TX-500	San Antonio/Bexar County	1,194	904	1,515
TX-603	El Paso City & County	68	0	378
TX-607	Texas Balance of State	3,859	1,061	5,107
TX-624	Wichita Falls/Wichita Counties	125	0	265
TX-701	Bryan/Brazos Valley	108	0	218
UT-500	Salt Lake City & County	237	118	540
UT-503	Utah Balance of State	53	0	173
UT-504	Provo/Mountainland	27	0	123
VA-500	Richmond/Henrico Counties	101	37	178
VA-501	Norfolk/Chesapeake Counties	132	70	224
VA-504	Charlottesville	11	0	48
VA-513	Harrisonburg/Western Virginia	24	0	72
VA-521	Virginia Balance of State	344	214	447
VA-602	Loudoun County	39	5	79
VT-500	Vermont Balance of State	135	51	272
VT-501	Burlington/Chittenden County	101	27	227
WA-500	Seattle/King County	3,601	960	4,973
WA-501	Washington Balance of State	2,202	1,096	2,866
WA-502	Spokane City & County	493	352	626
WA-503	Tacoma/Pierce County	673	421	921
WA-504	Everett/Snohomish County	269	161	421
WA-508	Vancouver/Clark County	519	382	671
WV-500	Wheeling/Weirton Area	2	0	48
WY-500	Wyoming Statewide	214	134	307

Note: Part 5 of 5. Complete list of 230 imputed 2021 unsheltered counts, sorted alphabetically by CoC ID. Median = bootstrap median prediction; p_{05} and p_{95} = 5th and 95th percentiles (90% prediction interval). All estimates from end-to-end clustered bootstrap (B = 300 draws). CoC names abbreviated for space. National total (imputed only): Median = 50,832; 90% PI: [24,021, 81,619]. Combined with 146 observed complete counts yields national 2021 unsheltered estimate of 195,191 [114,380, 255,978].