HUD & Homelessness Data Analysis

ARIEGE BESSON

September 12, 2025

Project Description

The goal of this research is to examine the determinants of city-level homelessness and to determine what may prevent it in the future. Our work builds on research presented in the book Homelessness is a Housing Problem, which uses data from HUD homelessless counts to posit that at the city level, homelessness is explained primarily by high rents and low vacancy rates. Our contributions to furthering this research includes gathering data on 115 HUD geographies (as opposed to 30) and using arguably better measures of serious drug use and of social services. We wish to contribute to an even more robust analysis of HUD and related data, and to expand on theories of how exactly the housing market affects homelessness.

Our Data

- HUD Point-In-Time homelessness count data
  - Homelessness is difficult to estimate. Almost certainly a lot of HUD PIT data are undercounted, and there may be systematic measurement differences both across places and within places over time. HUD tries to account for these errors, and we have accounted for some as well in our regressions, but it's a fact of life that these data are messy.
- Census data
- Census microdata (PUMS)
- Drug use data (CDC Wonder database)
- Climate data (NOAA)
- Public welfare spending data (Lincoln Institute's FiSC database & Willamette Government Finance Database)

Panel dataset used for analysis:

- 115 geographies (covering about 50% of the total homeless population as counted by HUD for each year)
- 7 years of data (2007, 2009, 2012, 2014, 2017, 2019, 2022)

Setup

- load packages
- clear the environment
- call in data

Analysis

Summary Statistics

Total homelessness over time

Graph total homelessness (represented in our data) by year and category. Note: Our sample represents about 60 percent of the total homelessness counted by HUD in 2022, and likely over 90 percent of non-rural homelessness in 2022.

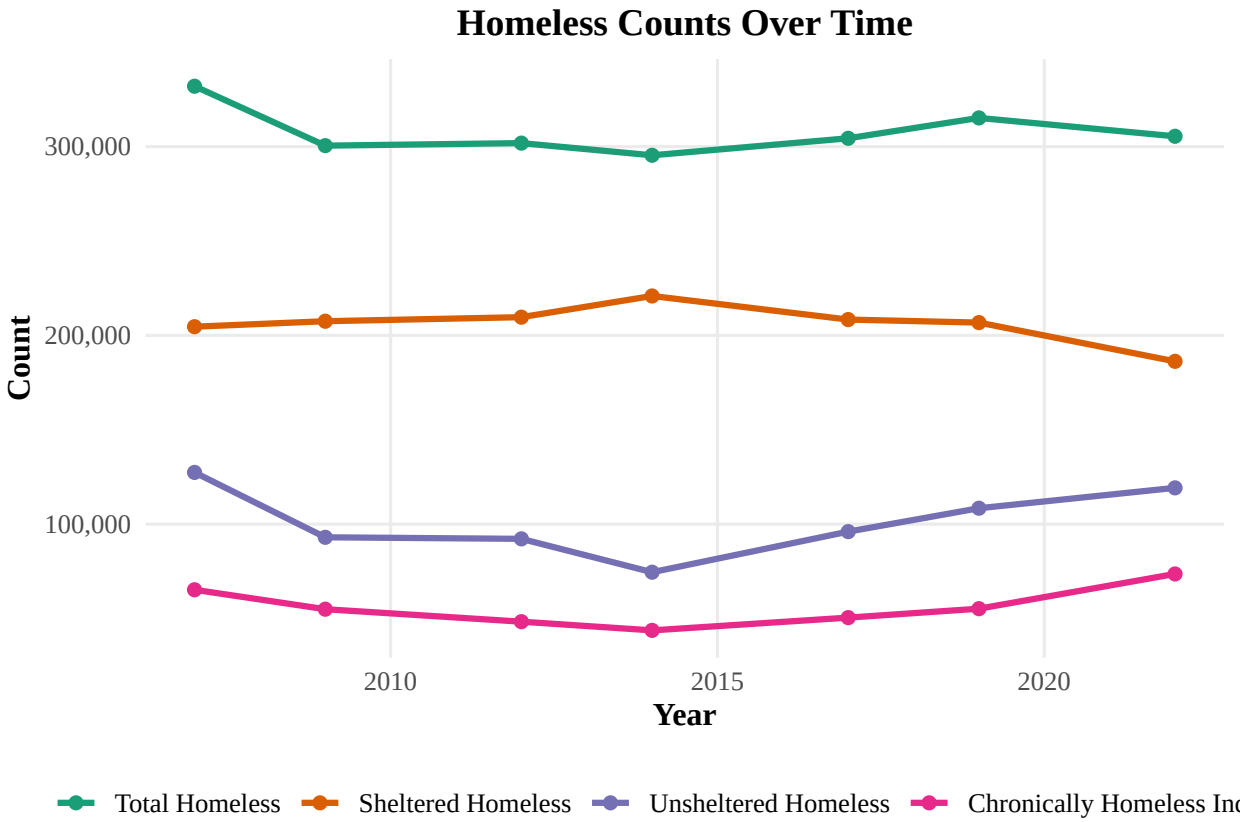


Table showing total homelessness and percentages of unsheltered and chronic homelessness over time

Table 1: Unsheltered and Chronic Homelessness Over Time

Year	Sheltered	Unsheltered	Total	Pct_Unsheltered	Chronic	Pct_Chronic
2007	204599	127397	331996	0.38	65227	0.20
2009	207499	93020	300519	0.31	54942	0.18
2012	209649	92204	301853	0.31	48325	0.16
2014	220878	74541	295419	0.25	43719	0.15
2017	208371	96040	304411	0.32	50479	0.17
2019	206772	108431	315203	0.34	55212	0.18
2022	186253	119218	305471	0.39	73611	0.24

Geographies with the highest percentages of total homelessness: New York and Los Angeles

What percentage of total homelessness per year, for the whole country as counted by HUD, does our sample represent?

The table below shows our sample's total homeless and our sample's total homeless as a percent of the national total provided by the HUD PIT counts.

Table 2: Coverage of Our Sample vs HUD PIT Counts

Year	HUD_PIT_Total	Our_Sample_Total	Pct_Our_Sample_Covers	NY_LA	Pct_NYLA_of_HUD
2007	647258	331996	0.51	103328	0.16
2009	630227	300519	0.48	87945	0.14
2012	621553	301853	0.49	92598	0.15
2014	576450	295419	0.51	105899	0.18
2017	550996	304411	0.55	131549	0.24
2019	567715	315203	0.56	137540	0.24
2022	582462	305471	0.52	130984	0.22

In 2022, New York and LA by themselves accounted for 43 percent of the count in our sample and 22 percent of the total HUD count. To put this in perspective, LA County is 3 percent of the US population and one tenth of one percent of its land area, and has some 12 percent of all its homeless.

Rates of homelessness

Rates of homelessness are also very uneven across space. In 2022 the mean homelessness rate is 17 per 10,000 residents, and the median is 12. But there are nine places with a rate over 50. These nine places are 60 percent of the homeless in the sample.

The table below shows the mean and median homelessness rates per 10,000 residents by year.

Table 3: Mean and Median Homeless Rates by Year

year	Mean_Homeless_Rate	Median_Homeless_Rate
2007	25.5	20.6
2009	24.1	19.9
2012	22.7	18.4
2014	20.3	15.5
2017	17.8	12.5
2019	17.8	12.3
2022	17.1	12.0

The table below shows places with a mean homelessness rate per 10k residents of over 50 for 2022.

Table 4: Places with High Homeless Rates in 2022

year	msa_name	state_abbr	pct_tot_homeless	rate_10k
2022	San Francisco	CA	0.0254	95.91347
2022	New York City	NY	0.2024	74.18518
2022	Los Angeles	CA	0.2264	71.12748
2022	Portland	OR	0.0171	65.75414
2022	Washington DC	DC	0.0144	65.64424
2022	Boston	MA	0.0160	63.66285
2022	San Francisco	CA	0.0319	59.83436
2022	Seattle	WA	0.0438	58.97329
2022	San Francisco	CA	0.0328	53.59858

In 2007, when homeless counts were at their highest, we see a similar pattern though with some additions: Detroit and Philadelphia loom larger, and Birmingham AL has a high rate (albeit a pretty small absolute number). NYC and SF remain dominant as shares of the US total.

The table below shows places with a mean homelessness rate per 10k residents of over 50 for 2007.

Table 5: Places with High Homeless Rates in 2007

year	msa_name	state_abbr	pct_tot_homeless	rate_10k
2007	Detroit	MI	0.0570	95.34527
2007	Washington DC	DC	0.0160	90.43128
2007	Birmingham	AL	0.0063	81.32281
2007	Boston	MA	0.0167	77.63842
2007	San Francisco	CA	0.0172	74.55136
2007	Jacksonville	FL	0.0037	70.56302
2007	Philadelphia	NJ	0.0014	70.43747
2007	New York City	NY	0.1517	60.87599
2007	Portland	OR	0.0118	55.81308
2007	Tampa	FL	0.0195	55.18729
2007	Austin	TX	0.0159	54.19940
2007	Los Angeles	CA	0.1595	53.60704
2007	Philadelphia	PA	0.0230	52.70296
2007	Seattle	WA	0.0104	51.01212

Variation

The fact that a small number of places account for most of the homelessness in the sample suggests potential problems with modelling homelessness statistically.

Variance decomposition

Variable: homeless rate

```
## total sum of squares: 27659570
##         id       time
## 0.80380647 0.02839059
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.21   85.73  149.57  207.56  253.42 1175.91
```

What this shows is that the variance in homeless rates is mostly explained by variance between places, as opposed to variance within place (and across years). ~80% of variance is between counties and ~3% of variance is across years, with the rest of the variance being residual.

The variance decomposition below, which is for the variable homeless counts, shows even more variance explained by place (95%) and very little by year (0.02%). This makes sense, as a count homelessness is very correlated with the size (population) of a place.

```
## total sum of squares: 46857960000
##           id         time
## 0.9503041660 0.0001682612
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      16     453    1034    2700    2186   78604
```

Within-place variation in homelessness rates

Another way to illustrate the stability of homelessness within places is by choosing a cutoff for "very high" homeless rates. We'll use the 90th percentile as a cutoff rate, which is 432 homeless people per 100,000 population (or 43 homeless people per 10,000 population) for the seven years of data we have. If we do this, we see the following:

- In the whole sample, there were 80 observations (out of 798) that met this threshold of very high homelessness at least once, representing 26 out of 114 places.
- 77 percent of places never exceeded this threshold
- 23 percent of places exceeded the threshold at least once (26/114)
- 81 percent of the places that were beneath this threshold once were always beneath it (88/109); 19 percent of places above it once were always above it (5/26)
- 98 percent of places that were below the threshold in one time period were below it in the next; 70 percent above it in one period were above it in the next

Let's look at the homeless rates in Los Angeles County and in Marin County, CA:

Table 6: Los Angeles County Homeless Rate

year	msa_name	state_abbr	rate_10k
2007	Los Angeles	CA	53.60704
2009	Los Angeles	CA	39.19776
2012	Los Angeles	CA	36.06018
2014	Los Angeles	CA	37.64961
2017	Los Angeles	CA	54.16241
2019	Los Angeles	CA	58.70642
2022	Los Angeles	CA	71.12748

Table 7: Marin County Homeless Rate

year	msa_name	state_abbr	rate_10k
2007	San Francisco	CA	40.38759
2009	San Francisco	CA	40.91725
2012	San Francisco	CA	28.31268
2014	San Francisco	CA	26.04027
2017	San Francisco	CA	42.80431
2019	San Francisco	CA	39.94962
2022	San Francisco	CA	43.78598

There's a fair bit of variance within Los Angeles (the rate in 2022 is double the rate in 2012) but we also know Los Angeles is something of an outlier. Marin County looks more like the average place, and we see much less within-place variance.

Average percent variation across years

Another method to measure within-place variation is the average percent variation across years. Here we can pick out places with high variation over time.

The average percent variation between years, across all places, for total homelessness (tot_homeless) is 7.50%. Places with the highest average percent changes include:

- Orleans Parish & Jefferson Parish LA (49.13%)

- Salem County NJ (-28.89%)
- Wayne County MI (-25.02%)
- Tompkins County NY (22.40%)
- St. Clair County IL (-20.50%)

All other geographies have less than |20|% average percent variation between years.

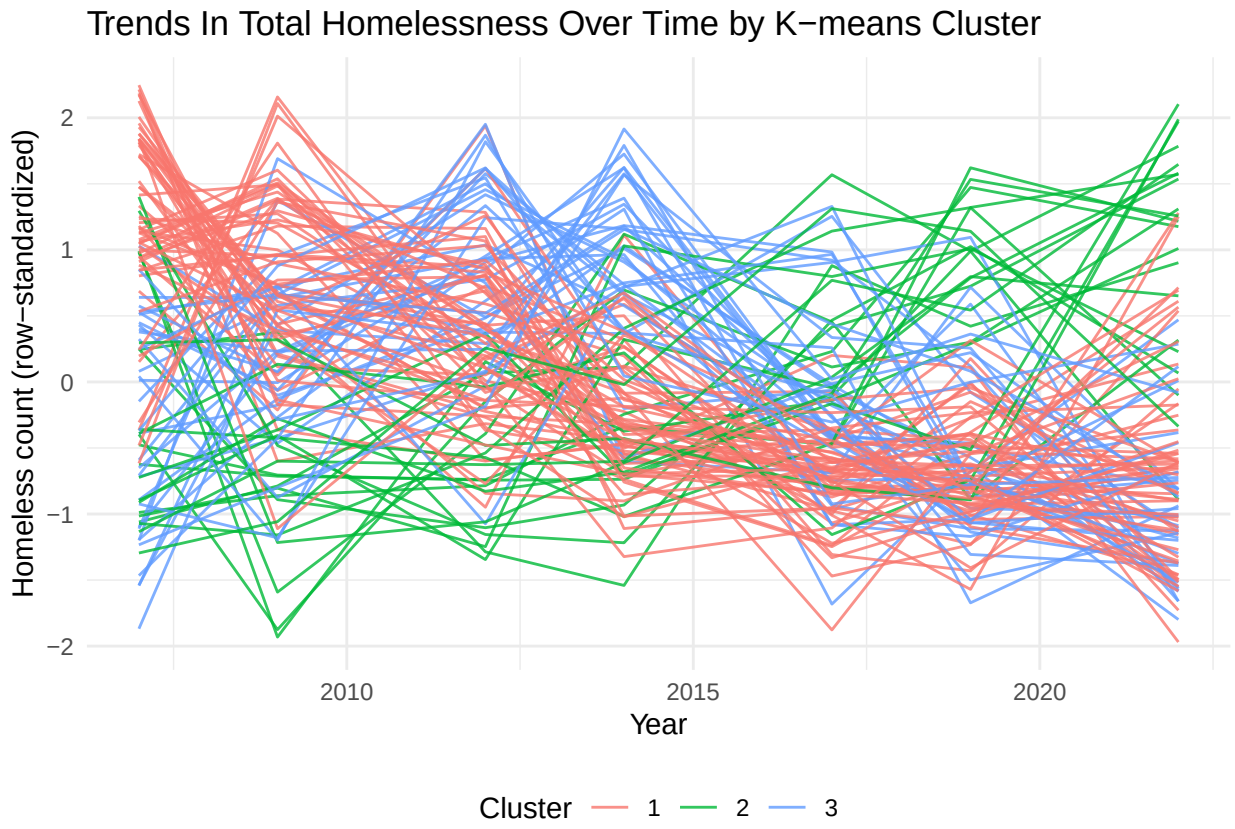
Coefficient of variation

Looking at the coefficient of variation, we see somewhat similar results. Coefficient of variation is the standard deviation divided by the mean and gives a non-directional, scale-independent measure of overall variability/stability across time. The coefficient of variation for total homeless and homeless rate is the same. The geographies with a coefficient of variation above 0.70 are:

- Salem County NJ (1.42)
- Wayne County MI (1.18)
- Orleans Parish & Jefferson Parish LA (0.95)
- Bergen County NJ (0.81)
- Hillsborough County FL (0.73)
- Rockland County NY (0.72)

Patterns in within-place variability

Grouping geographies by trends in homelessness over time: Here I used row wise standardization and then k-means to group places into three clusters based on the shape of homelessness over time.



That was mostly just for fun.

Correlations

What variables, if any, have a strong correlation with homelessness?

Within-county correlations

Within-place (demeaned) correlations measure whether two changes over time inside the same place are associated. Demeaning removes any cross-sectional (between-place) differences. However, we have only 7 data points for each place, resulting in low power, noisy estimates, and small correlations. These correlations can be used to see whether changes in a given variable within a place over time lead to changes in homelessness; causal language requires more assumptions.

By-year correlations

By-year correlations are influenced by variation between places (eg size, urban/suburban/rural, other characteristics of place), which we already know is high. However, these correlations can help pick out which characteristics different places (in the same year) are associated with higher homeless rates. This can be used for targeting policy or for descriptive comparisons.

Table 9: Correlations with Homeless Rate by Year (Filtered, Sorted Descending in 2022)

Variable	2007	2009	2012	2014	2017	2019	2022
homeless_rate	1.00	1.00	1.00	1.00	1.00	1.00	1.00
share_renter	0.48	0.49	0.56	0.55	0.59	0.62	0.58
tot_wages	0.24	0.20	0.29	0.39	0.51	0.58	0.57
avg_wage	0.23	0.19	0.24	0.36	0.43	0.52	0.54
welf_percap	0.39	0.51	0.52	0.62	0.63	0.60	0.48
share_pov_65plus	0.39	0.50	0.53	0.51	0.53	0.54	0.47
tot_employment	0.20	0.16	0.22	0.30	0.42	0.47	0.47
deaths	0.29	0.17	0.19	0.23	0.31	0.40	0.44
tot_pop	0.17	0.10	0.15	0.21	0.33	0.37	0.39
share_crowded	0.27	0.11	0.18	0.18	0.30	0.39	0.39
share_vacant_onmarket	0.13	0.28	0.28	0.12	0.14	0.23	0.36
share_vli_alone	0.45	0.50	0.59	0.54	0.44	0.36	0.33
share_vli_roommate	0.36	0.43	0.35	0.31	0.36	0.39	0.33
mgr	-0.01	0.01	0.04	0.12	0.20	0.29	0.32
mgr_80s	-0.03	-0.02	-0.04	0.08	0.23	0.29	0.31
renters_hh_inc	-0.09	-0.07	-0.08	0.01	0.14	0.29	0.30
share_no_hs	0.36	0.23	0.29	0.21	0.25	0.27	0.29
share_vli_crowded	0.27	0.08	0.10	0.13	0.22	0.36	0.29
share_vli_moved_far	0.25	0.16	0.27	0.32	0.37	0.29	0.29
crude_rate	0.34	0.28	0.13	0.09	0.12	0.15	0.29
share_pov_18to64	0.38	0.31	0.42	0.34	0.34	0.28	0.26
share_vli_moved	0.41	0.27	0.29	0.31	0.27	0.24	0.26
owners_hh_inc	-0.11	-0.11	-0.08	0.02	0.11	0.19	0.25
share_vli_disability	0.40	0.38	0.51	0.41	0.28	0.22	0.25
avg_temp_jan	0.20	0.35	0.31	0.20	0.12	0.18	0.25
lowerq_contractrent	-0.11	-0.08	-0.06	-0.03	0.04	0.13	0.22
share_unemployed	0.36	0.15	0.23	0.22	0.23	0.19	0.21
share_vacant	0.25	0.28	0.20	0.11	0.11	0.10	0.16
percap_inc	-0.18	-0.12	-0.12	-0.03	0.06	0.13	0.14
overall_hh_inc	-0.27	-0.27	-0.26	-0.15	-0.05	0.05	0.08
share_lbrforce	-0.24	-0.20	-0.14	-0.06	0.03	0.08	0.04
share_black	0.25	0.33	0.30	0.28	0.23	0.17	0.02
avg_pending							0.02
share_vli_gq	-0.03	-0.01	0.16	0.11	0.04	-0.05	-0.01
share_seasonal	0.02	-0.03	0.05	0.08	0.00	-0.02	-0.03
share_divorced	0.22	0.26	0.19	0.05	0.01	-0.08	-0.05
share_vli_rv	0.20	0.24	0.09	-0.11	-0.09	-0.07	-0.07
share_excess_bdrm_sfh	0.26	0.28	0.25	0.29	0.05	-0.07	-0.12
share_employed	-0.36	-0.15	-0.23	-0.22	-0.23	-0.19	-0.21
avg_temp_jul	-0.01	0.11	-0.08	0.04	-0.12	-0.22	-0.28
share_veterans	-0.10	-0.01	-0.12	-0.16	-0.25	-0.31	-0.30
share_excess_bdrm	0.01	0.02	-0.12	-0.13	-0.23	-0.32	-0.32
share_sfd	-0.36	-0.42	-0.51	-0.50	-0.56	-0.56	-0.47

Table 8: Within-County Correlations with Homeless Rate (demeaned by county)

	variable	cor	p.value	sig
homeless_rate	homeless_rate	1.00	<0.001	***
share_veterans	share_veterans	0.37	<0.001	***
share_no_hs	share_no_hs	0.33	<0.001	***
share_unemployed	share_unemployed	0.26	<0.001	***
share_vacant_onmarket	share_vacant_onmarket	0.24	<0.001	***
share_lbrforce	share_lbrforce	0.23	<0.001	***
share_vacant	share_vacant	0.20	<0.001	***
share_vli_moved	share_vli_moved	0.12	0.001	**
share_vli_rv	share_vli_rv	0.11	0.005	**
welf_percap	welf_percap	0.08	0.027	*
tot_wages	tot_wages	0.07	0.038	*
share_vli_gq	share_vli_gq	0.06	0.122	
share_black	share_black	0.02	0.595	
avg_temp_jan	avg_temp_jan	0.02	0.619	
tot_employment	tot_employment	0.01	0.712	
share_sfd	share_sfd	0.01	0.779	
share_pov_18to64	share_pov_18to64	0.01	0.853	
share_divorced	share_divorced	0.00	0.887	
share_vli_roommate	share_vli_roommate	0.00	0.962	
share_vli_crowded	share_vli_crowded	0.00	0.988	
overall_hh_inc	overall_hh_inc	-0.04	0.282	
percap_inc	percap_inc	-0.04	0.281	
mgr_80s	mgr_80s	-0.04	0.213	
owners_hh_inc	owners_hh_inc	-0.04	0.203	
avg_wage	avg_wage	-0.05	0.156	
share_crowded	share_crowded	-0.06	0.104	
share_vli_disability	share_vli_disability	-0.07	0.067	
renters_hh_inc	renters_hh_inc	-0.08	0.018	*
share_vli_moved_far	share_vli_moved_far	-0.10	0.015	*
deaths	deaths	-0.11	0.002	**
avg_temp_jul	avg_temp_jul	-0.14	<0.001	***
lowerq_contractrent	lowerq_contractrent	-0.14	<0.001	***
mgr	mgr	-0.14	<0.001	***
share_pov_65plus	share_pov_65plus	-0.15	<0.001	***
share_seasonal	share_seasonal	-0.17	<0.001	***
tot_pop	tot_pop	-0.17	<0.001	***
share_renter	share_renter	-0.18	<0.001	***
share_excess_bdrm_sfh	share_excess_bdrm_sfh	-0.20	<0.001	***
share_excess_bdrm	share_excess_bdrm	-0.20	<0.001	***
share_vli_alone	share_vli_alone	-0.26	<0.001	***
share_employed	share_employed	-0.26	<0.001	***
crude_rate	crude_rate	-0.27	<0.001	***
avg_pending	avg_pending		NA	

Excess bedroom variable

Table 10: Excess Bedroom Variables Correlations with Homeless Rate by Year

Variable	2007	2009	2012	2014	2017	2019	2022
share_excess_bd	0.26	0.28	0.25	0.29	0.05	-0.07	-0.12
share_excess_bdrm	0.01	0.02	-0.12	-0.13	-0.23	-0.32	-0.32

This variable was constructed as the number of bedrooms in excess of household members. Below are the by-year correlations for excess bedroom variables, pulled out from the table above. It appears that both measures of excess bedrooms (share_excess_bdrm_sfh, single-family homes only; and share_excess_bdrm, all units) show weak to modest positive correlations with homelessness. The trend in correlations flips over time for both variables from positive to negative, such that in later years more excess bedrooms is correlated with lower homelessness rates. Excess bedrooms could be acting as a proxy for other structural dynamics within housing markets, or excess bedrooms could indicate less tight housing markets where people have more space to host troubled friends or family who might otherwise experience homelessness.

Regressions

Linear regression with fixed effects

Model 1
Model: linear regression on log of a variable with fixed effects
Dependent variable: homeless rate (`log(homeless_rate)`)
Predictors: `mgr`, `share_pov_18to64`, `share_black`, `share_pov_65plus`, `welf_percap`, `avg_temp_jan`, `crude_rate`
Fixed effects: Year (`year`) and MSA (`msa_number`)
Errors: clustered robust standard errors at the CoC level (`coc_id`)

Table 11: Model 1: Linear FE OLS

Dependent Variable:	lhomeless
Model:	(1)
Variables	
mgr	0.0009***
	(0.0003)
share_pov_18to64	2.25
	(2.06)
share_black	1.32**
	(0.640)
share_pov_65plus	6.50***
	(1.63)
welf_percap	0.0003***
	(4.5×10^{-5})
avg_temp_jan	-0.002
	(0.005)
crude_rate	0.008***
	(0.002)
Fixed-effects	
year	Yes
msa_num	Yes
Fit statistics	
Observations	737
R^2	0.68958
Within R^2	0.47914

Clustered (coc_id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Quasi-Poisson regression

Model 2
Model: quasi-Poisson with a log link
Dependent variable: homeless count (`tot_homeless`)
Predictors: `mgr`, `share_pov_18to64`, `share_black`, `share_pov_65plus`, `welf_percap`, `avg_temp_jan`, `crude_rate`
Fixed effects: Year (`year`) and MSA (`msa_number`)
Errors: clustered robust standard errors at the CoC level (`coc_id`)

Table 12: Model 2: Quasi-Poisson GLM

	Term	Estimate	StdError	tValue
(Intercept)	(Intercept)	3.264*	1.302	2.507
mgr	mgr	0.003***	0.001	4.748
share_pov_18to64	share_pov_18to64	11.212**	4.252	2.637
share_black	share_black	2.64	1.917	1.377
share_pov_65plus	share_pov_65plus	16.004***	3.662	4.370
welf_percap	welf_percap	0	0.000	1.420
avg_temp_jan	avg_temp_jan	-0.002	0.009	-0.212
crude_rate	crude_rate	-0.01.	0.006	-1.654

Note:
Year and MSA fixed effects included but not shown.

Poisson regression

Model 3
Model: Poisson regression with a log link
*(*estimated via `fixest::fepois` in R; high-dimensional fixed effects and cluster-robust SEs)
Dependent variable: homeless count (`tot_homeless`)
Predictors: `mgr`, `share_pov_18to64`, `share_black`, `share_pov_65plus`, `welf_percap`, `avg_temp_jan`, `crude_rate`
Fixed effects: Year (`year`) and MSA (`msa_number`)
Errors: clustered robust standard errors at the CoC level (`coc_id`)

Table 13: Model 3: Poisson FE

Dependent Variable:	tot_homeless
Model:	(1)
Variables	
mgr	0.003***
	(0.0006)
share_pov_18to64	11.2**
	(4.43)
share_black	2.64
	(2.00)
share_pov_65plus	16.0***
	(3.82)
welf_percap	0.0004
	(0.0003)
avg_temp_jan	-0.002
	(0.009)
crude_rate	-0.010
	(0.006)
Fixed-effects	
year	Yes
msa_num	Yes
Fit statistics	
Observations	737
Squared Correlation	0.85257
Pseudo R^2	0.84723
BIC	756,931.9

Clustered (coc_id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Conclusion

Coming soon