Estimating the Annual Size of the Homeless Population in Los Angeles Using Point-in-Time Data
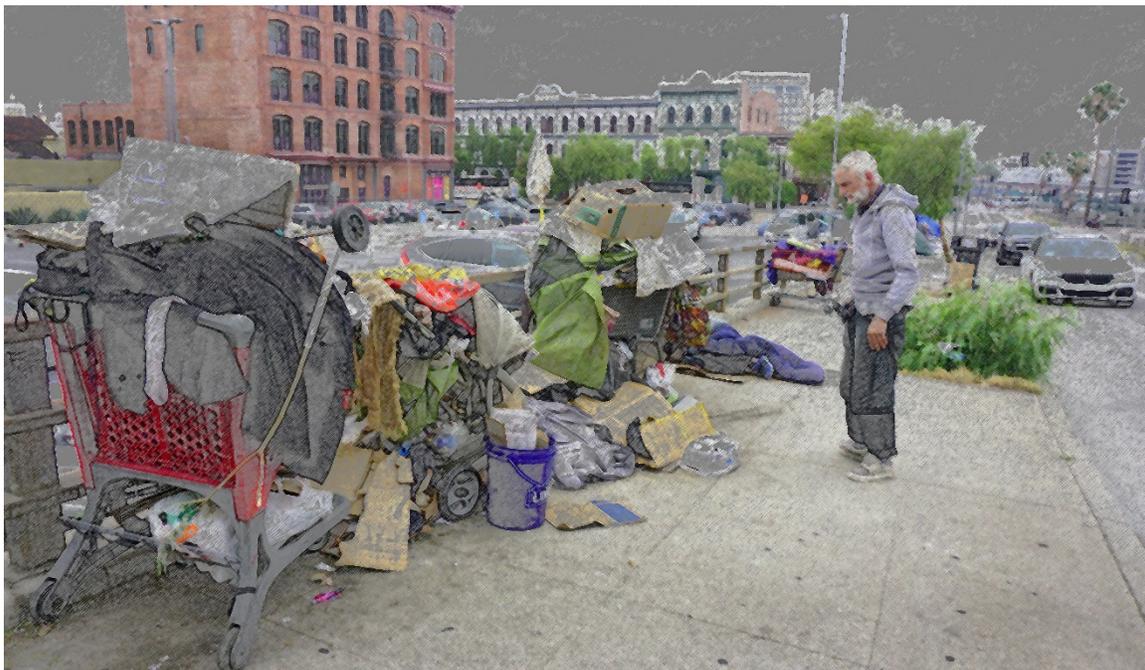
Jane Carlen

Economic Roundtable
A Public Benefit Research Organization
www.economicrt.org

Abstract

In this report we introduce a model-based framework for estimating the number of people who experience homelessness in Los Angeles over the course of a year based on point-in-time data. We refer to this as the annualized estimated of the population size. Unlike capture-recapture methods, this model is based on a single set of anonymous point-in-time data. This makes it viable for nearly all jurisdictions that conduct annual counts of their homeless population and gather basic information about the duration of homelessness within this population. In addition, it allows us to estimate the breakdown of the annualized population by duration of homelessness. A practical application of this is to quantify the distribution of homeless services needed to provide progressively more intensive levels of intervention for individuals with longer durations of homelessness. We incorporate measures of uncertainty in our final estimates, reflecting that demographic data is derived from a sample of the target population. As such, our model provides more complete information about the annualized population than the model developed by Burt and Wilkins (2005) that is currently in use in Los Angeles.

Introduction

Every year Los Angeles invests significant resources, including mobilizing thousands of volunteers, to estimate the number of people experiencing homelessness on a given night. The annual Los Angeles Homeless Count, which includes information collected from sheltered and unsheltered people, is the primary tool for measuring the size and demographics of this population and how they have changed over time. This information is critical to understanding the growing need for emergency shelter, housing services, and other interventions. With the availability of hundreds of millions of dollars in funding annually for housing and supportive services, through the county's Measure H and the city's Measure HHH, this information becomes even more vital to policymakers.

The primary goal of the Homeless Count is to measure the population that is homeless on a given night. The resulting population size estimate, which has been in the ballpark of fifty thousand for the last two years, is often taken as the definitive size of the homeless population. It is the figure most widely reported in local media, and when it declines, as reported in 2018, it is considered evidence that resources are being deployed successfully.

However, the size of the point-in-time (PIT) population, the number of people who are homeless on a given night, may not always be the primary data point for policymakers. When planning for housing and other services that are provided on an extended basis it is equally important to know how many people experience homelessness over the entire period.

In this report we address the question of how to estimate the number of people who experience homelessness over the course of a year, the annualized estimate, based only on PIT data. To

1

do this we draw on survey data that describes how long individuals have been homeless and how many stints of homelessness they have recently experienced.

We present an estimation method that can be flexibly applied regardless of the time intervals in which these measurements were gathered. Our model also allows us to present our estimate within a range of plausible estimates, reflecting limitations of the sampling methodology. In addition, it enables us to not only estimate the annualized population size, but the distribution of time spent homeless within the whole population. This is extremely important for capturing the prevalence of short-term homelessness. Although chronic homelessness is often the face of homelessness in cities like Los Angeles, more people experience shorter stints of homelessness over the course of a year. Bringing this to the forefront highlights the precarious housing situations that too many Angelenos face.

Research to date on estimating the annualized homeless population based on PIT data is sparse. In fact, we found only the model of Burt and Wilkins (2005), the model currently used Los Angeles, designed specifically for this purpose.

Related research has been conducted on modeling durations of homelessness based on shelter data and demographic surveys. For example, Metraux and Culhane (1999) present an overview of methods to analyze the lengths of shelter stays using administrative data. This work contains examples of constructing survival curves and hazard curves to describe the distribution of lengths of stays of individuals in shelter, drawing on data from New York City. More generally, this line of inquiry falls under the active research field of survival analysis.

In the next section we present the nuts and bolts of the model of Burt and Wilkins (2005) and address some of its shortcomings.

The Burt and Wilkins model

Although not widely publicized, the Los Angeles Homeless Services Authority (LAHSA) reports annualized estimates of the size of the homeless population and of many demographic subgroups. These estimates are calculated using the following model developed by Burt and Wilkins (2005).

$$Annualized\ estimate = A + 51 * B * (1 - C), \tag{1}$$

where A is the PIT count of the homeless population, B is the number of currently homeless people who became homeless in the counted area during the last week, and C is the proportion of currently homeless people who had a previous homeless episode during the last year.

This formula accounts for the fact that more people become homeless each week. If we were to conduct a PIT count in the remaining 51 weeks of the year we would initially add B people to the population estimate every week. But, according to the model, the proportion C of them would have been counted during a previous week.

The use of a consistent proportion C does not reflect that fact that if a new count were performed every week then only in the final week of the 52-week period would C approximate the proportion of individuals who were previously counted. On the other hand, if we assume that previous episodes of homelessness in a year are uniformly distributed, then in the second week of the 52-week period the proportion of people previously counted would be $\frac{C}{52}$. Summing over 52 weekly counts, the annualized estimate would be calculated as follows.

$$Modified\ annualized\ estimate = A + B(1 - \frac{C}{52} + 1 - \frac{2*C}{52} + ... + 1 - \frac{51*C}{52})$$
$$= A + 51 * B * (1 - \frac{1}{2} * C). \tag{2}$$

Thus, the original model in Equation (1) is an underestimate by $\frac{51}{2} * B * C$ of the annualized population, as compared to the revised estimate in Equation (2).

For example, in 2017 when B was estimated to be 1,513 people[1] and C to be 29.2 percent, the magnitude of underestimate was 11,264 or about ten percent.

In addition, the model introduces variability due to sampling of the unsheltered population. Although the Homeless Count attempts to count every person who is homeless on a given night, only a fraction of the unsheltered population completes demographic surveys. For example, in 2017 about 5,800 individual responses were gathered from unsheltered individuals containing information about their duration of homelessness. This represents about fifteen percent of the estimated PIT unsheltered population of 38,470 individuals. On the other hand, the survey of sheltered individuals is a near-complete census of the entire unsheltered population in a given time window, with a few minor exceptions such as domestic violence shelters.

While sampling is a necessary method for collecting data about any large and hard-to-reach population, it is important to relate the impact that it has on the precision of the resulting estimate. To illustrate this, let us make the simplifying assumptions that the true value of B is 1,000 for the unsheltered population and each individual's response is independent of others who are surveyed. Using the binomial distribution and the sampling parameters from above, our estimate of B has a 95 percent chance of being between 842 and 1,161. Our estimate of C has a 95 percent chance of being between 28.05 percent and 30.35 percent. Using our updated formula, we would expect that 95 percent of the time our derived annualized estimate would fall in the range from 111,611 to 125,352. This is a fairly wide range. Although the underlying estimate of measurement uncertainty is not perfect due to the simplifying assumptions, we advocate for presenting such a range to better reflect the data collection process.

Another weakness of the Burt and Wilkins (2005) model is that is it highly dependent on a single measurement, B, of the number of new entrants to homelessness in the previous week. Such a measure is impacted by the time of the month and clumping in reported durations

---

[1]This value of B is actually a composite estimate that we use for simplicity. In fact, since 2016 the annualized estimate calculated by LAHSA is a sum of seperate calculations for sheltered adults, sheltered youth, unsheltered adults, and unsheltered youth, where B and C are calculated for each sub-group.

of homelessness, among other factors. (Clumping in this context refers to the tendency to round off time estimates, such as if everyone who was homeless for about a week reports being homeless for one week. See Figure 1 for evidence of clumping.) The measurement of B is also sensitive to changes in data collection. For instance, in 2017 the youth-specific survey stopped asking questions about how long an individual was in Los Angeles County. Data from 2016 was therefore used to fill in missing pieces of the calculation (Cox et al. 2018).

The upshot of the sensitivity of the Burt and Wilkins model to sampling variation, measurement error and changes in data collection procedures is shown in the fluctuation of the annualized estimates. While estimates of the PIT homeless population in Los Angeles always increased in reported counts from 2011 to 2017, the annualized estimate has decreased in all but one year, as shown in Table 1. Part of the reason why there is such a large drop in the annualized estimate from 2016 to 2017 is that the method for calculating the length of stay in a shelter changed. In 2016, the length of stay in the shelter was used. In 2017, this length could be added to the length of stay in the prior living situation to determine whether the individual was homeless for a week or less (Cox et al. 2018).

Table 1: LAHSA PIT and annualized population estimates based on the Burt and Wilkins model for the Los Angeles CoC.

Year	PIT Estimate	Annualized Estimate
2011	34622	120,072
2013	35524	187,119
2015	41174	162,769
2016	43854	156,431
2017	52442	107,094

A final shortcoming of the Burt and Wilkins model that we address is that that it does not account for potentially unobserved very short-term homeless populations. The proportion of people homeless for at most a week is an underestimate because it does not account for people who are homeless for only a few days but not at the time of the survey. They would drive up the value of B in the model, which means that the reported value of B is likely an underestimate.

One way to account for this would be to base the model on the number of newly homeless people in the last day rather than in the last week, but this single-day measure would be even more prone to measurement error and natural variability than B. A related concern is that people who have only been homeless for a very short time may be the most likely to be skipped over by people conducting demographic surveys, even if they are homeless when the survey is conducted. Although the count of the PIT unsheltered population occurs overnight, the demographic survey is conducted largely during the day, when it is more difficult to determine who is homeless. This would also lead to an underestimate of B.

Admittedly, this unobserved group is difficult to measure directly. It would require identifying people who experience very short stints of homelessness, who may or may not be homeless when the survey is implemented. It may be possible to estimate the size of this group using data from shelters, for which we have records over an entire month. However, that would

require gathering reliable information about individuals' destinations after leaving a shelter. In 2017 most of that data was missing.

For these reasons, we propose a model that leverages a much larger pool of information about durations of homelessness to make inferences about the size of the short-term homeless population. It reflects our assessment that the percentage of individuals in the PIT population who have been homeless for a given duration should in general decrease as the duration increases, barring strong seasonal effects, policy changes or shocks to the economy or housing market. For instance, the number of people who have been homeless for two months is a subset of those who were homeless for one month a month ago. Therefore, we expect the PIT two-month group to be smaller than the one-month group. Given the minimal seasonality in Los Angeles, we assume that population entry and exit rates are fairly stable throughout the year.

Data

As mentioned above, annual Homeless Count data includes both the overnight count data used to estimate the size of the PIT population and demographic data that contains detailed information about a subset of the population. Throughout this report we use data that describes the Los Angeles Continuum of Care (CoC), which includes all of Los Angeles County except the cities of Pasadena, Glendale and Long Beach. If not otherwise specified, the data employed is from 2017, the most recent year for which we have demographic survey data.

The demographic data covers a wide array of topics, but the available variables depend on the year, whether the survey respondent was sheltered or unsheltered, and whether they were in the "youth" group (age 18-24). For our model, the most important variables are those describing time spent homeless. Specifically, we use the reported durations of current stints of homelessness and number of times homeless in the previous year or, if that is unavailable, three years. We find that the it is necessary to pre-process some of the information about durations of current stints, and we describe this process with the introduction of our model framework in the next section.

Model Framework

The most important advancement of the estimating methodology we present below over the Burt and Wilkins model is that we first model the distribution of the lengths of current stints of homelessness. We then use this fit curve to calculate an annualized estimate and derive other valuable information about the distribution of time spent homeless among the annualized population.
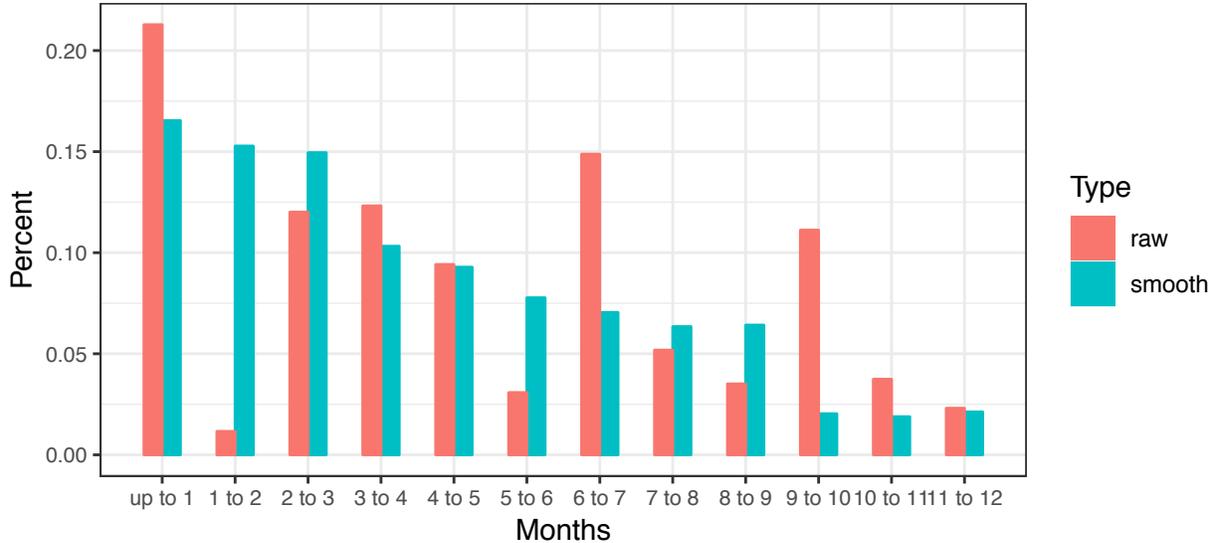
Figure 1: Raw and smoothed distribution of duration of homelessness (2017).

Smooth the duration data

Because of the clumping of survey responses mentioned previously, we must take steps to smooth the distribution of reported stint durations before modeling it. We restrict our attention to durations under a year as this is sufficient for our model, and detailed stint durations become less reliable for longer periods. Figure 1 shows the distribution of durations by month before and after this process. Before smoothing there are spikes in the distribution at one month, six months and nine months. This suggest a tendency for survey respondents to significantly round off the duration of their current stint of homelessness.

Accordingly, to smooth the data we first group the data into quarters of the year. We then fit a piecewise exponential distribution to each quarter. This agrees with our previous assumption that the likelihood of observing a certain stint duration decreases with its length. It also makes the simplifying assumption that exit rates from homelessness are constant within each quarter. Finally, the data is resampled from the piecewise distribution to reflect the natural variability in observed data. While we expect a decreasing curve overall, we know that in practice it will be spiky rather than smooth. The smoothed distribution shown in Figure 1 represents a single sample from the piecewise exponential, but we will ultimately draw on many samples to obtain our annualized estimate and measure uncertainty.

Note that this smoothing process was designed for our data, but it may not be optimal in all cases. In some cases this step may not even be necessary, for example if most homeless people are sheltered and duration data is carefully recorded rather than self-reported.

Model the distribution of durations

The next step in our modeling process is to fit a curve to the smoothed data. The class of distribution we employ to do this conveys certain assumptions about exit rates from homelessness over time.

We focus our attention on two commonly used and easily interpretable distributions used in survival analysis, the exponential distribution and the Weibull, which encompasses the exponential distribution as a special case. The "memoryless" property of the exponential distribution implies consistent exit rates from homelessness from month to month. In other words, the chance that a person who is homeless finds housing or leaves the area does not change over time. Although this is probably not realistic, it may be a useful conservative assumption if no additional information is available.

The Weibull distribution can capture increasing, constant or decreasing exit rates from homelessness over time depending on its parameters. The most appropriate trend in exit rates will vary by jurisdiction. We can accordingly place bounds on the parameters when optimizing based on local conditions. For example, Metraux and Culhane (1999) note that in New York City some households stayed in a shelter until they received a subsidized housing placement, but this would not occur until at least 90 days into the episode and typically not until 9 months to a year had elapsed. This would impact exit rates in a non-intuitive way. In that case other classes of distributions or non-parametric curves may be called upon to fit the duration data.

In the absence of such programmatic considerations in Los Angeles, we believe that exit rates are likely to decrease over the course of a year. The longer one is homeless, the more likely he or she is to face health problems, social isolation, or involvement with the criminal justice system that can create a more permanent barrier to finding housing. Accordingly, we place an upper bound of one on the estimated shape parameter of the Weibull distribution, which forces exit rates to be constant or decreasing over time.

In addition, we continue to restrict our attention to durations of up to a year for the reasons given above. Accordingly, our parameters are optimized based on a truncated distribution of data between one day and less than a year. In future applications we can account for underrepresentation of very short-term homeless people in demographic survey data by further truncating the fit region to exclude very short durations. For simplicity, we have not done this here, but it may be a useful strategy for basing estimates on the most trustworthy data.

Figure 2 plots the estimated Weibull distribution curve against the background of smoothed duration data. We found during estimation that our data did not support our hypothesis of significantly decreasing exit rates over time. Given our pre-determined parameter bounds, this resulted in fit distributions to resampled data with constant or near constant exit rates. In the plot above, the shape and scale parameters of the estimated Weibull distribution are 1 and 1.853, implying a constant monthly exit rate of 16.5 percent.
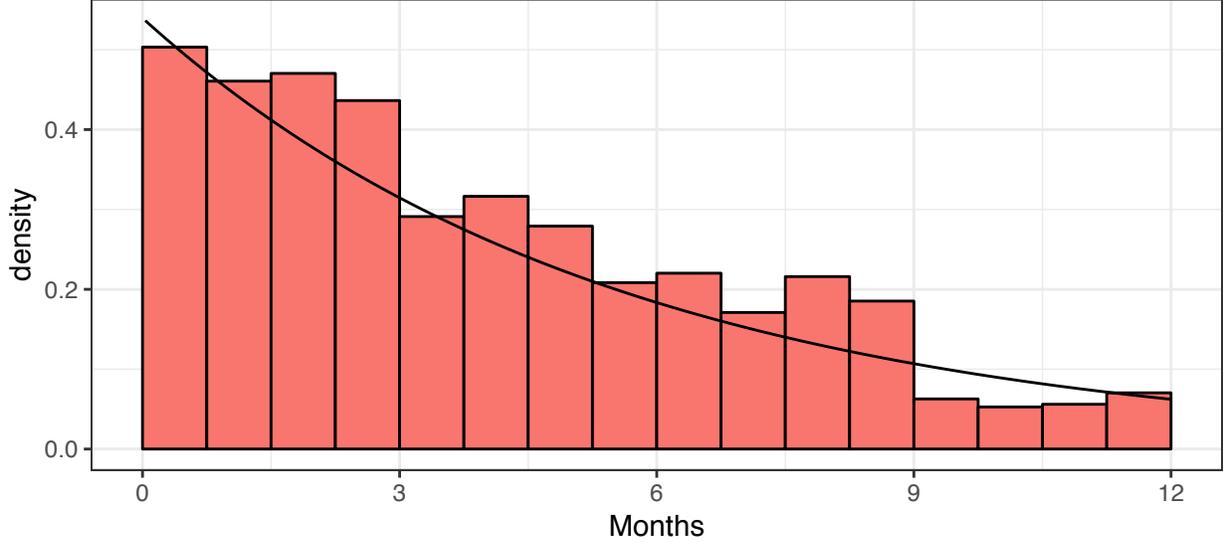
Figure 2: Estimated Weibull distribution shown against the smoothed data it models (2017).

Annualized estimate

We use our fit distribution to determine the number of newly homeless individuals in a single day or week more reliably than deriving this figure from survey responses. In fact, a benefit of our approach is that it is not tied to the time intervals in which the data was gathered. We adapt the formula in Equation (2) to accept as input a daily estimate of newly homeless people, D, instead of the weekly value (B) used by Burt and Wilkins.

$$Day\text{-}based\ annualized\ estimate = A + 364 * D * (1 - \frac{1}{2} * C). \tag{3}$$

We present the estimate based on this model in Table 2, along with a measure of uncertainty in our estimate.

Uncertainty

Our annualized estimate deviates from the true value for several reasons, including:

- Demographic survey respondents are a sample of the target population.

- There is natural variation in daily entries and exits from homelessness despite long-term trends.

- Demographic surveys of unsheltered people may not be representative. One factor is the challenge of surveying people who have been homeless for only a few days.

We reflect these sources of measurement error in our estimates in several ways. First, we use bootstrapping – repeated resampling of demographic survey data – to capture the variation

due to sampling of the unsheltered population. Second, we resample from the resulting piecewise exponential fit of the duration data to capture natural variation over time. Lastly, we restrict the Weibull shape parameter to be less than one to ensure that the duration distribution curve is always decreasing. This captures the fact that we expect to find more people on their first night of homelessness on any given night than on their second, and so forth, which helps to account for undercounting. As mentioned above, we can further account for undercounting by tailoring our truncated distribution, though we did not employ that strategy here.

Applying these resampling techniques is roughly equivalent to gathering new hypothetical data sets under the same conditions that generated our observed data. Each resampled data set leads to new estimates of the annualized population, the number of new entrants to homelessness each day (D), and the PIT estimate of people whose current duration of homelessness is at most a week (B).

We repeated this resampling process 1,000 times. Table 2 illustrates the 5th percentile, 50th percentile (median), and 95th percentile values out of the 1,000 estimates. The median values are used as "point estimates," single values we present as estimates of the annualized population, D and B respectively. The spread between the 5th and 95th percentile values is a measure of the uncertainty in our point estimates. We find that the variation is not very large, so our estimates are fairly precise.

Table 2: Percentiles of repeated estimates of the annualized population, D and B.

	Annualized population	D	B
5%	101,440	158	1,085
50%	104,755	168	1,157
95%	108,082	179	1,225

Our annualized population estimate of 104,755 shown in Table 2 differs from LAHSA's annualized estimate for 2017 by -2,339, and is well below LAHSA's three previously reported annualized estimates (see Table 1). We estimate B to be 1,157 and D to be 168. Multiplying our estimate of D by seven to make it a weekly estimate produces a value that is 22 higher than our estimate of B. This implies that using the snapshot measure B as a proxy for the number of new entrants to homelessness over the course of a week undercounts very short-term homeless people by 22 people per week. This was described earlier as a shortcoming of the Burt and Wilkins model, and our methodology allows us to quantify its impact.

Annualized distribution of duration of homelessness

Modeling current stint durations not only allows us to estimate the annualized population. We can also estimate how long those in the annualized population have experienced homelessness. Below we compare the distribution of duration of homelessness in PIT data, i.e. the length of the stint when surveyed, to the distribution from annualized data, where duration can include PIT stint length and time spent homeless during the year.

Table 3: PIT versus annualized distribution of duration of homelessness based on one simulation (2017). Duration for the annualized population may occur over several stints.

Duration (months)	PIT total	PIT percent	Annualized total	Annualized percent
up to 1	4,522	9%	10,626	10%
1-2	3,778	7%	8,604	8%
2-3	3,156	6%	8,475	8%
3-4	2,636	5%	7,215	7%
4-5	2,202	4%	6,514	6%
5-6	1,840	4%	5,474	5%
6-7	1,537	3%	4,883	5%
7-8	1,284	2%	4,183	4%
8-9	1,072	2%	3,458	3%
9-10	896	2%	3,040	3%
10-11	748	1%	2,492	2%
11-12	625	1%	2,173	2%
12+	28,145	54%	35,818	35%
total	52,442		102,955	

Table 3 shows that those who have been homeless for at least a year make up just over a third of the annualized population, as opposed to over half of the PIT population. If the shape parameter of the estimated Weibull distribution were estimated to be less than one, implying decreasing exit rates over time, we would find this pattern to be even more stark, and the annualized population to be significantly larger.

Discussion

In this report we have presented a model-based framework for annualized population estimation that builds on the model of Burt and Wilkins currently in use. Our method is more stable, more flexible for use with different data sources, and provides greater information about durations of homelessness experienced over the course of a year.

Practical implications

Reliable estimates of time spent homeless during a year support an evidence-based intervention framework. Understanding the prevalence of short-term versus persistent episodes of homelessness enables accurate allocation of resources based on differing levels of need. The prevailing approach to homelessness prevention and intervention is "progressive engagement." This entails treating new entrants into homelessness with a light touch and providing progressively more intensive services as needed to help individuals reestablish themselves in stable housing. The most intensive and costly intervention, permanent supportive housing,

is reserved for chronically homeless individuals for whom other interventions have failed. A population model, as shown in Figure 2, can be used to plan, budget and implement the range of services needed to help individuals with differing barriers to exiting homelessness.

Recommendations

We recommend that the estimation method presented in this report be adopted.[2] Alternatively, if that is not feasible, based on the findings above we recommend several improvements to the Burt and Wilkins model that can easily be implemented:

- The formula adjustment shown in Equation (2).

- Use averages of B and C over the last two to three years instead of a single-year's estimate, applying the most up-to-date method for calculating each. This will reduce year-to-year fluctuations, and is appropriate since we do not expect major changes in either from one year to the next. (An exception might be if shelter or housing policy changes significantly.)

- Publish the formulas and data sources used to calculate B and C, noting which survey variables were used. This makes the causes of year-to-year fluctuations more transparent.

- Present annualized estimates with measures of uncertainty, even if they rely on simplifying assumptions. This can be accomplished using bootstrapping or the binomial distribution.

- Ensure that the information needed to calculate the variables in the model is gathered for both sheltered and unsheltered people. For example, both surveys should ask about the number of times one has been homeless in the last year, the total amount of time spent homeless, and whether the individual was in Los Angeles during that time. Currently, the intake questionnaire for sheltered people does not ask about the number of times an individual was homeless in the previous year, but only in the previous three years. To compensate, analysts "adjust the percentage homeless more than once in the past three years for the shelter population by the ratio of the one year to three year measure taken from the data in the youth and adult demographic surveys" (Cox et al. 2018). There are significant reported differences between the sheltered and unsheltered populations which suggest this introduces bias. For example, in 2017 the vast majority of unsheltered homeless people reported being homeless only one time in the past three years, whereas for the sheltered population this figure was similar to the number reporting multiple instances of homelessness. We refer the reader to explore this and other differences through our interactive data visualization tool on the Economic Roundtable website, https://economicrt.org/los-angeles-homelessness-data-dive-app/.

---

[2]The code and data used to create this report are available at the Economic Roundtable github repository: https://github.com/economicroundtable/homeless_LA/.

Limitations and future work

Despite the advancements our method offers, there are many ways it can be improved upon in future work.

- The calculation of C could be adapted to specifically describe people whose current stint of homelessness has lasted a week or less. For simplicity and comparability with previous estimates we employed the estimate of C used by LAHSA. However, it seems to be biased. It is used to describe the likelihood that recently homeless individuals have already experienced homelessness during the annual period, yet it is based on the probability that anyone surveyed experienced homelessness during that time. Of course, those who have been homeless for a long time have no chance of a previous instance of homelessness during the year.

- Address seasonality of new entrants to homelessness. While it is somewhat reasonable to ignore seasonality in Los Angeles where outdoor sleeping is possible for nearly all of the year, many jurisdictions exhibit seasonality in the number of new entrants to homelessness and duration of stints of homelessness. To accurately extend our model to these jurisdictions would require knowledge of season effects in this context.

- Use our model framework to estimate annualized sub-populations. LAHSA provides annualized estimates for many sub-groups, and it would be interesting to apply our methodology to these groups and provide measures of uncertainty, especially for small sub-groups.

Finally, although we have attempted to correct for or measure the causes of error in our annualized estimate, there are deeper sources of error that we can do little about. Our annualized estimate depends on the PIT count, which is also subject to error. For example, the count is conducted by volunteers who may miss or miscount individuals. The multipliers used to count individuals in tents and vehicles are imperfect approximations. These issues are addressed in greater depth in Flaming and Burns (2017), and D. Smith and Smith (2018), for example. They affect both the Burt and Wilkins (2005) model and ours and are beyond the scope of this report.

Acknowledgements

We are also indebted to the volunteer organizers with DataKind who made the data dive possible: Wade Fuller, William High, Abhishek Kapatker, Sonali Murarka, Stacey Ronaghan, and Haitham Shammaa.

References

Burt, Martha R., and Carol Wilkins. 2005. "Estimating the Need: Projecting from Point-in-Time to Annual Estimates of the Number of Homeless People in a Community and Using This Information to Plan for Permanent Supportive Housing." Corporation for Supportive Housing. https://www.csh.org/wp-content/uploads/2013/08/Estimating-the-Need.pdf.

Cox, Robynn, Benjamin F. Henwood, Patricia St. Clair, and Jennifer Ailshire. 2018. "2017 Los Angeles Continuum of Care Homeless Count Methodology Report." University of Southern California.

Flaming, Daniel, and Patrick Burns. 2017. "Who Counts: Assessing Accuracy of L.A.'s Homeless Count." Economic Roundtable. https://economicrt.org/publication/who-counts/.

Metraux, Stephen, and Dennis P. Culhane. 1999. "Analyzing Shelter Stays and Shelter Stay Patterns Using Administrative Data: An Overview of Practical Methods." Homeless Data Users Conference #3. Washington, DC: US Department of Health; Human Services.

Smith, Dakota, and Doug Smith. 2018. "More Sidewalk Tents, but Fewer People Living in Them? The 2018 Homeless Count's New Math." *Los Angeles Times*, no. 15 (July). http://www.latimes.com/local/lanow/la-me-ln-homeless-count-analysis-20180711-story.html.