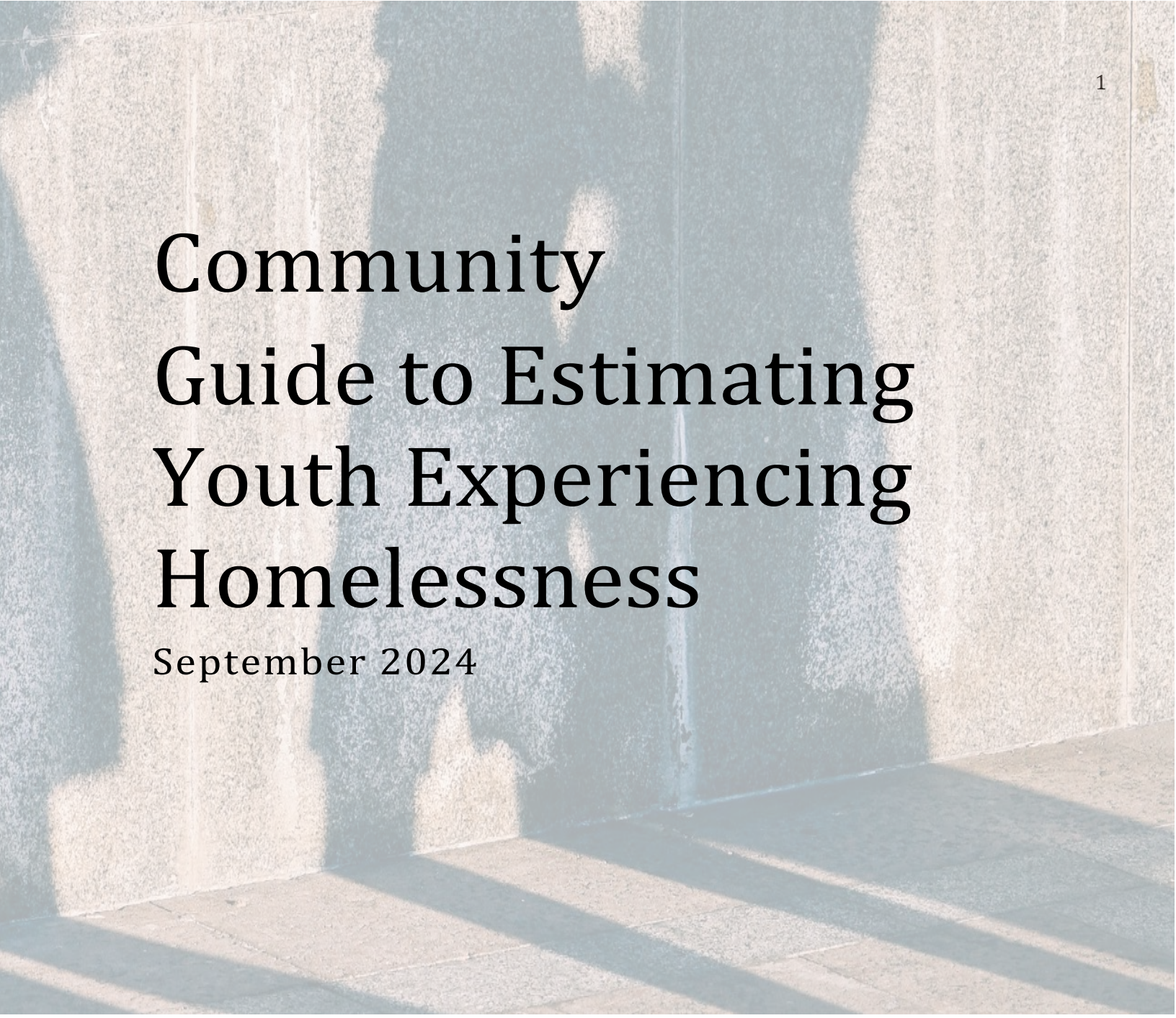


Community Guide to Estimating Youth Experiencing Homelessness

September 2024

Using community and program data to estimate the prevalence of youth ages 13-25 experiencing homelessness

Center on Poverty and Community Development
Case Western Reserve University

Table of Contents

Welcome

Welcome to the brief community guide for estimating a local population of youth experiencing homelessness.

The Center on Poverty and Community Development at Case Western Reserve University, under contract with the U.S. Department of Housing and Urban Development (HUD), conducted a study to assess the value of integrated data in estimating the prevalence and probability of youth experiencing homelessness in Cuyahoga County, Ohio.

This guide was developed from study findings to support other communities in estimating local populations of youth experiencing homelessness.

Background

At present, a primary way we know a young person is experiencing homelessness or housing insecurity is when they interact with the homeless services system (such as visiting a shelter location or interacting with an outreach worker). These interactions are tracked in a federally-funded homeless management information system, or HMIS. In the case of public school students, if they connect with a homeless education liaison, these interactions are tracked in McKinney-Vento data reported to the U.S. Department of Education.

HMIS is a helpful tool to understand people who connect with the homeless service system but can't account for all those who do not. For certain subpopulations, such as youth and young adults, we know they connect less frequently than older adults. Instead of shelter locations, youth and young adults frequently stay with friends, couch surf[1] or sleep in their cars.[2] This tendency for youth and young adults to leverage social support instead of formal homeless services has made it challenging to estimate the population.[3]

We designed this study to explore whether regional administrative data could complement HMIS and McKinney-Vento data to improve prevalence estimates of youth and young adults experiencing homelessness. To do this, we

[1] Curry, S. R., Morton, M., Matjasko, J. L., Dworsky, A., Samuels, G. M., & Schlueter, D. (2017). Youth Homelessness and Vulnerability: How Does Couch Surfing Fit? *American Journal of Community Psychology*, *60*(1–2), 17–24. https://doi.org/10.1002/ajcp.12156

[2] Ha, Y., Narendorf, S. C., Santa Maria, D., & Bezette-Flores, N. (2015). Barriers and facilitators to shelter utilization among homeless young adults. *Evaluation and Program Planning*, *53*, 25–33. https://doi.org/10.1016/j.evalprogplan.2015.07.001

[3] Morton, M. H., Dworsky, A., Matjasko, J. L., Curry, S. R., Schlueter, D., Chávez, R., & Farrell, A. F. (2018). Prevalence and Correlates of Youth Homelessness in the United States. *Journal of Adolescent Health*, *62*(1), 14–21. https://doi.org/10.1016/j.jadohealth.2017.10.006

leveraged administrative data (data collected by organizations and service providers), bringing separate data sets together to improve estimation and policy approaches for youth.

The population of youth and young adults in the data used for this project were from Cuyahoga County, Ohio. The county has the second highest population in the state and includes the city of Cleveland. Census estimates indicate almost 43% of the county identifies as non-white, with 11% of households speaking a language other than English at home. Additionally, 21% of the population is aged 17 and under, while 23% are between the ages of 18 and 34 and almost 40% are single-person households. Counties outside of Ohio that may be used for comparison include Hennepin County (Minneapolis, MN) and Erie County (Buffalo, NY). [4]

Roughly three-quarters of the sample of youth and young adults who were included in our analysis identified as non-white. Additionally, the sample exhibited indicators of adversity that may not be represented in samples found elsewhere. For example, between 62 and 66% were enrolled in state food assistance and roughly 20% were included in a child welfare report at some point. It is important to consider the unique population of your community in developing and conducting a local analysis.

This brief community guide walks through the research process, results, and how your community can implement a similar project. To complete the type of analysis described here, you must work through accessing individual-identifiable data on individuals in your own community. If this isn't an option, consider working with each of the entities described in

[4] Cuyahoga County Planning Commission (2023). Our county: The 2023 data book. Available at https://s3.countyplanning.us/wp-content/uploads/2023/12/Our-County-2023-reduced.pdf

this guide to obtain de-identified aggregate data. While you will not be able to conduct the analysis we did, you will have helpful descriptive information that is relevant to your community. Establishing connections with these entities will also develop relationships that you can build on as your community considers how to respond to youth homelessness.

A detailed comprehensive report with all analysis and findings from our study can be found on HUD's HUDUser research portal.

Steps to conduct your own analysis

Step one: Inventory existing data sources

Start by identifying all potential sources that collect or track information on housing unstable youth. If you're not sure, consider forming a workgroup or community board to talk through how a collaborative project can start locally. Consider the following questions as a starting point:

- In your community, who manages the Homeless Management Information System (HMIS)?
- Does your local Continuum of Care (CoC) share individually identifiable data with other CoC's in your state?
- Would they share data with your project? If so, what types of safeguards need to be considered? If not, are there ways to learn from communities that do share data to find solutions that can facilitate sharing?
- Does your state or local education system share information for research purposes? Would they consider sharing McKinney-Vento data for a specific project?

Additional information on data inventories is available from the Bloomberg Center for Government Excellence at Johns Hopkins University.[5] More information on data sharing is available through the National Neighborhood Indicators Partnership (NNIP)[6] and Actionable Intelligence for Social Policy (AISP).[7]

[5] Bloomberg Center for Government Excellence. Data inventory guide. https://labs.centerforgov.org/data-governance/data-inventory/
[6] National Neighborhood Indicators Partnership. NNIP lessons on local data sharing. https://www.neighborhoodindicators.org/library/guides/nnip-lessons-local-data-sharing
[7] Actionable Intelligence on Social Policy. Data sharing FAQ. https://aisp.upenn.edu/data-sharing-faq/

Note: Choose parameters for your data

It is critical to decide how your will define homelessness within your analysis. Will you only consider youth who enter shelter or receive HMIS-documented services? For our analysis, we used a broad approach and included youth considered housing unstable to generate our estimates. This decision was based on previous research finding youth less likely to interact with the formal homeless service system.

What age range of youth and young adults will your project want to estimate and why? If you choose youth under the age of 18, consider the availability of data in your community. Your team will also need to choose the date range of your estimates. These can follow calendar years (January 1st to December 31st), school years (August through June) or some other time frame. It is useful to consider nuance within the specific data sources if there are gaps in data collection or specifics that should be considered. For example, McKinney-Vento data aren't always accurately recorded during summer months, as staff struggle to keep up with families when kids are out of school. For our analysis, we had robust data from 2017, 2018, and 2019, so these were the three years used for analysis.

Steps to conduct your own analysis

Step two: Talk meaningfully with those with lived experience of homelessness and homeless service providers

Talk with youth and young adults with lived experience as well as data managers or program administrators to understand 1) how, when, and why individuals do or do not interact with the homeless system of care, and 2) how data are collected by system providers when they do receive services.

Here are some questions you might include in your conversations with individuals with lived experience:

- What factors influence or impact the decision to engage with homeless services?
- Were there notable experiences with service providers that made you feel heard or supported?
- Were there notable experiences with service providers that made you feel dismissed or disrespected?
- When service providers collected information about you at intake, did they ask you how you identified your race, gender identity, or sexual orientation? If so, what choices (if any) did you have to make about these parts of your identity?

Here are some questions you might include in your conversations with data providers and administrators:

- How do you collect personal information from people when they come to you for services?
- Which staff are asking those questions (or entering them into a data form)?
- What type of training and quality control are used to ensure the data are being entered accurately?
- Do staff receive training in any of the following areas?
    - Trauma-informed care
    - Cultural sensitivity
    - Client-centered care
    - Racial equity
    - Sexual and gender minority care
- Is there a way to know if someone is housing unstable or experiencing homelessness in the current data system?
- If so, what fields do you currently have that can provide that information?
- How often are data updated in the system?
- How long are historical data files kept?

Steps to conduct your own analysis

Note: Consider the security and privacy of data early

The data you will be accessing contains personally identifiable information (PII) and should be stored in a secure environment. Remember, if you are unable to obtain PII to replicate this analysis, consider working with your community partners to receive de-identified aggregate data instead. This information can provide you with rich description of your community and establishes (or expands) existing connections between providers, policy makers, researchers, and those with lived experience. Security measures that consider federal, state, and local regulations and guidelines are required to access and analyze PII. Be sure to consult appropriate legal counsel before requesting data so your team can get the necessary safeguards in place. Additional information on legal considerations can be found at AISP.[8]

Step three: Establish data use agreements with trusted partners

Data use agreements are contracts between two parties who decide to share information. They are typically developed and managed through legal counsel who

[8] Actionable Intelligence on Social Policy. Finding a Way Forward: How to Create a Strong Legal Framework for Data Integration. https://aisp.upenn.edu/legal-framework/

understand the appropriate security and privacy regulations described above. Language in the agreements include the type of data and reason for sharing, length of time they will be used, storage and use stipulations, as well as any program-specific limitations.

A web-based manual maintained by the Abdul Latif Jameel Poverty Action Lab has helpful information on establishing data use agreements when linking administrative data.[9]

[9] Abdul Latif Jameel Poverty Action Lab. Handbook on Using Administrative Data for Research and Evidence-based Policy. admindatahandbook.mit.edu/

Steps to conduct your own analysis

Step four: Access and link data

Once you've established appropriate data use agreements and located a safe repository for storing data, it's time to access it. Providers may choose to share data through a secure access site, file transfer, or direct access to their respective data systems for download. It is critical to have identifiable information in each of these data sets so they can be linked together.

As there is typically no universal identifier to know if an individual is the same person in one data set (food stamp recipient data) as the other (HMIS), having full name, date of birth, gender and racial identity can help in linkage. Our project used deterministic and probabilistic matching techniques to link individuals across data sets. Your team may prefer one technique over the other.

At the end of the matching process, you will have one master file with all individuals you've decided to include in your project, with variables that show which data sources they appeared in. For example, if Jill Brown was in the food stamp recipient and HMIS data but not in any other data sources, the columns for HMIS and food stamp recipient indicators will be filled in and the other columns will be blank or have a no indicator.

Step five: Analyze the data

Now that your team has a linked data set (we called ours a registry) of individuals that shows what data sources each individual appears in, you are able to conduct an analysis to assess who is more likely to not have homeless service provider records, yet still be experiencing homelessness based on other data. You can also estimate the potential number of individuals who do NOT appear in any of the data you have. Using multiple systems estimation (MSE) allows for doing so.

Multiple systems estimation (MSE)

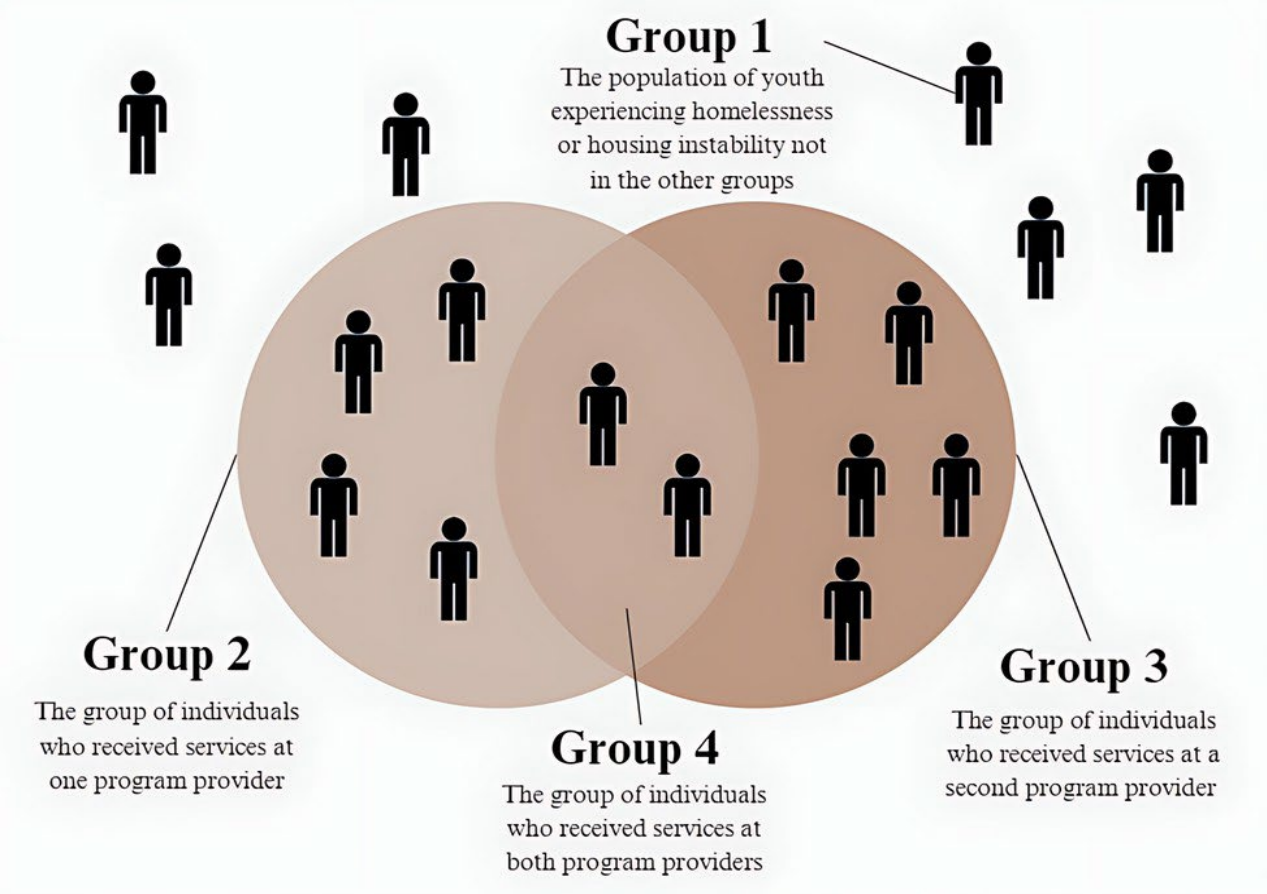
MSE is a statistical method of calculating the likelihood of individuals who aren't included in data from multiple sources. A linked data set of youth found in any of the programs (the registry) allowed our research team to conduct multiple systems estimation (MSE) to approximate the amount of youth experiencing homelessness or housing instability who were not included in the data - think of it as a way to understand who isn’t being counted.

Estimating all individuals who are included and excluded from a population allows for a full picture. MSE allows us to potentially understand the total group of youth experiencing housing instability - those who seek formal services and those who do not.

Multiple systems estimation

Using lists of individuals from each data source, the number of those who are found on each separate list (groups 2 and 3) as well as those who are on both lists (group 4), along with assumptions on how the lists are generated and relate to each other, allow for a statistical estimation of the group of individuals who aren't on any list (group 1).

Our registry provided the data input needed for the MSE analysis. There are a number of assumptions and data elements to consider with this type of analysis and it's useful to have help from a statistician or trained researcher. If you have a college or university in your community, consider reaching out to their humanities or math departments about professors or students who may be willing to work with you. An article by Chan, Silverman and Vincent (2021) was very helpful to our team.[10]



[10] Lax Chan, Bernard W. Silverman & Kyle Vincent (2021) Multiple Systems Estimation for Sparse Capture Data: Inferential Challenges When There Are Nonoverlapping Lists, Journal of the American Statistical Association, 116:535, 1297- 1306, DOI: 10.1080/01621459.2019.1708748

What we found

Our analysis identified approximately 6,000 youth between the ages of 13 and 25 who experienced homelessness or housing instability in each study year.

Results from the MSE indicated that there **may be up to three times** more youth facing homelessness than are documented in administrative data systems.

Next steps

1 | Interpret

Once you've conducted your own analysis of data on youth in your community, it's important to interpret what your results mean.

2 | Share!

Consider sharing preliminary findings with community members, those with lived experience of homelessness, and service providers. They will have important feedback, insights, and context for some of what you see in the results. Be sure to develop a process for incorporating feedback into your work. Not doing so can quickly erode the relationships you've worked hard to develop.

3 | Get to work

Once your team believes you've developed a deep understanding of the findings, consider how your community can use them to affect change.
Importantly, consider the weight of your findings - if there really are three times more youth facing homelessness in your community, how can you intervene?

No community is alone

Youth experiencing homelessness and housing instability appear across the Unites States, in urban, suburban and rural locations as well as in schools, shelters, and hotels. During the time when they are learning who they are, housing instability forces them to make decisions and consider options that may be new to them.

Youth and young adults are developmentally learning to plan for the future, trust or be wary of the people around them, and situate their individual identities within the world. We can support these developmentally appropriate life stages by providing additional options, presenting alternate approaches to problem solving, being reliable and predictable older adults, and remembering that they won't have it all figured out on their own.

If you get stuck in trying to estimate counts and prevalence estimates for your community, ask for help! Actionable Intelligence for Social Policy (AISP) has compiled a searchable dataset[11] of partner entities that share and integrate data. Our team at the Center on Poverty and Community Development is happy to share ideas, talk through problems, or offer suggestions.

[11] Actionable Intelligence for Social Policy. https://aisp.upenn.edu/data-sharing-landscape/

This research was performed under contract with the U.S. Department of Housing and Urban Development.

The work that provided the basis for this publication was supported by funding under an award with the U.S. Department of Housing and Urban Development Grant Number H-21693CA. The substance and findings of the work are dedicated to the public. The authors are solely responsible for the accuracy of the statements and interpretations contained in this publication. Such interpretations do not necessarily reflect the views of the Government.

Thank you to our valued community partners!

APLACE4ME

Greater Cleveland Food Bank

Ohio Department of Job & Family Services

Ohio Department of Education & Workforce

Health and Human Services Cuyahoga County, Ohio

Photo credit: Matt Shiffler Photography; mural by Justin Michael W

An example of MSE

Based on the methods proposed by Lax Chan, Bernard W. Silverman & Kyle Vincent in Multiple Systems Estimation for Sparse Capture Data: Inferential Challenges When There Are Nonoverlapping Lists, Journal of the American Statistical Association, 116:535, 1297-1306, DOI: 10.1080/01621459.2019.1708748 # https://rdrr.io/cran/SparseMSE/man/estimatepopulation.0.html

Here we illustrate an approach for organizing administrative data to utilize a method for estimating the true size of a population when that is difficult to observe directly, called Multiple Systems Estimation (MSE). In order to take advantage of MSE, it is necessary to have information about the existence of at least some members of the target population recorded in multiple data sources, and for the observations made in one data source to be linkable to the other data sources, such that it is possible to determine whether a person recorded in data source A is only recorded in A, or if they are also recorded in some combination of B, C, and D. Our analysis was completed using R.[12]

The R Markdown file for this example, with accompanying sample data, can be accessed at: https://github.com/PovertyCenterCLE/Multiple-Systems-Estimation

[12] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

An example of MSE

To demonstrate this approach, we created an example synthetic dataset, mse_synth_data.csv, of youth facing homelessness in Cuyahoga County, Ohio in 2017, 2018, and 2019, as recorded in four linked administrative data sources (named list1, list2, list3, and list4). It includes a series of binary flags that indicate if the individual was identified in one of four administrative databases as having experienced homelessness or housing instability in a given year (e.g., list1_17).

Each record also includes basic personal characteristics used to stratify youth into homogenous subgroups along the lines of race (white, non-white), sex assigned at birth (M, F), and age group (13-18, 19-25). A date of birth was randomly assigned to each record based on age group.

This code accomplishes three things. First, it organizes and reshapes a linked administrative dataset into the format needed to carry out MSE. Second, it demonstrates an efficient method for carrying out the same set of analyses on numerous subsets of the data simultaneously, and without excessive code repetition. Third, it extracts some key pieces of information from the MSE output to allow researchers to begin drawing conclusions about the population of homeless youth in Cuyahoga County, the degree to which the homeless population is 'hidden' and thus unlikely to be able to be reached by service providers, and for better understanding differences between different subpopulations of homeless youth, along the lines of race, sex, and age.

Table 1 below shows the initial data structure.

Table 1 – Initial Data Structure

sex	race_cat	byear	dob	id	list1_17	list1_18	list1_19	list2_17	list2_18	list2_19	list3_17	list3_18	list3_19	list4_17	list4_18	list4_19
M	NONWHITE	1997	1997-05-01	14,950	1	0	0	0	0	0	0	0	0	0	0	0
M	NONWHITE	2003	2003-06-18	12,286	0	0	1	0	0	0	0	0	0	0	0	0
F	NONWHITE	1992	1992-01-16	28,121	0	0	0	0	0	0	1	0	0	0	0	0
F	NONWHITE	2002	2002-11-01	2,754	0	0	1	0	0	1	0	0	0	0	0	0
F	NONWHITE	2000	2000-04-03	27,031	0	0	0	0	0	0	0	0	0	0	0	0
M	WHITE	2003	2003-03-01	28,654	0	0	0	0	1	0	0	0	0	0	0	0
M	NONWHITE	2002	2002-01-04	11,152	0	1	0	0	0	0	0	0	0	0	0	0
F	NONWHITE	2002	2002-02-15	21,490	0	0	0	0	0	0	0	0	0	0	0	0
F	NONWHITE	1994	1994-11-16	18,611	0	0	0	0	0	0	0	0	0	0	0	0
M	NONWHITE	1995	1995-08-29	19,237	0	0	0	0	0	0	0	0	0	0	0	0

An example of MSE

To begin, we identify and retain the records for individuals observed in at least one list in a year in which they were between the ages of 13 and 25. We discard records for individuals who never appear on any list, or who only appear on a list when they are outside the age range for this study.

```
dat2 <- dat %>%
  mutate(
    elig17 = ifelse(dob %within% interval(as_date("1992-01-01"), as_dat
e("2004-12-31")), 1, 0),
    elig18 = ifelse(dob %within% interval(as_date("1993-01-01"), as_dat
e("2005-12-31")), 1, 0),
    elig19 = ifelse(dob %within% interval(as_date("1994-01-01"), as_dat
e("2006-12-31")), 1, 0),
    observed17 = ifelse(list1_17 + list2_17 + list3_17 + list4_17 > 0,
1, 0),
    observed18 = ifelse(list1_18 + list2_18 + list3_18 + list4_18 > 0,
1, 0),
    observed19 = ifelse(list1_19 + list2_19 + list3_19 + list4_19 > 0,
1, 0),
    # id = row_number()
  )
```

An example of MSE

From the initial pool of records, we identified those for youth who were recorded as facing homelessness in a year in which they were in the age range for this study.

```
dat3 <- dat2 %>%
  filter(
    (elig17 == 1 & observed17 == 1) |
    (elig18 == 1 & observed18 == 1) |
    (elig19 == 1 & observed19 == 1)
  )
```

Next, we reshape the dataset from “wide” (all information about a person contained in a single row of data, with more columns) to “long” (with one row for each administrative list and year in which an individual was recorded, such that a person recorded as facing homelessness in all four lists every year from 2017-2019 span 12 data rows). This is done to make it easier to retain only information about people in the years in which they were recorded.

```r
dat_elig <- dat3 %>%
  select(id, starts_with("elig")) %>%
  pivot_longer(cols = -id) %>%
  mutate(source = str_sub(name, 1, -3),list_yr = as.numeric(paste0("20"
, str_sub(name, -2)))) %>%
  rename(eligible = value) %>%
  select(id, list_yr, eligible) %>%
  distinct()


dat_long <- dat3 %>%
  select(-starts_with("elig")) %>%
  pivot_longer(cols = c(starts_with("list1_"), starts_with("list2_"), s
tarts_with("list3_"), starts_with("list4_"))) %>%
  mutate(source = str_sub(name, 1, -3), list_yr = as.numeric(paste0("20
", str_sub(name, -2)))) %>%
  rename(seen_in_source = value) %>%
  mutate(
    age = list_yr - byear,
    age_cat = if_else(
      list_yr - byear <= 18, "13-18", "19-25"
    )) %>%
  select(sex, byear, race_cat, age, age_cat, list_yr, source, seen_in_s
ource, id)


dat_long2 <- dat_long %>%
  left_join(dat_elig) %>%
  filter(seen_in_source == 1 & eligible == 1) %>% # keep only person-ye
ar records in which an individual was observed while within the study a
ge range
  select(-eligible) %>%
  distinct()

autofit(flextable(dat_long2 %>% slice_sample(n = 10)))
```

An example of MSE

Table 2 shows the reshaped data.

Table 2 Long Form Table

sex	byear	race_cat	age	age_cat	list_yr	source	seen_in_source	id
M	1998	NONWHITE	19	19-25	2,017	list1_	1	14,515
F	1999	NONWHITE	19	19-25	2,018	list1_	1	7,117
F	1997	WHITE	21	19-25	2,018	list1_	1	8,897
M	1998	WHITE	20	19-25	2,018	list3_	1	17,714
M	2001	WHITE	16	13-18	2,017	list2_	1	13,444
F	1997	WHITE	21	19-25	2,018	list3_	1	8,865
M	2003	WHITE	16	13-18	2,019	list3_	1	13,967
M	2001	NONWHITE	17	13-18	2,018	list3_	1	11,003
F	2000	WHITE	19	19-25	2,019	list1_	1	9,051
F	2002	NONWHITE	16	13-18	2,018	list1_	1	2,363

Next, we split the dataset into strata of relatively homogenous subgroups of youth, defined by every unique combination of sex, race, age group, and year (e.g., sex == MALE, race = NONWHITE, age group = 13-18, list year == 2017).

```
make_strata <- dat_long2 %>%
  group_by(sex, race_cat, age_cat, list_yr) %>%
  summarise(
    n_kids = n_distinct(id)
  ) %>%
  ungroup() %>%
  mutate(strata = row_number())

dat_long3 <- make_strata %>%
  left_join(dat_long2)
```

We then reshape our dataset again, from “long” to “wide” to prepare it for the analysis step.

```
dat_wide <- dat_long3 %>%
  pivot_wider(names_from = source, values_from = seen_in_source, values
_fill = 0)

autofit(flextable(dat_wide %>% slice_sample(n = 10)))
```

An example of MSE

The final data-reshaping step is to turn the person-level dataset into an aggregated strata-level dataset needed for estimating population sizes with MSE. First, we create a 'split' data frame with 24 rows- one for each stratum. This is accomplished with the base::split function, which groups observations according to the specified variable (strata in this case) and collapses the subset of records within each subgroup into a new column named value by default. In other words, whereas tables are typically organized to include a single piece of information per cell, each cell is a value that is an entire dataset with multiple columns.

Structuring the records in this way comes in handy when the same set of analyses are to be performed on many subsets of the data. For one, the code only needs to be written once and run once, as it is applied simultaneously to each subset of the data. In other words, it eliminates the need for tedious (and error-prone) copying, pasting, and re-running to run the code on each subgroup. Second, it saves the output of each analysis step as a new column or set of columns in the data frame, making it easy to directly compare the results for each subgroup. Table 3 shows the resulting strata.

Table 3 Strata

sex	race_cat	age_cat	list_yr	n_kids	strata	byear	age	id	list1_	list2_	list3_	list4_
M	NONWHITE	19-25	2019	823	18	1999	20	16,664	1	0	0	0
M	NONWHITE	19-25	2019	823	18	1999	20	16,686	1	0	0	0
M	NONWHITE	19-25	2018	998	17	1993	25	15,722	0	0	1	0
F	NONWHITE	13-18	2018	1,246	2	2002	16	1,603	0	1	1	0
M	NONWHITE	19-25	2017	1,069	16	1993	24	15,345	0	0	1	0
F	WHITE	13-18	2019	370	9	2006	13	4,843	0	1	0	0
F	NONWHITE	19-25	2018	1,245	5	1995	23	7,311	0	0	1	0
F	NONWHITE	13-18	2017	1,137	1	2001	16	785	0	1	0	0
F	NONWHITE	13-18	2018	1,246	2	2005	13	2,071	1	1	0	0
F	WHITE	19-25	2018	297	11	1993	25	8,781	0	0	1	0

```r
by_strata <- dat_wide %>%
  split(.$strata) %>%
  enframe()

vars <- dat_wide %>%
  select(strata, sex, age_cat,list_yr, race_cat, n_kids) %>%
  distinct()

by_strata <- vars %>%
  bind_cols(by_strata)
```

An example of MSE

The function make_matrix, defined below, takes our split dataset and the strata id (named 1:24) as inputs. For each stratum, it drops any observation list in which no individuals were observed, and creates a matrix of the form needed to run the MSE analysis.

```
make_matrix <- function(dw, strata){

# Create a dataframe showing the combination of list observations for e
ach record. The variable lp indicates the list pattern for each indivia
ual, based on the column order of the lists.

d2<- dw %>%
  unite(lp, -c(sex:id), remove = FALSE) %>%
  select( lp:ncol(.))


# For each subgroup, count the number of unique contributions made by e
ach list (i.e., the number of individuals observed in HMIS alone, and n
ot in combination with any other list.). Drop any lists that do not mak
e at least one unique/independent contribution

d2$nlists <- rowSums(d2[,-1])



todrop <- d2 %>%
  filter(nlists == 1) %>%
  summarise(across(-c(lp, nlists), sum)) %>%
  t() %>%
  as.data.frame() %>%
  rownames_to_column() %>%
  filter(V1 == 0)

todrop <- todrop$rowname



# remove any zero-contribution lists from the dataset, re-create the va
riable lp to show the observation patterns of only the included lists.
d2 <- d2 %>%
  select(-c(all_of(todrop), nlists)) %>%
  select(-lp) %>%
  unite(lp, remove = FALSE)



d3 <- d2 %>%
  group_by( lp) %>%
  tally() %>%
  ungroup()



dt <- merge(d2, d3)
d0<- distinct(dt, lp, .keep_all = TRUE) %>%
  mutate(across(everything(), \(x) replace_na(x, 0)))

d00 <- d0 %>%
  select(-lp)

d00

}
```

An example of MSE

In the following step, we begin adding new columns to the split dataset. Specifically, the variable mat contains the observation matrix for each stratum, while vars shows the homelessness lists in which individuals within that stratum were observed at least once. Depending on the population of interest and the available homeless indicator databases, vars may or may not be relevant. In the example data, we can see that there were no individuals observed in the list4_ (juvenile delinquency & department of child & family services) databases in several of the age 19- 25 strata. Thus, this list does not appear in the list history matrices for these strata. The observation matrix for stratum 1, Non-white females, aged 13-18 in 2017, is shown below in Table 4.

```
by_strata2 <- by_strata %>%
  mutate(mat = map(value, ~make_matrix(.x))) %>%
  mutate(vars = map(mat, names)) %>%
  select(-c(value))

autofit(flextable(by_strata2[[8]][[1]]))
```

Table 4 Matrix with Counts

list1_	list2_	list3_	list4_	n
0	0	0	1	17
0	0	1	0	76
0	1	0	0	542
0	1	0	1	5
0	1	1	0	19
0	1	1	1	3
1	0	0	0	363
1	0	0	1	1
1	0	1	0	12
1	0	1	1	7
1	1	0	0	75
1	1	1	0	12
1	1	1	1	5

An example of MSE

The function mse_fun applies multiple functions from the SparseMSE package to each stratum, and adds the output of those functions as new columns in the split dataframe. For more details on these functions, please refer to the journal article and R documentation referenced at the beginning of this document.

checkident performs a test to determine whether it is possible to generate a consistent estimate of population size given the observation history for a given set of observations. A value of 0 means that the conditions for consistent estimation are met, while values of 1, 2, and 3 indicate issues of MLE existence, identifiability, or both.

The estimatepopulation.0 function estimates the population for each stratum, including the 'hidden figure' of homeless youth who are not observed on any list.

The output of this function is saved as the variable est_stepwise. The variable est_pt_sw is the point estimate of the total population, including both the observed and unobserved (hidden) cases. We use this value to construct additional useful variables, such as the number of unobserved youth, and the ratio of unobserved:observed youth. See Table 5 for example results.

```
mse_fun <- function(df) {
by_strata2 <- df %>%
  mutate(checkident = map_dbl(mat, ~checkident(.x, mX=0,verbose=FALSE))
)


by_strata3 <- by_strata2 %>%
  mutate(est_stepwise = map(mat, ~estimatepopulation.0(.x,  quantiles=c
(0.025,0.975)))) %>%
  mutate(
    est_low_sw = map_dbl(est_stepwise, ~pluck(.x, "estimate", "2.5%" ))
,
    est_pt_sw = map_dbl(est_stepwise, ~pluck(.x, "estimate", "point est
." )),
    est_hi_sw = map_dbl(est_stepwise, ~pluck(.x, "estimate", "97.5%" ))
)

test <- by_strata3 %>%
  mutate(across(est_low_sw:est_hi_sw, round),
         est_unlisted = est_pt_sw - n_kids,
         unlisted_to_listed = round(est_unlisted/n_kids, 1)) %>%
  select(sex, age_cat, list_yr, race_cat, n_kids, est_pt_sw, est_low_sw
, est_hi_sw, est_unlisted, unlisted_to_listed, checkident)



}

by_mse <- mse_fun(by_strata2)

autofit(flextable(by_mse))
```

An example of MSE

Table 5 Final Table

sex	age_cat	list_yr	race_cat	n_kids	est_pt_sw	est_low_sw	est_hi_sw	est_unlisted	unlisted_to_listed	checkident
F	13-18	2017	NONWHITE	1,137	3,785	3,192	4,550	2,648	2.3	0
F	13-18	2018	NONWHITE	1,246	4,235	3,603	5,036	2,989	2.4	0
F	13-18	2019	NONWHITE	1,366	4,714	4,075	5,504	3,348	2.5	0
F	19-25	2017	NONWHITE	1,183	2,974	2,642	3,382	1,791	1.5	0
F	19-25	2018	NONWHITE	1,245	3,746	3,270	4,333	2,501	2.0	1
F	19-25	2019	NONWHITE	1,095	2,993	2,581	3,520	1,898	1.7	0
F	13-18	2017	WHITE	373	1,823	1,226	2,838	1,450	3.9	0
F	13-18	2018	WHITE	411	2,822	1,690	4,955	2,411	5.9	0
F	13-18	2019	WHITE	370	2,671	1,555	4,836	2,301	6.2	0
F	19-25	2017	WHITE	314	1,281	925	1,844	967	3.1	0
F	19-25	2018	WHITE	297	971	736	1,330	674	2.3	0
F	19-25	2019	WHITE	248	569	449	760	321	1.3	0
M	13-18	2017	NONWHITE	1,222	4,845	4,001	5,946	3,623	3.0	0
M	13-18	2018	NONWHITE	1,277	5,414	4,460	6,654	4,137	3.2	0
M	13-18	2019	NONWHITE	1,352	5,356	4,470	6,494	4,004	3.0	0
M	19-25	2017	NONWHITE	1,069	3,817	3,248	4,536	2,748	2.6	0
M	19-25	2018	NONWHITE	998	3,500	2,975	4,165	2,502	2.5	0
M	19-25	2019	NONWHITE	823	1,941	1,673	2,294	1,118	1.4	1
M	13-18	2017	WHITE	383	1,765	1,253	2,578	1,382	3.6	0
M	13-18	2018	WHITE	423	1,176	929	1,544	753	1.8	0
M	13-18	2019	WHITE	406	1,945	1,368	2,870	1,539	3.8	0
M	19-25	2017	WHITE	355	992	777	1,317	637	1.8	1
M	19-25	2018	WHITE	271	720	569	947	449	1.7	0
M	19-25	2019	WHITE	203	723	514	1,074	520	2.6	2