Constructing Center-Based Cluster-Level Metrics to Use in Household Level Analysis: A Tutorial for NSECE Data

Introduction

This tutorial illustrates the process of creating cluster-level aggregates using Center-based providers and how these metrics are integrated into the NSECE Household dataset for analysis. Due to the disclosure risk associated with data at the provider cluster-level, analyses using NSECE clusters require Level 2 restricted access. To learn more about the Level 2 data access process, interested users can send an email to the NSECE project staff at nsece@norc.org.

This tutorial assumes that the user is familiar with the NSECE survey design - particularly the concept of the "provider cluster" as detailed in document the "PSU and Cluster Weights User Guide." It also assumes that the user is aware of the application and review process that all Level 2-restricted use (L2) requests go through, as outlined in document "Application for Access to NSECE Levels 2 or 3 Restricted Use Data (revised 6/10/16)."

The tutorial walks the reader through the basic steps in the data management process that the NSECE team carries out when data users request cluster-level aggregate information. Although the tutorial focuses on center-based providers, users may apply the same logic and step-by-step instructions to create cluster-level aggregates with the home-based provider file. The tutorial provides a step-by-step description of critical points and provides supplementary Excel files with household and center-based provider example cases (not the entire set of center-based or household observations). The intention of the Excel files is to show the reader how datasets are constructed and what to expect in the data that will be delivered. While the NSECE data support team may be able to provide analytic suggestions for supply and demand analysis, the selection of variables as well as the interpretation of results is ultimately the user's decision and responsibility.

As per our disclosure guidelines, it is important to know that NORC only provides users with a dataset that includes simulated cluster IDs. This simulated cluster file allows users to create variables at the cluster-level that are similar, but not exactly the same, to the real ones. Once the user generates the cluster aggregates they are interested in, NORC will run the user's syntax based on real cluster IDs. NORC will merge this data file to the household data file and this supplemented household file will be made available to the user. (A cluster-level file of aggregates generated from the real IDs is not provided directly to the research team.) This tutorial also uses a dataset that has been altered to illustrate the process. The process consists of four steps:

1. Select Center-Based (CB) Provider questionnaire variables for cluster-level metrics

2. Calculate cluster-level metrics
3. Inspect provider cluster-level summary statistics
4. Merge data onto the Household dataset

This tutorial is part of the support that NORC provides to NSECE users interested in conducting supply and demand analyses.

Supporting Materials

The reader is encouraged to consult the following documentation and guidelines:

National Survey of Early Care and Education Project Team. (2016). Households' Geographic Access to Center-based Early Care and Education: Estimates and Methodology from the National Survey of Early Care and Education. OPRE Report # 2016-08, Washington DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Service. Available at:
https://www.acf.hhs.gov/sites/default/files/opre/hh_geoaccessto_cb_ece_toopre_042916_b508.pdf

Additional information on cluster-level and county-level sampling weights is available in NSECE PSU and Cluster Weights Users' Guide (NSECE, 2016).

https://www.researchconnections.org/childcare/resources/34792/pdf

Step 1: User selects Center-based variables for cluster-level metrics

In this step, the user identifies variables of interest from the CB data file, based on his or her research question and submits a request to the NSECE support team with said variables. For this tutorial, we selected CB_SERVE_0TO3YRS and CB_B1_5_STDRATE_HRLY_2YR_TC, as shown in Table 1 below. The user may also choose to create a variable by recoding existing variables, for instance, into a dummy variable. Typically, the user supplies programming code in SAS or Stata (or the equivalent logic) to NORC in order to recode variables if the desired variable does not exist.

Table 1: Variables used for tutorial

Variable	**Description**
CB_SERVE_0TO3YRS	Indicator for whether center-based provider serves:Age Category 0-3 years old
CB_B1_5_STDRATE_HRLY_2YR_TC	Hourly rate for highest price of care that the provider charges for full time enrollment:Age Category 2 year

As will become apparent throughout the tutorial, the selection of categorical variables or continuous variables will have a critical impact on the quality of cluster level measures. For instance, the number of centers per provider cluster may be as few as one or a couple of providers, depending on how the

variable is defined and whether all or only a small subset of centers are included in the created measure. The user is strongly encouraged to think about the statistical properties of the variable of interest in light of the possible interpretation of the end result.

The file named "**Data_A.xls**" shows the initial setup of the dataset with centers as the unit of observation, which in turn are linked to one or more provider clusters. This reduced dataset is based on a subset of clusters (which will serve as an illustration). It is comprised of 1,221 rows; namely, 1,120 unique centers within 50 provider clusters. As noted before, the sample data in this tutorial have been simulated to protect sensitive information while retaining measurement properties of the clusters. The simulated data will illustrate how the rest of the steps work. The variables are as follows (Table 2).

Table 2: Variables included in Data_A.xls

Variable	**Description**
L2_PCLUSTER_FALSE_ID_2	False cluster id for provider clusters
CB_METH_FALSE_CASEID	False case id for sample centers
L2_STRATA_F	Altered strata from the provider sampling frame
L2_CB_WEIGHT_PCLUSTER_FALSE	False weights used for sample CB cases
CB_SERVE_0TO3YRS_F	False version of variable of interest
CB_B1_5_STDRATE_HRLY_2YR_TC_F	False version of variable of interest

Exercises for users:

1.1. How many unweighted center-based providers does cluster 777003 have?

1.2. How many unweighted center-based providers does cluster 777033 have?

1.3. How does the difference in the number of centers per cluster affect the calculation of cluster-level metrics?

Answer Key:

1.1. Provider cluster 777003 has 14 unique center-based providers.

1.2. Provider cluster 777033 has 1 unique center-based provider.

1.3. Some estimates at the cluster level cannot be calculated if the number of centers per cluster is low, as step 2 in this tutorial explains. Note that the key issue is not just the number of centers, but the number of centers with specific characteristics. For instance, provider cluster 777003 has 14 centers, but only 4 of them have a valid price for 2 year-olds.

Step 2: User generates cluster-level metrics using altered Cluster IDs

In this step, the user creates metrics aggregated at the cluster level based on original or recoded variables at the provider level. The file named "**Data_B.xls**" shows metrics at the cluster provider level based on altered data. The file has 51 rows, with one record for each of the 50 provider clusters. The variables are shown in Table 3. These variables are computed calculating unweighted counts, weighted counts, and weighted proportions or means depending on whether the selected variable is discrete or continuous. The user may ask for other metrics that s/he deems suitable for analysis. Calculations are performed as outlined in the document "PSU and Cluster Weights User Guide."

Table 3: Variables includes in Data_B.xls

Variable	**Description**
L2_PCLUSTER_FALSE_ID_2	False cluster id for provider clusters
UWT_N_CLUSTER	Unweighted sample size of centers per provider cluster
WGT_N_CLUSTER	Weighted sample size of centers per provider cluster

UWT_N_0TO3YRS	Unweighted number of centers serving children 0 to 3 years old per provider cluster (based on CB_SERVE_0TO3YRS)
WGT_N_0TO3YRS	Weighted number of centers serving children 0 to 3 years old per provider cluster (based on CB_SERVE_0TO3YRS)
WGT_PROP_0TO3YRS	Weighted proportion of centers serving children 0 to 3 years old per provider cluster (based on CB_SERVE_0TO3YRS)
UWT_N_STDRATE_HRLY_2YR	Unweighted number of cases used to calculate the mean of prices charged by centers for care of 2 year old children (based on CB_B1_5_STDRATE_HRLY_2YR_TC)
WGT_MEAN_STDRATE_HRLY_2YR	Weighted mean of prices centers charge for care of 2 year old children (based on CB_B1_5_STDRATE_HRLY_2YR_TC)

Dataset named "**Data_C.xls**" combines the file Data_A.xls (center-level information) with Data_B.xls (provider cluster level information) using the cluster identifier (L2_PCLUSTER_FALSE_ID_2). The file has 1,221 rows, with one record for each of the 1,220 centers within 50 provider clusters. While cluster level metrics were computed through SAS programming, we provide a few manual illustrations to show how the numbers were estimated.

Figure 1. Snippet of Data_C.xls (subset of variables included in data file)

	A	B	C	D	E	G	H	I	J	K
1	L2_PCLUSTER_FALSE_ID_2	CB_METH_FALSE_CASEID	L2_STRATA_F	L2_CB_WEIGHT_PCLUSTER_FALSE	CB_SERVE_0TO3YRS_F	UWT_N_CLUSTER	WGT_N_CLUSTER	UWT_N_0TO3YRS	WGT_N_0TO3YRS	WGT_PROP_0TO3YRS
2	777001	22003238	3	2	0	1	2	0	0	0
3	777002	22000456	1	1.989	0	13	37.40644	4	14.56	0.38925
4	777002	22000501	1	3.978	1	13	37.40644	4	14.56	0.38925
5	777002	22000641	1	1.989	0	13	37.40644	4	14.56	0.38925
6	777002	22001990	1	1.989	1	13	37.40644	4	14.56	0.38925
7	777002	22002352	1	3.978	0	13	37.40644	4	14.56	0.38925
8	777002	22002778	1	1.989	0	13	37.40644	4	14.56	0.38925
9	777002	22005842	1	1.989	0	13	37.40644	4	14.56	0.38925
10	777002	22007151	1	3.978	1	13	37.40644	4	14.56	0.38925
11	777002	22008200	1	1.989	0	13	37.40644	4	14.56	0.38925
12	777002	22001472	2	2	0	13	37.40644	4	14.56	0.38925
13	777002	22000173	3	4.6154	0	13	37.40644	4	14.56	0.38925
14	777002	22000213	3	2.3077	0	13	37.40644	4	14.56	0.38925
15	777002	22004381	3	4.6154	1	13	37.40644	4	14.56	0.38925

Figure 1 shows that for the first provider cluster (777001) there is just one center (22003238, UWT_N_CLUSTER =1), but it represents 2 centers according to the weights and with a value of CB_SERVE_0TO3YRS_F=0 (i.e., center does not serve Age Category 0-3 years old). Consequently, the unweighted number of cases that are of interest is 0 (UWT_N_0TO3YRS) and the weighted number is also zero (WGT_N_0TO3YRS), similarly to the weighted proportion of centers serving children 0 to 3 years old (i.e., WGT_PROP_0TO3YRS).

Figure 1 also displays the second cluster (777002). This center has 13 unique centers (UWT_N_CLUSTER=13) and 4 of them serve children 0-3 years old (UWT_N_0TO3YRS=4). For that cluster, the weighted proportion of centers serving children 0 to 3 years old is 0.38925. Figure 2 illustrates how this number (WGT_PROP_0TO3YRS=0.38925) is calculated. One could visualize the approximate calculation as shown in the last column of Figure 2 below.

Figure 2. Calculations for Provider cluster: 777002(Proportion of centers serving children 0 to 3 years old)

Center	Weight	Indicator for Centers serving children 0 to 3 years old	Calculations
22000456	1.989	0	$= \frac{\text{Sum of weights for 4 centers serving children } 0-3 \text{ in provider cluster } 777002}{\text{Sum of weights for all 13 centers in provider cluster } 777002}$ $= \frac{(1.989(0)+3.978(1)+\ldots+2.3077(0)+4.6154(1))}{1.989+\ldots+4.6154}$ $= \frac{14.5604}{37.4065} = 0.38925$
22000501	3.978	1	
22000641	1.989	0	
22001990	1.989	1	
22002352	3.978	0	
22002778	1.989	0	
22005842	1.989	0	
22007151	3.978	1	
22008200	1.989	0	
22001472	2	0	
22000173	4.6154	0	
22000213	2.3077	0	
22004381	4.6154	1	

Figure 1 also displays a snippet for the same cluster 777002, illustrating results for the weighted mean of prices centers charge for care of 2 year old children. In this instance, the weighted average price is 0.9822. This number was calculated based on prices from three centers with valid information out of thirteen centers in the cluster (22000501, 22001990, and 22004381). Note that this average includes rates of $0 and that this cluster had two centers with a $0 rate. That is,

$$\frac{(3.978 * 2.613) + (1.989 * 0) + (4.6154 * 0)}{3.978 + 1.989 + 4.6154} = \frac{10.39451}{10.5824} = 0.982$$

Exercises for users:

2.1. Think about what other estimates you could have calculated at the cluster level based on the two variables available at the center-level.

2.2. Try to calculate these alternative cluster-level estimates based on the data provided.

Answer Key:

2.1. Some examples are (i) whether there is at least one center serving children 0-3 in the cluster (1=at least one center, 0=otherwise) and (ii) proportion of centers that have a price for 2 years old above a certain threshold (let's say, the national or state average for 2 year olds).

2.2. You would need to use file "Data_A.xls" and aggregate data by L2_PCLUSTER_FALSE_ID_2. The weights should be used to generate weighted metrics.

Step 3: User inspects provider cluster-level summary statistics

In this step the user produces a set of summary statistics for metrics generated at the cluster level. They are reported in file named "**Data_D.xls**." These statistics come from the 50 provider clusters we are using in this exercise (i.e., Data_B.xls). Data_D.xls shows basic measures of central tendency and dispersion for variables previously described in Table 3. To facilitate interpretation, each measure under the column "Statistic" in Data_D.xls is identified with a prefix; namely, "COUNT_" to indicate sample size, "MEAN_" to indicate the mean value of the variable, "MAX_" the largest value in the dataset, "MIN_" the smallest value in the dataset, "Q1_" the lower quartile, "Q3_" the highest quartile and "MED_" to indicate the median. The dataset (Data_D.xls) includes both a column for results based on simulated provider cluster IDs (column RESULT_SAMPLE_FALSE), and information based on real provider cluster IDs (column RESULT_SAMPLE).

This tutorial includes summary statistics based on simulated and real provider cluster IDs because during the actual process of requesting cluster-level metrics (to be appended later on to the Household dataset), the requestor is expected to generate metrics based on simulated IDs and NORC will reproduce this summary statistic based on real IDs. With simulated IDs, the user will be able to produce summary statistics (similar to those in Data_D.xls --RESULT_SAMPLE_FALSE) and have a realistic view of measurement properties of cluster-level metrics when compared to real IDs. In other words, the comparison illustrated here with simulated and real IDs helps to illustrate how different results would be relative to results based on real IDs.

Summary statistics (Data_D.xls) based on provider cluster do not have any analytic value; they are intended only to show the user how robust the cluster-level calculations would be. For instance, the

variable called "MEAN_UWT_N_0TO3YRS" represents the average number of unweighted centers per cluster calculated on the basis of "UWT_N_0TO3YRS," which in turn was created for the variable CB_SERVE_0TO3YRS_F (centers serving children 0 to 3 years old per provider cluster). That variable (i.e., MEAN_UWT_N_0TO3YRS) shows that on average there are 13.92 centers serving children 0-3 in each cluster per cluster and this average is the same for the sample with simulated and real IDs.

Please note that the analysis of cluster-level data is intended in combination with household level data, as detailed in the following step. Cluster-level variables are only intended to be used combined with the household survey for an analysis at the household or child level.

Exercises for users:

3.1. What is the weighted median and mean price charged for 2-year olds in the sample of provider clusters with Simulated Provider Cluster IDs?

3.2. How does this price vary across provider clusters in the sample with Simulated Cluster IDs?

Answer Key:

3.1. Mean: MEAN_WGT_MEAN_STDRATE_HRLY_2YR = $7.156

Median: MED_WGT_MEAN_STDRATE_HRLY_2YR = $ 7.075

3.2. 25th percentile: Q1_WGT_MEAN_STDRATE_HRLY_2YR= $4.179

75th percentile: Q3_WGT_MEAN_STDRATE_HRLY_2YR= $10.039

Step 4: NORC merges cluster aggregates onto the Household dataset

To provide a sense of the number of household units associated with cluster-level metrics, we present "**Data_E.xls.**" This dataset is an updated version of Data_B.xls with one record per provider cluster. This file has one additional column that shows the unweighted number of households in each cluster (HH_COUNT). For instance, we see that for cluster 777001 (with just one center) there are 21 households, whereas cluster 777002 (whose composition includes 13 centers) is associated with 24 households. Variable HH_COUNT indicates that overall, the 50 provider clusters included in these tutorial working files may be matched to 754 sampled households (unweighted count of households). These counts include households of any type in the NSECE Household Survey. If a user were to restrict the analysis to households with a child under 5 or low-income households, the count would be different, lower than what is reported here. This data file aims to help users understand the linkage between

cluster aggregates and the household file. Level 2-restricted use teams will not have access to a file like this. Instead, they will only access a household-level file, described in the next section.

Metrics calculated at the cluster-level (Data_B.xls) can now be merged into the household dataset using the cluster identifier. To demonstrate how cluster level metrics will look once they have been merged with HH data, we use a sample of 300 households with simulated cluster data from the HH dataset (Data_F.xls). Please note that this data file corresponds to a subset of the households matched to the provider clusters. As mentioned before, the 50 provider clusters in Data_E.xls could be matched to 754 households. Only 300 of those households are included in Data_F.xls. Data_F also includes several characteristics of the household such as its poverty level and car ownership status (see Table 4). The file also includes characteristics of where the household is located, such as its region and HH_COM_POV_DENS, the concentration of nearby population living in poverty as documented in the American Community Survey data. These variables will allow us to link supply with demand characteristics, which are relevant for our analyses.

Table 4: Variables included in Data_F.xls	
Variable	Description
HH_COM_POV_DENS	The community poverty density variable identifies whether the community where the household is located has a high, medium or low concentration of population living in poverty (i.e. at or below the Federal Poverty Line)
HH_REGION	Region of the country where child's household is located: Northeast; Midwest; South; West.
HH_ECON_PARWORK	Count of the number of parents who live in the HH *and* report working last week at HH_D1A_WORK_X
HH_ECON_INCOME_POVRATCAT	Ratio of annual income for the calendar year 2011 to poverty level.
HH_ECON_OWNCAR	Car ownership
HH_HHCOMP_ANY_CHILD_0THRU5YRS	HH has any children 0-5 years

The file Data_F.xls has 301 rows, with one record for each of the 300 households included in this tutorial sample. The file shows how the 11 households associated with cluster 777001 are displayed at the household level (you may recall this cluster could be matched to a total of 21 households, but only 11 of those 21 households are included in this file). These 11 households have no values for the weighted mean of CB_B1_5_STDRATE_HRLY_2YR_TC (as shown in the missing value in column WGT_MEAN_STDRATE_HRLY_2YR). As mentioned above, all households in this provider cluster are associated with a single center that does not serve children ages 0-3 years old. Therefore, in this cluster, the weighted proportion of centers serving children 0-3 is zero. Note that this file shows that all households in this cluster have values of WGT_PROP_0TO3YRS= 0.

A dataset with a structure similar to Data_F.xls (whose unit of analysis is the household, using HH weights), should be the basis for data users interested in conducting supply and demand analyses. We could, for example, analyze the relationship if the household income level and the prices charged by centers, or the relationship between the concentration of people living in poverty and the level in which 0 to 3-year olds are served. When Level-2 research teams work with provider cluster data, they will access a file like this, that is, a file in which cluster provider aggregates are merged as attributes of the household or children in the household. An important difference between Data_F and what Level-2 research teams actually have access to has to do with Provider Cluster IDs. The household or child-level file that Level-2 research teams access will not include any Provider Cluster IDs. Part of the disclosure protection NORC has established is that Level-2 data files do not provide identifiers that would allow researchers to know which households share a provider cluster.

Exercises for users:

4.1. How does the availability of centers serving children 0-3 years old vary by household income?

4.2. Do the prices charged for 2-year olds vary by community poverty density?

Answer Key:

4.1. Calculate average WGT_PROP_0TO3YRS by HH_ECON_INCOME_POVRATCAT, using household-level weight HH_METH_FALSE_WEIGHT

Poverty	Average
1 (<100% FPL)	57%
2 (100 to 199% FPL)	50%
3 (200 to 299% FPL)	54%
4 (>=300% FPL)	54%
Total	54%

4.2. Calculate average WGT_MEAN_STDRATE_HRLY_2YR by HH_COM_POV_DENS, using household-level weight HH_METH_FALSE_WEIGHT

Poverty	Average
1 (Low)	6.94
2 (Moderate)	6.36
3 (High)	7.39
Total	7.08