Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society

A Model-based Disaggregation Method for Estimation of Adult Competency

by Andreea L. Erciulescu, Weijia Ren, Jianzhu Li, Tom Krenzke, Leyla Mohadjer and Robert Fay

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2022
Data Disaggregation: Building a more representative data portrait of society

A Model-based Disaggregation Method for Estimation of Adult Competency

Andreea L. Erciulescu, Weijia Ren, Jianzhu Li, Tom Krenzke, Leyla Mohadjer, Robert Fay[1]

Abstract

Estimation at fine levels of aggregation is necessary to better describe society. Small area estimation model-based approaches that combine sparse survey data with rich data from auxiliary sources have been proven useful to improve the reliability of estimates for small domains. Considered here is a scenario where small area model-based estimates, produced at a given aggregation level, needed to be disaggregated to better describe the social structure at finer levels. For this scenario, an allocation method was developed to implement the disaggregation, overcoming challenges associated with data availability and model development at such fine levels. The method is applied to adult literacy and numeracy estimation at the county-by-group-level, using data from the U.S. Program for the International Assessment of Adult Competencies. In this application the groups are defined in terms of age or education, but the method could be applied to estimation of other equity-deserving groups.

Key Words: Allocation; Adult Competency; Official Statistics; Program for the International Assessment of Adult Competencies, Small Area Estimation.

1. Introduction

Making effective evidence-based policies and laws relating to adult education requires sound research based on reliable data that are most relevant to jurisdictions such as counties, states, and demographic groups within counties and states. As a multicycle international study involving over 30 countries under the leadership of the Organization for Economic Cooperation and Development, the first cycle of the Program for the International Assessment of Adult Competencies (PIAAC) was designed to provide national estimates of the proficiency of adult literacy, numeracy and problem-solving skills. The PIAAC survey provides high-quality national estimates through a multistage probability design with in-person data collections that include a screener questionnaire, a background questionnaire, and an assessment of adult skills. In the United States, PIAAC is sponsored by the National Center for Education Statistics at the Institute of Education Sciences. From 2012 to 2017, some 12,330 U.S. adults ages 16 to 74 living in households were surveyed for U.S. PIAAC. In the rest of the paper, we will refer to the U.S. PIAAC as PIAAC.

Because the PIAAC sample size was too small to support the production of estimates at disaggregate levels, Krenzke et al. (2020) and Li et al. (2022) developed hierarchical Bayes model-based small area estimation (SAE) methodology to produce county, state, and state by age and education groups estimates for average scores for literacy and numeracy, and various proficiency levels. By using PIAAC survey data in conjunction with data from the American Community Survey (ACS), the SAE estimates provide reliable U.S. official statistics of adult literacy and numeracy skills in all 50 states, all 3,141 counties, and the District of Columbia, and in all 50 states and the District of Columbia by six age groups, 16-24, 25-34, 35-44, 45-54, 55-64, and 65-74 year old, and four education groups, less than high school, high school diploma or general education diploma (GED), some college (no degree or attained associate's degree), and bachelor's degree or higher. These estimates are available in the Skills Map at https://nces.ed.gov/surveys/piaac/skillsmap/ for all outcomes of interest, the proportions at or below Level 1 (P1), at Level 2 (P2), and at or above Level 3 (P3), and averages.

[1]Andreea L. Erciulescu, Westat, 1600 Research Blvd, Rockville, MD, USA, 20850 (AndreeaErciulescu@westat.com); Weijia Ren; Meta, USA; Jianzhu Li, FINRA, 1735 K Street, NW, Washington DC, USA, 20006; Tom Krenzke, Westat, 1600 Research Blvd, Rockville, MD, USA, 20850; Leyla Mohadjer, Westat, 1600 Research Blvd, Rockville, MD, USA, 20850; Robert Fay, Westat, 1600 Research Blvd, Rockville, MD, USA, 20850. The work of Dr. Ren and Dr. Li was conducted while they were with Westat.

This paper addresses county by group estimation of adult proficiency, where the quantities of interest are the same as the ones considered in Krenzke et al. 2020 and Li et al. (2022), and the groups of interest are the same ones considered in Li et al. (2022). A deterministic method is used to allocate the state by group estimates to the county by group level, for all the counties nested within the corresponding state, using the ratios of county to state estimates. The allocation is applied at the posterior sample level, resulting in pseudo-posterior distributions for the quantities of interest.

The rest of the paper is organized as follows. In Section 2, we describe the data available from PIAAC, the Skills Map, and the ACS, serving as input into the county by group-level estimation process. The allocation method is presented in Section 3 and validation metrics are provided in Section 4. A summary is given in Section 5. Unless otherwise noted, selected results are presented throughout the manuscript only for one quantity of interest, the proportion at or below Level 1 literacy, and for one demographic group, the population with less-than-high-school education. Results for all quantities of interest and for all domains of interest are comparable and are available in the Skills Map and in the report Erciulescu et al. (2022).

2. Data available at various levels of aggregation

Person-level survey data are available from PIAAC and are used to produce survey estimates for the county by group domains with sample data. Because these data are sparse at disaggregated levels such as county by group, an indirect estimation approach is developed. The indirect estimation approach uses data produced as the output of SAE models for higher levels of aggregation, i.e., county, state, and state by group. The survey estimates and selected auxiliary data are used in the validation process of the indirect county by group estimates. Details about the PIAAC survey, model, and auxiliary data are provided in the rest of this section.

2.1 Survey estimates

Following Li et al. (2022), the multiple imputation approach is implemented for constructing the survey estimates and the associated variance estimates. The PIAAC micro-level data includes ten plausible values generated from a posterior distribution by combining the item response theory scaling of the cognitive items with a latent regression model using information from the background questionnaire in a population model. For each plausible value, Háyek-type estimates and associated Taylor-series approximated variance estimates are constructed for the county by group-level quantities of interest, with survey weights calibrated to 2013-2017 ACS control totals for age groups, education levels, gender, and race/ethnicity within state. Then, the estimates for the ten plausible values are averaged to produce the county by group-level survey estimates. The variances of the county by group-level estimates are estimated using the multiple imputation approach by combining the within-imputation and between-imputation variances.

The PIAAC data includes 185 counties with at least one respondent in each county, but some counties have no respondents for one or more age/education groups. The number of county by group domains with sample data and the distributions of county by group domain sample sizes are provided in Table 2.1-1. With a small number of county-level survey estimates available for the groups, an indirect estimation approach is deemed necessary to produce estimates for all the county by age/education group domains. The medians of the county by group domain sample sizes range from 6.5 to 16, indicating that the survey estimates are subject to great uncertainty in most county by group domains. In addition to producing estimates for all the domains of interest, the indirect estimation approach described in the next section helps reduce the uncertainty of the survey estimates.

2.2 Model-based estimates

Posterior distributions for the adult proficiency quantities of interest at the county, state, and state by group levels were produced in Krenzke et al. (2020) and Li et al. (2022). Although only point estimates with associated variance estimates and credible intervals are directly available to download from the Skills Map, 4,500 posterior samples for the adult proficiency quantities of interest are also available upon request. We work with these posterior samples to construct the county by group-level estimates.

Table 2.1-1
County-level sample size distributions for age and education groups: 2012/2014/2017 PIAAC

Age and education groups	Number of counties with sample	Number of respondents					
		Minimum	10th percentile	Median	90th percentile	Maximum	Mean
16-24	172	1	2	10	31	82	15
25-34	177	1	3	11	33	96	16
35-44	180	1	2	9.5	21.5	56	11
45-54	178	1	2	10	21	72	11
55-64	180	1	2	10	18	45	10
65-74	176	1	1	6.5	14	31	7
Less than high school	175	1	2	8	25	67	12
High school diploma or GED	180	1	4	16	42.5	86	20
Some college	182	1	3	16	38	114	19
Bachelor's degree or higher	179	1	3	13	36	115	18

Let $Y_j^{(b)}, Y_{jk}^{(b)}$, and $Y_{g,j}^{(b)}$ denote the state-level, county-level, and state by group-level posterior samples, respectively. The variable Y denotes any of the quantities of interest, P1, P2, P3, or average, j is an index for the states, k is an index for the counties, g is an index for the groups, and b is an index for the posterior samples. These samples are available for all the counties and states, and for all the age/education groups of interest. The quality of the county by group-level estimates is dictated by the quality of the county-level, state-level, and state by group-level estimates, since the latter serve as the only inputs into the model used to produce the former, to be described in the next section.

2.3 Auxiliary data

Two ACS variables related to adult proficiency are now selected in the validation of the county by group-level estimates. They are available from the initial pool of variables considered by Krenzke et al. (2020) for the SAE models. The first variable is the proportion of population below poverty and the second variable is the proportion of population employed. These variables are available for all the county by group domains of interest.

3. Allocation method

The county-level estimates for groups are constructed using an allocation method. This method consists of using the county-level and state by group-level model-based estimates and allocating them to the county by group domains of interest via a deterministic model. Like the SAE modeling process developed for county, state, and state by age/education group estimation, model-based estimates for P1 and P3 are constructed first, and then their sum is subtracted from one to obtain the model-based estimates for P2. Also, literacy and numeracy measures are estimated independently. The model output from the SAE models developed for the estimation of county, state, and state by age/education group proficiency measures serve as input into the county by group estimation process.

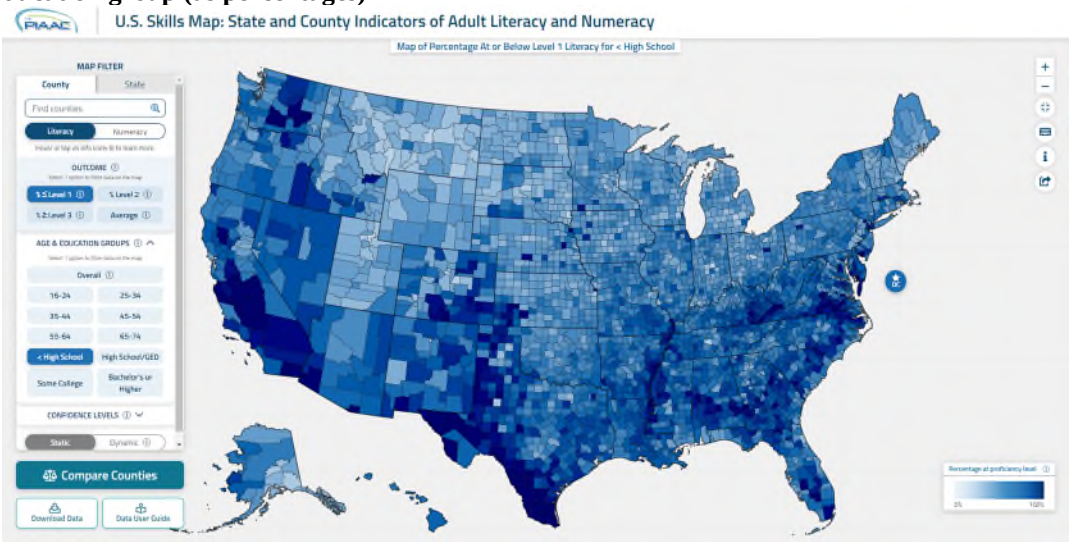
Following the notation from the previous section, let the pseudo-posterior samples for the county by group-level quantities of interest be defined as

$$Y_{g,jk}^{(b)} := Y_{g,j}^{(b)} \frac{Y_{jk}^{(b)}}{Y_j^{(b)}}, \tag{1}$$

with a note that Y now denotes either P1, P3, or the average. The implicit working assumption is that the county-to-state ratio of posterior samples is constant across the age and education groups. Pseudo-posterior samples for the county by group-level P2 are then defined as $1 - P1_{g,jk}^{(b)} - P3_{g,jk}^{(b)}$. Posterior summaries, such as means, variances, and credible intervals are constructed using the pseudo-posterior distributions described above. Posterior means for P1, P2, and P3 that are below 0 or above 1 are set equal to 0 or 1, respectively.

Among the 175 counties with sample data for the less-than high-school education group, there were 3 P1 model-based literacy estimates greater than 1. Among the 2,967 counties without sample data for the less-than-high-school education group, there were 42 P1 model-based literacy estimates greater than 1. None of the county-level P1 model-based estimates for the less-than-high-school education group were below 0. No adjustment was needed for the average score model-based estimates because all were positive. The P1 literacy model-based county by group-level estimates for the population with less than high school education are illustrated in the map in Figure 3-1.

Figure 3-1
Map of county-level model-based estimates for the proportion at or below Level 1 literacy for less-than-high-school education group (as percentages)



4. Validation of final estimates

The survey estimates constructed for the county by age/education groups serve as input into the validation process, along with the selected ACS estimates related to proficiency described in Section 2. Three validation metrics are considered: 1) visual displays of the magnitude and direction from survey estimates to model-based estimates, 2) visual displays of the relationship between the proficiency estimates and related ACS variables, and 3) uncertainty measures for the model-based estimates.

The shrinkage plots in Figure 4-1 show the magnitude and direction from survey estimates to model-based estimates of literacy proportions for the less-than-high-school education group, by sample size. Short arrows correspond to small differences between the survey and the model-based estimates and long arrows correspond to large differences between the survey and the model-based estimates. The shrinkage is more substantial in domains with smaller sample sizes than those in domains with larger sample sizes. Arrows pointing upward correspond to negative differences between the survey and the model-based estimates and arrows pointing downward correspond to positive differences between the survey and the model-based estimates.

Scatterplots of estimates of literacy P1 for the less-than-high-school education group are illustrated in Figure 4-2, against the selected ACS variables (proportion of population below poverty, proportion of population employed). In general, similar range of the survey estimates and model-based estimates, and similar relationship between the ACS variables and the estimates are observed, with this relationship being clearer between the model-based estimates and the ACS variables than between the survey estimates and the ACS variables. This result is expected because the county by group-level model-based estimates are functions of the county, state, and state by group-level model-based estimates, which are themselves functions of, or related to, the ACS variables considered here.

Figure 4-1
Literacy proportion (less than high school) - Shrinkage plots of point estimates by sample size

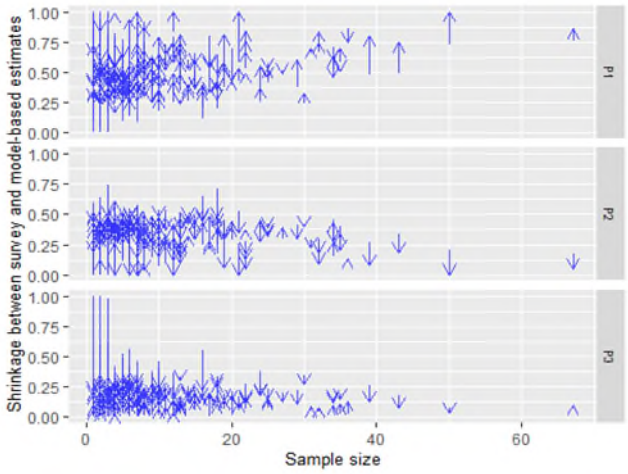
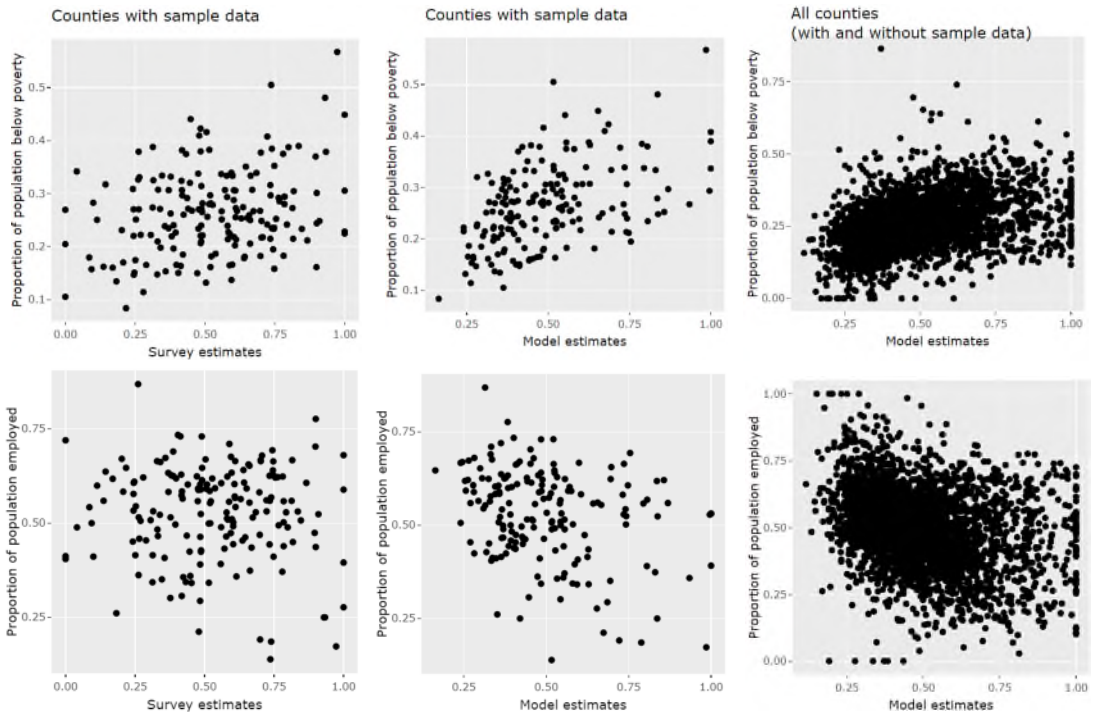


Figure 4-2
Literacy proportion at or below Level 1 for less-than-high-school education group versus selected ACS variables - County by group-level estimates: 2012/2014/2017 PIAAC



As reported in Table 4-1, the median credible interval width is 24.5 percent for county-level estimates of literacy P1 for the less-than-high-school education group, where the median is taken over all the county by group domains in the U.S. When these county by group domains are categorized by sample availability, the median credible interval widths are 22.7 percent and 24.7 percent for in-sample and not-in-sample domains, respectively.

The coefficients of variation (CVs) for the county-level model-based estimates of literacy P1 for the less-than-high-school education group are also summarized in Table 4-1 by sample availability. For most of these county by group domains, the CVs are lower than 20 percent. All CVs reported in Table 4-1 for in-sample domains are less than half

of the corresponding CVs for the survey estimates reported in Table 2-2 in Erciulescu et al. (2022). This result is expected because the uncertainty in the model-based county-level, state-level, and state by group-level estimates used as inputs into the allocation method was lower than the uncertainty in the corresponding survey estimates (see Krenzke et al. 2020 and Li et al. 2022). Also, the CVs reported in Table 4-1 are larger than the CVs for the state-level model-based estimates of literacy P1 for the less-than-high-school education group reported in Table 3-5 in Li et al. (2022): the state-level estimates by group have a median CV of 11.8 percent compared to 12.9 percent for the county-level estimates by group. This result is expected because the level of aggregation in this paper is finer than the level of aggregation in the cited report (county by group versus state by group).

Table 4-1
Distribution of credible interval widths and coefficients of variation for county-level model-based estimates for less than high school for literacy proportion at or below Level 1: 2012/2014/2017 PIAAC

	Percentile				
Statistics for less than high school	20	40	50 (Median)	60	80
County estimates for all domains					
95 percent credible interval width (percent)	21.1	23.4	24.5	25.9	29.4
Coefficient of variation (percent)	10.7	12.1	12.9	14.0	17.7
County estimates for in-sample domains					
95 percent credible interval width (percent)	19.6	21.5	22.7	24.0	27.9
Coefficient of variation (percent)	10.3	11.6	12.3	12.8	16.3
County estimates for not-in-sample domains					
95 percent credible interval width (percent)	21.2	23.5	24.7	25.9	29.5
Coefficient of variation (percent)	10.7	12.1	13.0	14.1	17.7

5. Summary

An allocation approach was developed to produce county-level model-based estimates for six age groups and four educational attainment groups. Previous model-based SAE estimates, constructed at the county, state, and state by group levels, served as inputs into the allocation approach. The resulting estimates are available in the Skills Map.

References

Krenzke, T., Mohadjer, L., Li, J., Erciulescu, A., Fay, R., Ren, W., Van de Kerckhove, W., Li, L., and Rao, J.N.K. (2020). "Program for the International Assessment of Adult Competencies (PIAAC): State and County Estimation Methodology Report (NCES 2020-225)." U.S. Department of Education, National Center for Education Statistics, Washington, DC: U.S. Government Printing Office. Retrieved [10-26-2020] from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020225.

Li, J., Krenzke, T., Ren, W., Mohadjer, L., Fay, R., and Erciulescu, A. (2022). "Program for the International Assessment of Adult Competencies (PIAAC): State-level Estimation for Age and Education Groups Methodology Report (NCES 2022-050)." U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved [9-2-2022] from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2022050.

Erciulescu, A., Ren, W., Li, J., Krenzke, T., Mohadjer, L., and Fay, R. (2022). "Program for the International Assessment of Adult Competencies (PIAAC): County-Level Estimation for Age and Education Groups Methodology Report (NCES 2023-004)." U.S. Department of Education, National Center for Education Statistics, Washington, DC: U.S. Government Publishing Office. Retrieved [10-31-2022] from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2023004.