Replication Variance Estimation for Balanced Sampling: An Application to the PIAAC Study

Jianzhu Li [1], Sixia Chen[1], Tom Krenzke[1]
and Leyla Mohadjer[1]
1600 Research Blvd, Rockville, MD 20850

Abstract

This paper describes the application of a replication method for variance estimation to a sample drawn from a balanced sampling design in the Programme for the International Assessment of Adult Competencies (PIAAC) Round 1 study. One of the participating countries, France, used the Cube method to select a balanced sample of the first stage sampling units, Interviewer Action Areas (IAAs), within administrative regions. To be consistent with other participating countries, the variance estimation should use a replication method with 80 replicate weights. Typical methods for constructing replicate weights are not valid for the balanced design. Within each stratum we applied a method outlined by Fay (1984) to generate replicate weights such that the replication variance estimator produces algebraically equivalent results to the approximate estimator proposed by Deville and Tillé (2005). Across strata, the replicate weights were allocated in a block-diagonal fashion to reduce the correlation.

Key Words: Spectral decomposition; jackknife variance estimation; balancing conditions

1. Introduction

In many large-scale complex surveys, replicate weights are generated and included in the data files to provide valid variance estimates. The concept of the replication method is to draw replicate samples from the original sample following a specific resampling scheme. The variability of the estimates among the replicate samples is then used as a replication-based estimator of variance. The typical replication methods include bootstrap, jackknife, and balanced repeated replication. These methods became very popular for variance estimation since it can be relatively easily applied for a variety of survey estimators regardless of their complexities. Relevant modules or functions have been developed in many widely-used statistical software to implement such methods. The modern computational capacity makes the implementation feasible and practical even for surveys that use a large number of replicates. The validity of variance estimates computed from most replication methods is conditional on sample design features. The typical replication methods may not be applicable for certain types of sample designs. In this paper, we presented a case study in which the replicate weights were generated for survey data using the balanced sampling approach in its first stage of sample design.

In Section 2, we briefly review the balanced sampling technique and the proposed variance estimator. In Section 3, we discuss a general procedure, outlined by Fay (1984) and Fay and Dippo (1989), to construct replicate weights for any given sample design. In

Section 4, we report the creation of replicate weights for one of the participating countries in the Programme for the International Assessment of Adult Competencies (PIAAC) study, followed by a summary in Section 5.

2. Balanced Sampling and Variance Approximation

Deville and Tillé (2004) developed a general method, the Cube method, which allows the selection of a multivariate balanced sample on a given set of auxiliary variables. In other words, the Horvitz–Thompson estimates of some auxiliary (x) variables are exactly or nearly equal to their population totals X:

$$\hat{X} = \sum_{k \in S} w_k x_k = \sum_{k=1}^{N} x_k = X,$$

where $w_k = 1/\pi_k$ is the design weight or the inverse of the selection probability for case k.

The balanced sampling approach draws representative samples randomly, not in a purposive fashion. It integrates calibration into the sample design stage to achieve high efficiency leading to variance reduction. The method allows unequal inclusion probabilities and a large set of categorical or continuous balancing variables. The Cube method is carried out in two phases. In the flight phase it attempts to select a random sample that satisfies the balancing constraints exactly. If no sample is obtained in the flight phase, a sample is selected to respect the constraints as closely as possible in the landing phase.

For variance estimation, Deville and Tillé (2005) proposed to approximate the variance of the weighted total of a variable Y ($\hat{Y}$) under the balanced sampling approach by a Generalized Regression estimator (GREG) under conditional Poisson sampling. That is,

$$Var_b(\hat{Y}) = Var_p(\hat{Y}|\hat{X} = X). \quad (1)$$

where b denotes balanced sampling and p denotes Poisson sampling.

Using the same approximation, Breidt and Chauvet (2011) derived the variance estimator for the weighted total $\hat{Y}$ as

$$\widehat{var}_b(\hat{Y}) = \frac{n}{n-q} \sum_s \pi_k (1 - \pi_k) \left(\frac{y_k}{\pi_k} - \frac{\tilde{y}_k}{\pi_k} \right)^2. \quad (2)$$

In the formula above, n is the number of cases in the sample, q is the number of balancing conditions, and the subscript s refers to the sample of size n. The term $\tilde{y}_k$ can be expressed as $\tilde{y}_k = x_k' \hat{\beta}$, where

$$\hat{\beta} = x_k' \left(\sum_s (1 - \pi_k) \frac{x_k x_k'}{\pi_k^2} \right)^{-1} \sum_s (1 - \pi_k) \frac{x_k y_k'}{\pi_k^2}. \quad (3)$$

is derived from a linear regression model.

If we denote $\boldsymbol{z_s} = \left(\frac{y_1}{\pi_1}, \frac{y_2}{\pi_2}, \ldots \frac{y_n}{\pi_n} \right)$, formula (2) can be re-written in a quadratic form as

$$\widehat{var}_b(\hat{Y}_\pi) = \mathbf{z}_s' \mathbf{D}_s \boldsymbol{z}_s. \quad (4)$$

In formula (4), $\mathbf{D_s}$ is an *n*x*n* symmetric matrix with each element D_{kl} (*k*=1,…,*n*; *l*=1,…,*n*) defined as

$$\mathrm{D_{kl}} = \begin{cases} \mathrm{c_k - c_k a_k'(\sum_{i\in S} c_i a_i a_i')^{-1} a_k c_k} & \mathrm{k = l} \\ \mathrm{-c_k a_k'(\sum_{i\in S} c_i a_i a_i')^{-1} a_l c_l} & \mathrm{k \neq l'} \end{cases}$$

where $\mathbf{a_k} = \frac{x_k}{\pi_k}$ and c_k can be approximated by $(1-\pi_k)\frac{n}{n-q}$. It should be noted that matrix $\mathbf{D_s}$ is only dependent on the sample units, the auxiliary variables, and the inclusion probabilities, but not the actual *y* values.

3. Replicate Weight for Balanced Sampling

Kim and Wu (2013) used Fay's Method (Fay 1989, Fay and Dippo, 1989) to derive replicate weights based on formula (2). Fay's method is a methodology for generating a replication variance estimator which replicates a quadratic form variance estimator. Any symmetric positive semi-definite matrix can be decomposed in terms of its eigenvalues and eigenvectors. This is called spectral decomposition. Therefore, matrix $\mathbf{D_s}$ can be expressed as

$$\mathbf{D_s} = \mathbf{EQE'}$$

where **Q** is an *n*x*n* diagonal matrix with *n* eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ on the diagonal and 0s off the diagonal, and **E** is an *n*x*n* matrix with *n* eigenvectors $\nu_1, \nu_2, \ldots, \nu_n$ as columns of the matrix. With this matrix decomposition, the variance estimator (4) reduces to the following:

$$\widehat{var}_b(\hat{Y}) = \sum_{i=1}^{n} \lambda_i \, (\mathbf{v_i' z_s})^2.$$

A typical jackknife variance estimator for $\hat{Y}$ takes the form

$$\widehat{var}_r(\hat{Y}) = c \sum_{r=1}^{R} (\hat{Y}(r) - \hat{Y})^2,$$

where $\hat{Y}(r)$ is the replicate estimate using the *r*th replicate weights. The constant *c* equals 1 for the paired jackknife estimator (Rust and Rao, 1996).

Under Fay's Method, we define the r^{th} replicate weights $\boldsymbol{w_r}$ by linking it to the *r*th eigenvalue λ_r and the corresponding eigenvector $\boldsymbol{\nu_r}$ as follows:

$$\boldsymbol{w_r} = \begin{bmatrix} (1+\sqrt{\lambda_r}\,\nu_{r1})w_1 \\ (1+\sqrt{\lambda_r}\,\nu_{r2})w_2 \\ \vdots \\ (1+\sqrt{\lambda_r}\,\nu_{rn})w_n \end{bmatrix},$$

Then, the difference between the r^{th} replicate weights $\boldsymbol{w_r}$ and the full sample weight $\boldsymbol{w}$ is

$$\boldsymbol{w_r} - \boldsymbol{w} = \sqrt{\lambda_r} \begin{bmatrix} \nu_{r1} w_1 \\ \nu_{r2} w_2 \\ \vdots \\ \nu_{rn} w_n \end{bmatrix},$$

and the jackknife replicate variance estimator reduces to

$$\widehat{var}_r(\hat{Y}) = \sum_{r=1}^{n} (\{\boldsymbol{w_r} - \boldsymbol{w}\}' \mathbf{y_s})^2 = \sum_{r=1}^{n} \lambda_r (\boldsymbol{\nu_r}' \mathbf{z_s})^2 = \widehat{var}_b(\hat{Y}).$$

It indicates that using $\boldsymbol{w_r}$ in the paired jackknife formula would produce the same variance estimate as the approximate estimator. The number of replicates R is equal to the number of eigenvalues, or the number of cases, *n*, if matrix $\mathbf{D_s}$ is of full rank. The approach may generate enormous number of replicate weights for large scale surveys. Kim and Wu (2013) discussed reducing the number of replicates using a weight-calibration method.

3. Application to PIAAC Study

3.1 Overview of PIAAC Study

The Programme for the International Assessment of Adult Competencies (PIAAC) study is an international survey, sponsored by the Organisation for Economic Cooperation and Development (OECD) that assesses and compares adults' proficiency in literacy, numeracy, and problem solving around the world. It collects data from the non-institutionalized population between 16-65 years old. PIAAC is a complex study since the data collection is conducted in numerous countries with diverse populations, cultures, languages, education and life experiences. Twenty four countries participated in the first round of data collection in 2011-2012, and nine countries are taking part in the second round in 2014. The participating countries designed their sampling plans following the quality assurance standards and guidelines set by the OECD and the PIAAC Consortium[1]. The International Public Use Files (PUFs) for Round 1 of PIAAC were released in 2013[2].

The PIAAC standards and guidelines require the use of replication method for variance estimation, and set the maximum number of replicates to 80 for each country. This decision was made mainly for two reasons: First, for such a large scale international survey, the data users may be interested in different types of analyses, even some complicated nonlinear statistics, e.g., competency scores or plausible values. Not like the

[1] The design and implementation of PIAAC is the responsibility of an international consortium of well-established institutions from North America and Europe led by the Educational Testing Service in the United States. The other partners of this consortium are Westat in the United States; cApStAn in Belgium; the Research Centre for Education and the Labour Market (ROA) at the University of Maastricht in the Netherlands; and the GESIS - Leibniz Institute for the Social Sciences, the German Institute for International Education Research (DIPF), and the Data Processing Centre of the International Association for the Evaluation of Educational Achievement (IEA) in Germany.

[2] http://www.oecd.org/site/piaac/publicdataandanalysis.htm (accessed September 16, 2014)

traditional test scores, plausible values can be viewed as a set of quantities generated from multiple imputations based on responses to the test items and other background information (Mislevy, Johnson and Muraki, 1992). Replication methods are convenient to handle variance estimation for all the statistics including plausible value analysis. Second, an online platform, the International Data Explorer (IDE), was developed by OECD to allow the users to create statistical tables and charts online in real time. PIAAC PUF data can be analysed by country and by demographic characteristics, social and economic status, education level, employment status, etc. This system was designed in a way that requires all the countries use the same or very similar variance estimation techniques. Replicates were created following the standard approaches for all participating countries except France. France applied the balanced sampling technique in selecting their sample. Typical replication methods such as jackknife or BRR cannot be directly applied in this case. We implemented Fay's Method, as described in Section 2, to create the replicate weights for France.

3.2 Sample Design: France

France has adopted the Cube method to draw balanced samples in many of its national surveys to improve efficiency. In the PIAAC study, the sample was selected using a two-stage stratified design. In the first stage, the municipalities were aggregated to create the French master sample Interviewer Action Areas (IAAs), which served as the Primary Sampling Units (PSUs). The frame came from the census data file. The sample was drawn with probabilities proportionate-to-size, where the size measure was the number of residences in the PSUs. The certainty PSUs were first identified. Next, a balanced sample of non-certainty PSUs was selected using the Cube method. The balancing variables included the number of dwelling units in rural, semi-urban, and urban areas, respectively, as well as total income. Sampling was conducted independently in each of the 22 strata defined by geographical regions. A few regions had no dwelling units in rural areas and therefore used only three balancing conditions. In the second stage, the individual taxation files, updated annually, served as the frame to select persons from the certainty PSUs and the sampled non-certainty PSUs. The frame was stratified by types of housing and administrative districts. Persons were sampled using systematic random sampling to achieve a target of about 5,000 completed interviews.

3.3 Weighting Process for France

In this section we will discuss the weighting process using the French data and provide the details about creating replicate weights for the balanced sample of non-certainty PSUs. In total, there were 79 certainty and 488 non-certainty PSUs in the sample. Since there was no variance contribution at the PSU level from the certainty PSUs, the replicate weights were created directly at the person level. The number of sampled non-certainty PSUs ranged from 3 to 84 in 22 strata (regions). At the PSU level, we independently applied Fay's Method to generate the replicate base weights for non-certainty PSUs within strata. In each stratum, the number of replicate weights was determined by the number of non-zero eigenvalues, or the number of sampled PSUs minus the number of balancing conditions. The number of replicate weights for non-certainty PSUs by stratum is presented in the last column of Figure 1.

There was one small region (stratum 22) that contained only three PSUs. Deville and Tillé's variance estimator, i.e., formula (2), cannot be computed when the number of balancing conditions equals or exceeds the number of PSUs. We collapsed this region with one of its neighboring regions (stratum 7) with twelve PSUs. The variance in

stratum 7 only increased slightly after collapsing. In the matrix decomposition, there were a handful of negative values of very small magnitude in eigenvectors due to rounding errors. They were trimmed at zero to avoid negative replicate weights. The largest region (stratum 1) had 81 non-zero eigenvalues. To be consistent with other PIAAC countries, its smallest eigenvalue was dropped so that only 80 replicate weights were created. The impact of dropping one replicate weight for stratum 1 was small enough to ignore. The evaluation results showed that the variance of $\widehat{N}$ (sum of PSU weights) was underestimated by only 0.04%.

Once the replicate weights were created for non-certainty PSUs in each stratum, we randomly sorted them within strata and allocated them in a diagonal format, as shown in Figure 1. The yellow blocks denote the active replicate weights created by Fay's Method, whereas the blanks were filled in by full sample weights which do not contribute to the estimated variance. For example, stratum 1 had 80 active replicate weights; stratum 2 had 8 active replicate weights numbered as 1 through 8; stratum 3 had 13 active replicate weights numbered 9 through 21, and so on. Once it reached 80, the numbering of replicate weights went back to 1 and continued. The smallest stratum (region 22) shared 11 active replicate weights with stratum 7. Allocating the replicate weights diagonally minimized the correlation between strata. Although the active replicate weights were still stacked occasionally for some strata and may cause overestimation of variance, the impact was quite small. We computed the estimated variance of $\widehat{N}$ using the 80 replicate weights in Figure 1 (with correlation between strata) and compared it to the sum of estimated variances for all strata (without correlation between strata). The results showed that the relative difference was less than 0.01%. After the PSU-level replicate base weights were generated, the person-level replicate base weights were computed in a straightforward way as the PSU-level replicate base weights multiplied by the inverse of person-level conditional selection probabilities.

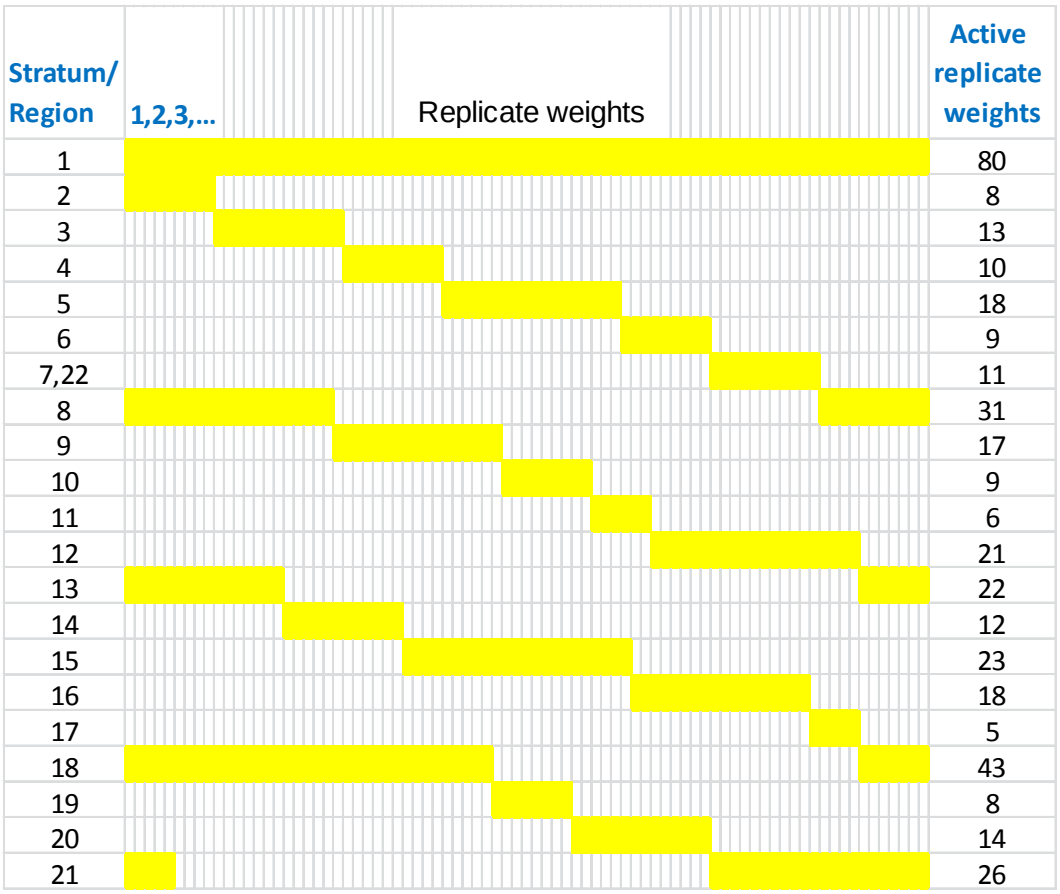

Figure 1: Allocation of replicate weights across strata (regions) for the records in the non-certainty PSUs.

As mentioned earlier, for certainty PSUs, the replicate weights were created at the person level directly. We implemented the paired jackknife approach so that the same variance formula can be used once we combined the certainty and non-certainty PSUs. In this process, the person records in the certainty PSUs were first sorted as they were ordered on the frame. Next, the variance strata were assigned sequentially from 1 to 80. Two variance units were nested within each variance stratum with each having two person records. Figure 2 shows the allocation of active replicate weights across strata. The blue blocks denote the active replicate weights, whereas the blanks were filled in by full sample weights. As shown in the last column of Figure 2, the largest region (stratum 1) had 80 active replicate weights, whereas in many of the remaining strata the certainty PSUs contained a small number of sampled persons and therefore had only a few active replicate weights.

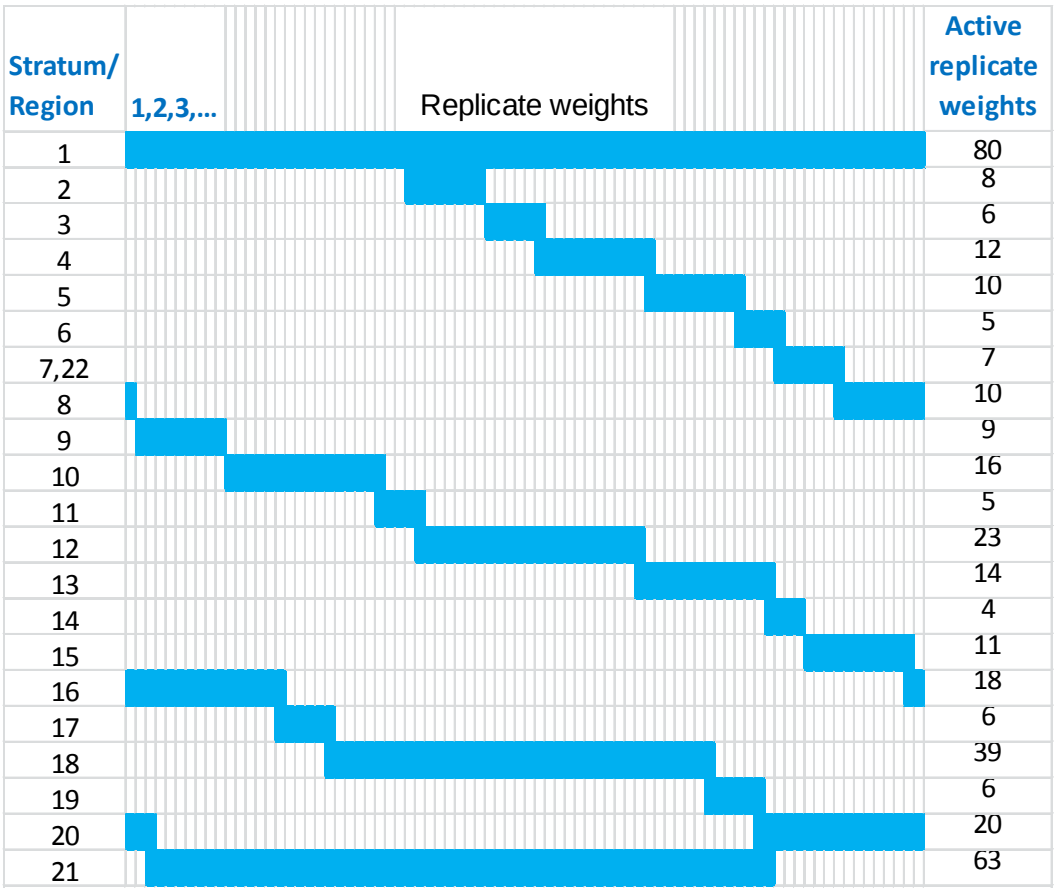

Figure 2: Allocation of replicate weights across strata (regions) for the records in the certainty PSUs.

Figure 3 shows the allocation of the active replicate weights when the person records from the certainty and non-certainty PSUs were pooled together. The green blocks indicate the overlapping of active replicate weights between the certainty and non-certainty PSUs within each stratum. The overlapping occurred in five strata and was quite minimal except in stratum 1. The correlation between the certainty and non-certainty PSUs was therefore small. This correlation can be further reduced if estimates by strata are of interest, but this is not the goal for the PIAAC study. Once the person level base weight and replicate base weights were generated, we continued to conduct other weighting steps such as nonresponse adjustment and calibration. The replicate weights were handled in the same way as the full sample weight in each of these adjustment steps.

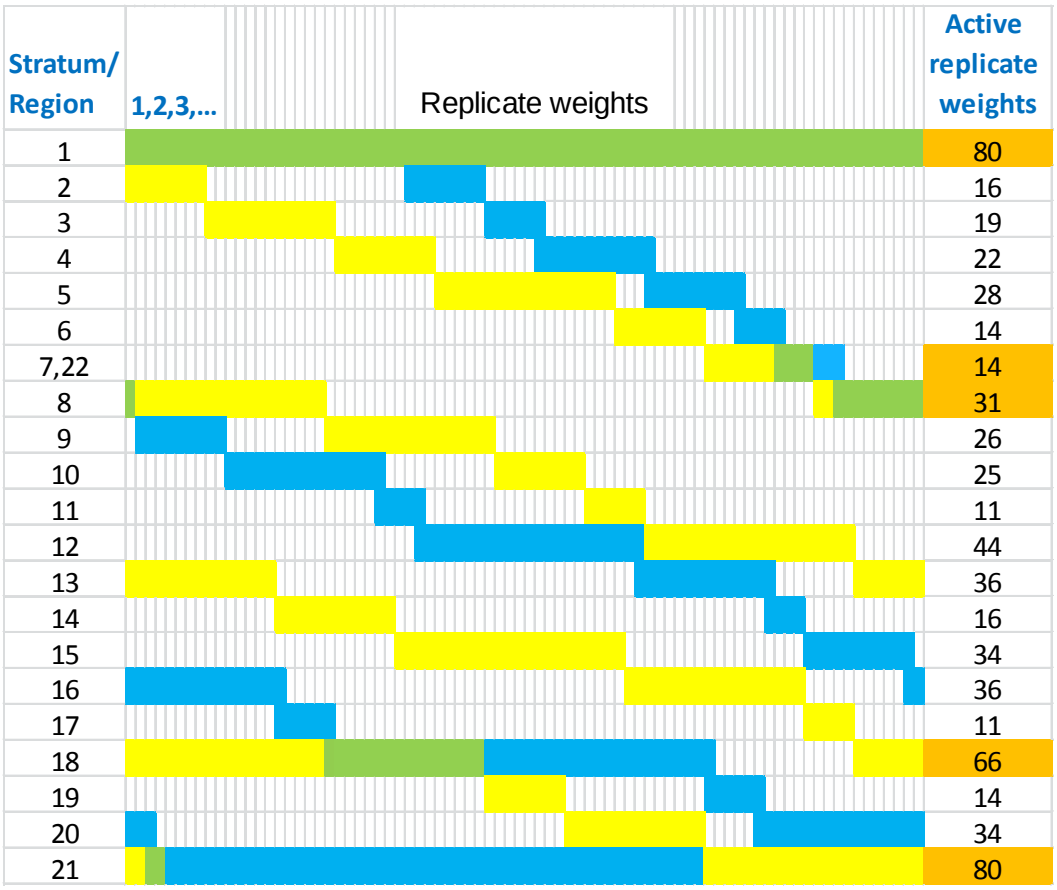

Figure 3: Allocation of replicate weights across strata (regions) for the records in all PSUs.

3. Summary

In the PIAAC study, France applied the Cube method to draw a balanced sample of PSUs. Fay's Method works well for generating replicate weights for balanced samples. The variance formula using the replicate weights created by Fay's Method takes the same form as the paired jackknife approach, which makes it convenient to combine the sample cases from the certainty and non-certainty PSUs. Most of the PIAAC countries used the paired jackknife approach. France is consistent with these countries for variance estimation though a special type of sampling was implemented. When allocating the active replicate weights, we reduced the degree of overlapping among strata, as well as between the certainty and non-certainty PSUs. This allocation scheme reduces the overestimation of variance. When the number of replicate weights is much larger than desired, the methods proposed by Kim and Wu (2013) may be considered for a reduction. In the PIAAC study for France, the overall number of replicate weights is quite large, but the number of replicate weights within stratum is within the limit of 80. In this case, the reduction of replicates can be achieved by allocating the replicates diagonally and folding them at the maximum number allowed.

References

Breidt, F.J. and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 411-425.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, Vol. 91, No. 4, 893–912.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 411-425.

Fay, R.E. (1984). Some properties of estimators of variance based on replication methods. Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC, 495-500.

Fay, R.E. and Dippo, C.S. (1989). Theory and application of replicate weighting for variance calculations. Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC, 212-217.

Kim, J.K. and Wu, C. (2013). Sparse and efficient replication variance estimation for complex surveys, *Survey Methodology*, Vol. 39, No. 1, 91-120.

Mislevy, R.J., Johnson, E.G.,and Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics*, Vol. 17, No. 2, 131-154.

Rust, K. and Rao, J.N.K (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, Vol. 5, 283–310