A Comparison of Two Clustering Criteria for Stratifying Primary Sampling Units[1]

Padraic Murphy
U.S. Census Bureau, Washington, DC 20233

Abstract

In a previous paper, we compared two methods of improving the stratification (clustering) of primary sampling units in a two-stage sample design (Murphy & Chesnut). We tentatively concluded that one of the two methods was preferable; but our conclusion depended largely on an assumption that one of two possible criteria for evaluating stratifications was better. In this paper, we show that by one specific objective measure, that assumption is justified.

Key Words: Clustering, multi-stage sampling, stratification

1. Introduction

In a previous paper, we compared two methods for identifying good (though not necessarily optimal) stratifications of Primary Sampling Units (PSUs) (Murphy & Chesnut). The first method is a "hill-climbing" algorithm due to Friedman and Rubin (Friedman & Rubin, 1967). The second combines k-means clustering with an integer programming method for solving an assignment problem (King, Schilp, & Bergmann, 2011). The first method uses a criterion function that is a weighted mean of the coefficients of variation of the stratification variables, labeled MCV, while the second uses a criterion function labeled $trace(\boldsymbol{W})$. Here, $\boldsymbol{W}$ is the within-cluster component of the total scatter matrix, which is similar to a covariance matrix. The total scatter matrix can be expressed as the sum of $\boldsymbol{W}$ and the between-cluster component, $\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B}$**.** Note that the total scatter is constant; therefore minimizing $trace(\boldsymbol{W})$ is equivalent to maximizing $trace(\boldsymbol{B})$.

One might use either criterion function (MCV or trace($\boldsymbol{W}$)) to evaluate a given stratification, independent of the method used to create that stratification. In the previous paper, we tentatively preferred MCV, because it takes into account the relative sizes of PSUs, and it does not require scaling of the stratification variables before beginning a search. However, as one might expect, the method using $trace(\boldsymbol{W})$ as a criterion function tended to out-perform the method using MCV if the criterion used to evaluate final stratifications is $trace(\boldsymbol{W})$, and vice versa. Therefore, the conclusion we reached was somewhat dependent on the choice of evaluation criterion. We wanted to answer the natural question: Is there any way to decide objectively which criterion is better? (We may sometimes use the term "metric" – this is interchangeable with "criterion" or "criterion function".)

[1] This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

2. Methodology

One of the primary objectives in PSU stratification is to minimize the first stage component of sampling variance for some population characteristic we are measuring. The example we use here is the unemployment level at a certain point in time, which is the number of persons in the labor force who are not employed as of the reference date. In our case, we assume estimates at the state level.

2.1 Notation and Sample Design Assumptions

In order to compare our two metrics (MCV and trace($\boldsymbol{W}$)) we look at the population coefficient of variation (CV), using a synthetic PSU population for which we know the true total population and unemployment level for each PSU. Here, the CV denotes the square root of the sampling variance of a total estimator divided by the population total. If Y and $\hat{Y}$ denote the true population total and total estimator for the target characteristic, respectively, and M_h and M_{hi} denote the number of population totals in stratum h and PSU i within a stratum, respectively, and we use the analogous subscripts for subdomain totals of Y, then we have the following expressions (1) and (2) for the total estimator and its CV, respectively:

$$\hat{Y} = \sum_{h=1}^{L} \frac{M_h}{M_{hi}} Y_{hi} \tag{1}$$

$$CV(\hat{Y}) = \frac{\sqrt{V(\hat{Y})}}{E(\hat{Y})} = \frac{\sqrt{\sum_{h=1}^{L} M_h \left(\sum_{i=1}^{N_h} M_{hi}\left(\bar{Y}_{hi} - \bar{\bar{Y}}_h\right)^2\right)}}{Y} \tag{2}$$

Here, L is the number of first stage strata, N_h is the number of PSUs in stratum h, and $\bar{\bar{Y}}_h$ and $\bar{Y}_{hi}$ are the true population means for Y in stratum h and PSU i, respectively; so $\bar{\bar{Y}}_h = Y_h/M_h$, and $\bar{Y}_{hi} = Y_{hi}/M_{hi}$.

It is important to note that we assume here a design in which we select one PSU from each stratum with probability proportional to size (as measured by the number of second stage units). Note that with this understanding, the estimator $\hat{Y}$ is unbiased. In addition, since we are only concerned with the first stage of sampling, we make the simplifying assumption that the second stage component of sampling variance will be zero. We also ignore the fact that many of the large demographic surveys really have an intermediate stage of sampling where they select households, with the final sample of persons taken from within the sample households, but this should not have any impact on conclusions about the first stage sample design.

2.2 Description of Experiment

Our two metrics, MCV and $trace(\boldsymbol{W})$, are calculated as functions of the stratification variables alone, since it is assumed the values of the target variable are unknown at the time of stratification; but each stratification also results in a latent CV value for the target variable. Ideally, we would like to select the stratification that has the lowest CV for the *target variable*. Thus, there is a "true" ranking of the possible stratifications based on the CV of the target variable, and there is a ranking for each of the metrics based on their respective values as functions of the stratification variables. We can compare these rankings to see how well each metric does with respect to the "true" ranking. Furthermore, we can calculate the Pearson correlation of the CV values of the target variable with the values of $trace(W)$ and MCV, respectively. As we shall see, this Pearson correlation tends to be higher (closer to 1) for MCV than for $trace(W)$.

We simulated values of both a target variable (the characteristic being measured by the survey, e.g., unemployment level) and several stratification variables (e.g., historical unemployment levels, number of persons employed in manufacturing) at the PSU level for 346 PSUs. We based the simulated unemployment level values on actual survey data, with the objective of having distributions in our synthetic data set reflect distributions one might actually find in the real world. This is how we arrived at this particular number of PSUs. However, since data is only available for a sample of PSUs, and to preserve survey confidentiality, we grouped these 346 PSUs into 7 synthetic states, which approximately correspond to certain regions in the United States. Table 1 summarizes some state level data.

Table 1: Summary Statistics for Synthetic States

Synthetic State	*PSU count*	*Labor Force Level (in thousands)*	*Unemployed Level (in thousands)*
1	46	5,276	313
2	53	6,471	287
3	59	4,491	169
4	60	5,690	316
5	41	3,774	217
6	48	4,359	220
7	39	2,286	122

Source: Current Population Survey (CPS) 2015-2017

For the stratification variables, we selected a set of characteristics available to the Current Population Survey (CPS) for PSU stratification in the sample redesign following the 2010 Census. For each synthetic state, we selected the four stratification variables with the highest Pearson correlations with respect to the unemployment level. Here, both target and stratification variable values are at the PSU level within each synthetic state. Table 2 has a list of the stratification variables. Table 3 shows the Pearson correlation value with the target variable for each stratification variable and synthetic state. The starred values are the four highest in each state.

The next step was to create a benchmark stratification of the PSUs in each of our synthetic states. For a given synthetic state, this was done as follows:

- For each PSU, calculate the 27-month average unemployment rate, dividing the unemployment level by the labor force level. (These are the PSU-level analogues of those columns in Table 1.)
- Sort the PSUs by this unemployment rate.
- Determine the maximum number of strata that can be formed such that each stratum contains at least two PSUs. Let this maximum number be M. Let L be the number of strata to form in a state. For each value of L in the range from 3 up through M, perform the following steps:

  - Create an initial stratification that will not necessarily satisfy the requirement that the measures of size of all strata be within ten percent of the average stratum size. Starting with the first PSU in the sorted list, add PSUs to the first stratum until its total measure of size is greater than 95 percent of the average stratum size. Continue building strata this way through the end of the PSU list.
  - The number of strata created in the previous step could be less than L, but in that case just create one or more empty strata until there are L strata.
  - Check whether all strata satisfy the size constraint. If not, identify the stratum with the largest measure of size, and move the smallest PSU from this stratum to the smallest stratum. Continue iteratively until all strata satisfy the size constraint, or until a maximum number of iterations are done.
  - If the maximum number of iterations is reached without satisfying the size constraint, do not attempt stratification for this value of L. Otherwise, the current assignment of PSUs to strata is the candidate benchmark stratification for this value of L.
  - If a satisfactory stratification for this value of L exists, calculate the coefficient of variation for the unemployment level estimate that would be obtained by sampling one PSU from each stratum.
- For all the values of L with a satisfactory stratification, compare the resulting CV values for a state unemployment level estimate. Let L^* be the value of L resulting in the lowest CV. Then L^* will be the number of strata formed for all of the alternative stratifications to be ranked. Also, the corresponding PSU stratification is the benchmark stratification for this state.

Table 2: Stratification Variable Descriptions

Variable Name	*Description*
TOTID184	Households in Census 2000 with at least one person 0+ in Poverty within housing unit
TOTID448	Families with female heads, no husband present
TOTID464	Population age 0 to 5
TOTID480	Owner-occupied housing units with value $90,000 to $99,999
TOTID520	Families with female heads
TOTID529	Persons age 16-19 unemployed or not in labor force
TOTID536	Related children in poverty (age 0 to 17)
TOTID541	Occupied housing units with that are electric heated
TOTID547	Population for whom poverty status is undetermined
TOTID550	Related children (age 0 to 17)
TOTID597	Average number of reported motor vehicle thefts 2002 to 2009
TOTID610	Number of reported aggravated assaults in 2009
TOTID615	Number of reported violent crimes in 2009
TOTID654	Number of households with female head-of-household
TOTID685	Persons in poverty, age 0+
TOTID688	2010 Unemployed Total 16+

Table 2: Stratification Variable Descriptions

Variable Name	*Description*
TOTID689	2010 Unemployed Female 16+
TOTID702	Renter-occupied housing units with rent less than $700

Source: U.S. Census Bureau, Demographic Statistical Methods Division, 2010 Sample Redesign Requirements

Table 3: Correlations of Selected 2010 MSPF Stratification Variables (Weighted) with CPS 27-month Average Unemployment Level, by Synthetic State

Stratification Variable	*Synthetic State*						
	1	**2**	**3**	**4**	**5**	**6**	**7**
TOTID184	0.6458	0.3948	0.7057	0.5091	-.0106	*0.4811	0.7214
TOTID448	0.8104	*0.5698	*0.8110	*0.5647	0.2676	*0.5260	0.8298
TOTID464	0.8173	0.3504	*0.7824	0.4415	0.3159	0.4091	0.7924
TOTID480	0.1922	0.2359	0.6454	0.4955	*0.4091	0.1164	0.4193
TOTID520	0.7865	*0.5632	0.7658	*0.5828	0.1877	*0.5087	0.7485
TOTID529	*0.8254	0.3343	0.7266	0.4911	0.0746	0.3171	*0.9008
TOTID536	0.7973	0.4378	0.7726	0.5003	0.0005	0.2479	*0.8948
TOTID541	0.4832	0.2495	0.5319	0.3538	*0.4422	0.2102	0.7240
TOTID547	0.7778	*0.3638	0.7783	0.5220	0.1190	0.2260	0.8611
TOTID550	*0.8238	0.3523	0.7560	0.4894	0.3133	0.4410	0.8218
TOTID597	0.6557	0.3372	0.5968	0.3244	0.0978	0.2150	*0.9124
TOTID610	0.5599	0.1525	0.6192	0.3000	*0.4250	*0.4710	0.6424
TOTID615	0.6153	0.1956	0.6577	0.3372	*0.4371	0.4273	0.6633
TOTID654	0.6766	*0.5541	0.7347	*0.5902	0.3011	0.4292	0.5856
TOTID685	0.7778	0.3638	*0.7783	0.5220	0.1190	0.2260	0.8611
TOTID688	*0.8280	0.3980	0.7333	0.4911	0.2013	0.3682	0.8103
TOTID689	*0.8454	0.3119	0.7084	0.4964	0.2332	0.3206	*0.8873
TOTID702	0.6374	*0.4803	0.7083	*0.5419	0.3720	0.2177	0.5323

Source: Current Population Survey (CPS) 2015-2017

With a benchmark stratification for each state, the next step was to form alternative stratifications (all with L^* strata) with higher unemployment estimate CV levels. We created ten alternative stratifications for each state, swapping PSUs of similar size between strata until the CV increased to a given value, and doing this ten times for ten CV values that were evenly spaced (roughly) over an appropriate interval for each state. We also randomly generated 200 distinct stratifications for each state (all with L^* strata) that satisfied the size constraints. There were no target CV values for the random stratifications, the idea being to get a sense of how CV values are distributed across the population of all possible stratifications.

In addition, for each of the four stratification variables used for each state, we created one stratification (with L^* strata) in the same way we created the L^*-strata candidate for the

Benchmark, except we used the stratification variable in place of the target variable. We noticed when calculating the MCV metric for the Benchmark and other stratifications that the Benchmark MCV tended to fall towards the middle of the range of MCV values we were seeing. This was also true for the CV values of each stratification variable. Our intuition is that for any variable, the distribution of its CV values across all possible stratifications is approximately normal, so that most stratifications would result in CVs in the middle part of the range. This means that even though a given stratification may have a CV in the lower tail of the distribution for variable A, it is likely that for a different variable B, with the same stratification, the CV value for B will be in the middle part of variable B's CV distribution. If B is strongly correlated with A, the CV value for B will tend to be below average, but still towards the middle. We were seeing this when A was the target variable and B was a stratification variable; and we wanted to check our intuition by seeing if the same thing happens when A is a stratification variable and B is the target variable. And in fact, this does seem to be the case. Secondarily, this gave us four additional stratifications for our experimental sample.

Finally, having created a total of 215 PSU stratifications for each synthetic state, we were ready for the final steps of our experiment, which we describe as follows for one state:

- Calculate the value of $trace(W)$ for each of the 215 stratifications, using the stratification variables indicated by the shaded cells in Table 3.
- Calculate the Pearson correlation of $trace(W)$ with the unemployment level CV (CV_{UE}) across the sample of 215 stratifications
- Repeat the three previous steps, replacing $trace(W)$ with MCV.

3. Results

The results of our experiment are shown in Table 4, and illustrated by the paired scatter plots in Figures 1-7.

Table 4: State-level Comparison of the Correlations of the Stratification Metrics [$trace(W)$ versus MCV] with CV_{UE}

Synthetic	$\rho(trace(W), CV_{UE})$		$\rho(MCV, CV_{UE})$	
State	Estimate	95% c.i.	Estimate	95% c.i.
1	-0.059	(-0.191, 0.076)	0.397	(0.278 , 0.504)
2	-0.006	(-0.139 , 0.128)	0.284	(0.156 , 0.402)
3	-0.185	(-0.311 , -0.053)	0.647	(0.562 , 0.719)
4	-0.382	(-0.490 , -0.261)	0.433	(0.317 , 0.536)
5	0.047	(-0.088 , 0.179)	-0.131	(-0.261 , 0.003)
6	0.109	(-0.025 , 0.239)	0.188	(0.055 , 0.314)
7	-0.195	(-0.320 , -0.063)	0.585	(0.489 , 0.666)

Shading indicates that we could not reject the null hypothesis that population correlations corresponding to the two sample estimates are equal.

Source: Current Population Survey (CPS) 2015-2017

Note: We calculated the correlation estimates and 95% confidence limits in Table 4 with the SAS® procedure PROC CORR, using the "FISHER" option in the PROC CORR statement, which invokes the Fisher z-transformation. We were also able to use output from the procedure to test whether the population correlations corresponding to these sample estimates are different. We used a significance level of 0.05 for these tests.

The test we used assumes the two correlations come from two independent samples. In this case, the two correlations are from the same sample, so clearly are not independent. However, we believe the consequence of having dependent samples is that the test statistic is smaller than it should be, due to ignoring a negative covariance term in its denominator. This makes the test overly conservative – that is, less likely to detect a significant difference – but we do not think it affects our results.

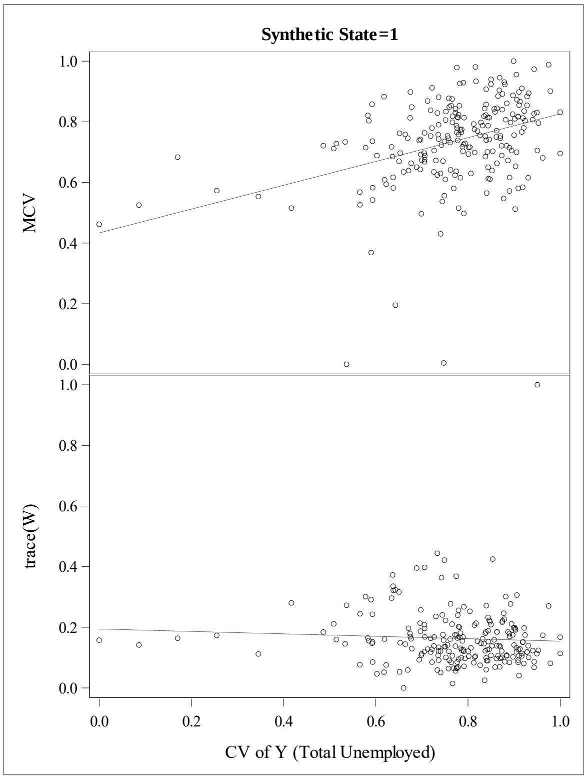

Figure 1: Comparison of MCV and $trace(W)$ Correlations with CV_{UE} for Synthetic State 1

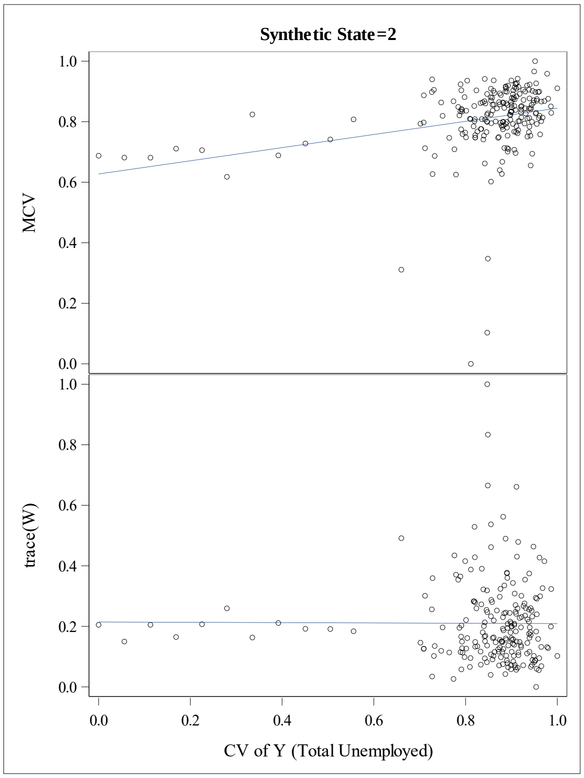

Figure 2: Comparison of MCV and $trace(W)$ Correlations with CV_{UE} for Synthetic State 2

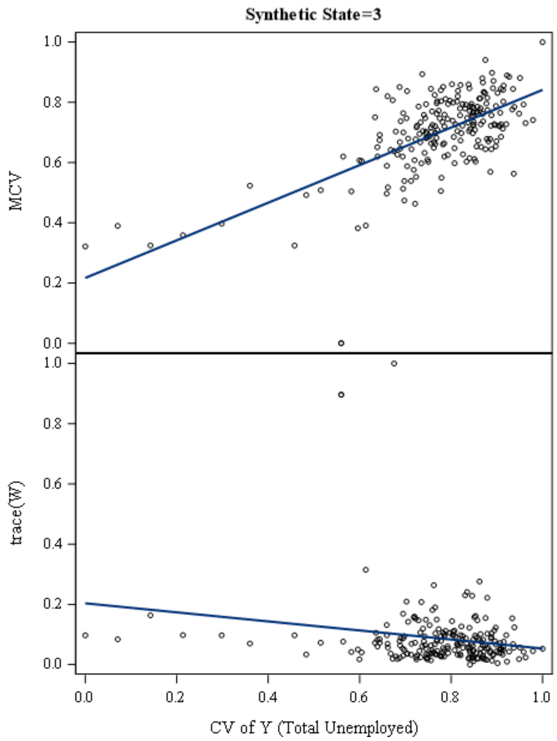

Figure 3: Comparison of MCV and $trace(W)$ Correlations with CV_{UE} for Synthetic State 3

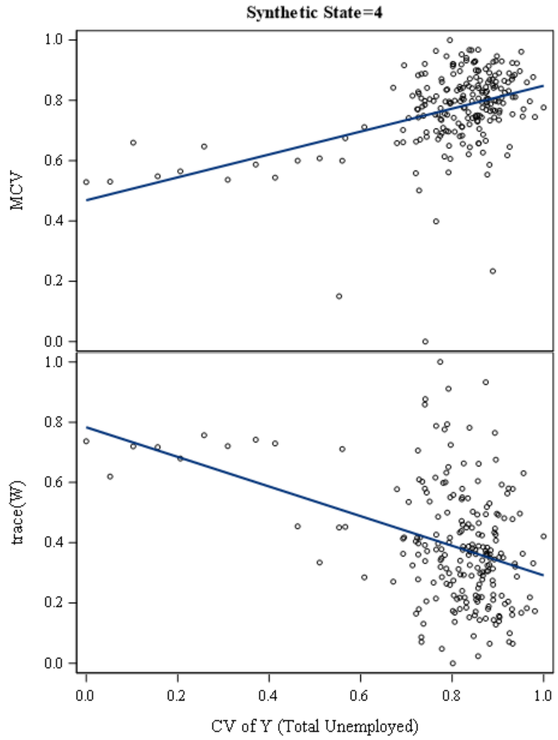

Figure 4: Comparison of MCV and $trace(W)$ Correlations with CV_{UE} for Synthetic State 4

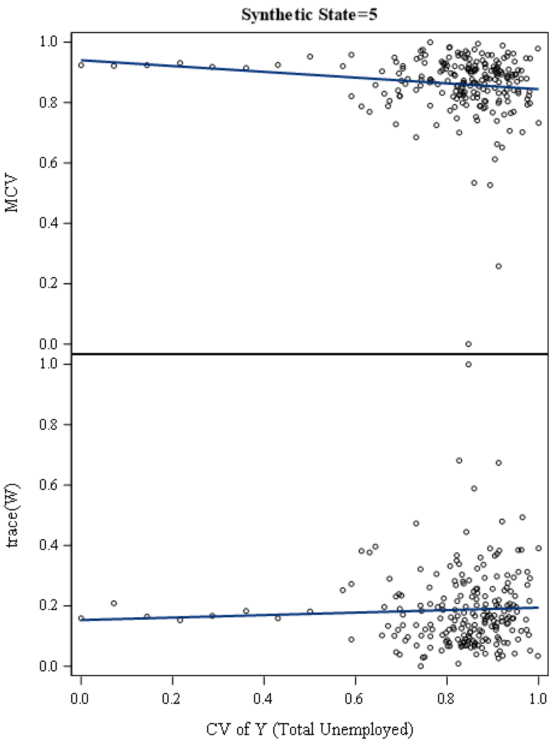

Figure 5: Comparison of MCV and $trace(W)$ Correlations with CV_{UE} for Synthetic State 5

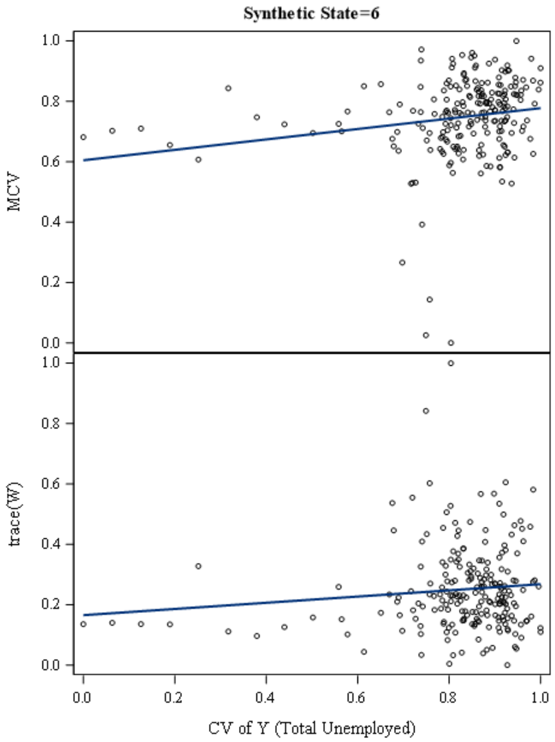

Figure 6: Comparison of MCV and $trace(W)$ Correlations with CV_{UE} for Synthetic State 6

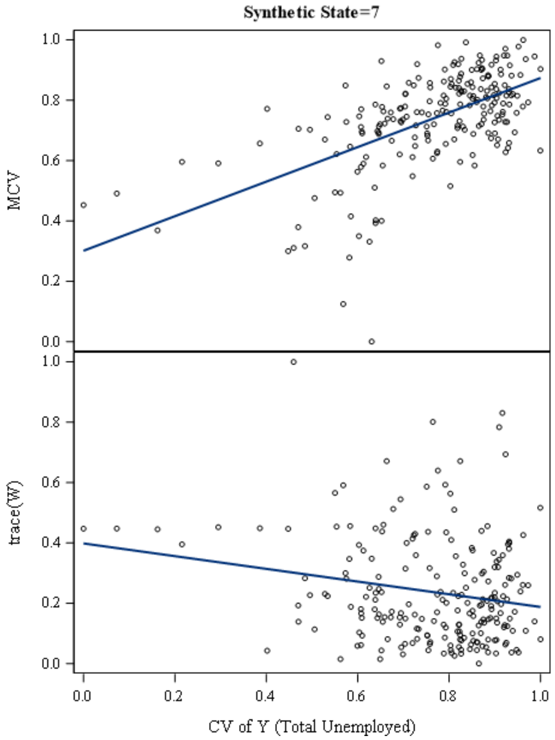

Figure 7: Comparison of MCV and $trace(W)$ Correlations with CV_{UE} for Synthetic State 7

4. Discussion of Results

For all of the synthetic states except 5 and 6, it is clear from Figures 1-7 and Table 4 that the MCV metric is more highly correlated with CV_{UE} than $trace(W)$. In Table 4, note that the 95% confidence interval for the MCV correlation is above and does not overlap the corresponding $trace(W)$ interval at all, except in states 5 and 6. In state 5, the intervals for both metrics include zero. In state 6, only the $trace(W)$ interval includes zero, and the MCV correlation point estimate is slightly higher, but there is a lot of overlap between the intervals. In both states 5 and 6, we failed to reject the null hypothesis.

The relatively poor performance of the MCV metric in states 5 and 6 is explained by the relatively low values of the stratification variable correlations with the target variable in those two states. From Table 3, note that the best stratification variable has a correlation value of 0.44; and the corresponding value for state 6 is 0.53. The average correlation values across the four stratification variables for states 5 and 6 are 0.43 and 0.50, respectively. In contrast, the average values in the other states range between 0.54 and 0.90, and the best values range between 0.59 and 0.91.

Considering the performance of $trace(W)$, note that the confidence interval for the metric's correlation with CV_{UE} includes zero in all seven states, and the point estimates are negative for all but one state. Essentially, this means that using $trace(W)$ to select a stratification is just as likely to result in a poor result as a good result with respect to the target variable, even when the stratification variables are highly correlated with the target variable.

We believe this is very strong evidence that MCV is superior to $trace(W)$ as a PSU stratification metric, in the following sense: If a survey has a single key target variable it is estimating, and if the stratification variables have reasonably high positive correlation with the key target variable, then MCV will be more highly correlated with the coefficient of variation (CV) of the key target variable than $trace(W)$.

We believe that $trace(W)$ does poorly relative to MCV because it ignores the PSU measure of size. As evidence of this, consider the CV values for the 2010 PSU measure of size shown in Table 5. The states with the three lowest values of the measure of size CV – states 2, 5, and 6 – are also the states with the three smallest absolute differences between the MCV and $trace(W)$ point estimate correlation values from Table 4. (The difference is greater for state 2 because of higher correlation of its stratification variables with the target.)

While the MCV metric does well relative to the $trace(W)$ metric, it is obvious from the scatter plots in Figures 1-7 that the stratification with the lowest MCV value is not the Benchmark for any of the states. In each scatter plot, the Benchmark is represented by the point furthest to the left. Also, the lowest point in each scatter plot (closest to the horizontal axis) represents the stratification that would be selected as "best" from this sample if the metric on the vertical axis were the selection criterion.

Table 5: Coefficient of Variation Values CPS 2010 Measure of Size

Synthetic State	MOS CV	$\|\rho(MCV, CV_{UE}) - \rho(Tr(W), CV_{UE})\|$ from Table 4
1	38.7%	0.46
2	17.2%	0.29
3	52.8%	0.83
4	24.9%	0.82
5	18.1%	0.18
6	17.1%	0.08
7	69.5%	0.78

Source: Current Population Survey (CPS) 2015-2017

Table 6 shows the MCV and $trace(W)$ ranks for the Benchmark, as well as the Unemployment Level CV (CV_{UE}) ranks for the stratifications with the lowest MCV and $trace(W)$ values, by synthetic state.

Table 6: Ranks by Value of CV_{UE}, MCV, and $trace(W)$ for Selected Stratifications

Stratification Description	*Metric*	*Synthetic State*						
		1	*2*	*3*	*4*	*5*	*6*	*7*
Benchmark (lowest CV_{UE})	MCV	6	17	3	6	165	53	15
	$trace(W)$	121	128	157	203	107	37	190
lowest MCV	$trace(W)$	194	196	213	212	215	215	164
	CV_{UE}	11	45	10	41	100	62	37
lowest $trace(W)$	MCV	41	205	162	195	207	203	74
	CV_{UE}	34	197	138	78	36	181	144

Source: Current Population Survey (CPS) 2015-2017

Note that except for state 5, the CV_{UE} rank is better for the lowest MCV stratification than for the lowest $trace(W)$ stratification. Also, except for states 5 and 6, the MCV rank for the Benchmark is under 20, out of 215, putting it in the lowest ten percent.

References

Friedman, H. P., & Rubin, J. (1967). On Some Invariant Criteria for Grouping Data. *Journal of the American Statistical Association*.

King, S. L., Schilp, J., & Bergmann, E. (2011). Assigning PSUs to a Stratification PSU. *Proceedings of the Joint Statistical Meetings - Section on Survey Research Methods*, (pp. 2235-2246).

Murphy, P. A., & Chesnut, J. (n.d.). A Comparison of Clustering Algorithms Used For Multivariate Stratification of Primary Sampling Units. *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*, (pp. 1477-1495).