GOVERNMENTS DIVISION REPORT SERIES

(Research Report #2010-5)

Alternative Sample Approach for the Annual Survey of Public Employment and Payroll

Casey Corcoran
Yang Cheng

U.S. Census Bureau
Washington, DC 20233

CITATION: Corcoran, Casey and Cheng, Yang. 2010. Alternative Sample Approach for the Annual Survey of Public Employment and Payroll. Governments Division Report Series, Research Report #2010-5

Report Completed: September 23, 2010
Report Issued: October 14, 2010

Disclaimer: This report is released to inform interested parties of research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Alternative Sample Approach for the Annual Survey of Public Employment and Payroll

Casey Corcoran, Yang Cheng
Governments Division
U.S. Census Bureau[1], Washington, D.C. 20233-0001

Abstract

The Governments Division of the U.S. Census Bureau publishes estimates of many aggregates. To achieve more accurate results with lower cost one goal of sampling is to minimize the number of smaller and non-contributory units in the overall design. Unfortunately, little research has been conducted to determine how to reduce smaller units. In this paper, we propose a numerical procedure based on minimizing the sum of mean squared errors to identify the breakpoint for determining the size-based strata. We then vary the amount of small stratum reduction to find an optimal balance between cutpoint and reduction. Data from the 2007 Census of Governments: Employment are used for this study.

Key Words: survey design, decision-based estimation, mean squared error

1. Introduction

The annual survey of Public Employment and Payroll collects data on the number of federal, state, and local government civilian employees and their gross payrolls. The 2007 Census of Governments yielded a total of 89,527 governments (including the federal government). Their breakdown is as follows:

Table 1: Breakdown of 2007 Census of Government Employment

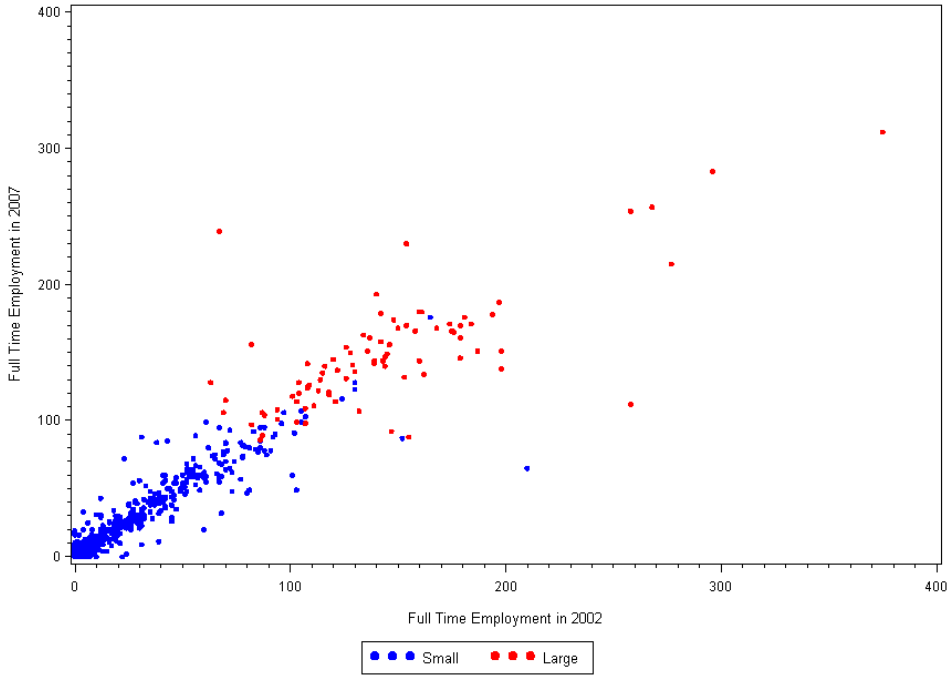
Government Type	*N*	*Total Employees*	*Total Payroll*
States	50	5,200,347	$17,788,744,790
Counties	3,033	2,928,244	$10,093,125,772
Cities	19,492	3,001,417	$11,319,797,633
Townships	16,519	509,578	$1,398,148,831
Special Districts	37,381	821,369	$2,651,730,327
School Districts	13,051	6,925,014	$20,904,942,336
Total	**89,526**	**19,385,969**	**$64,156,489,693**

Source: U.S. Census Bureau, 2007 Census of Governments: Employment.

We learn from table 1 that special districts and townships are larger in number and do not contribute much to payroll and employees totals compared to their counterparts. In addition, our census and surveys have demonstrated that smaller governments are less likely to be respondents and have less variability than larger governments. To illustrate these points, the scatter plot in Figure 1 shows the concentration of smaller governments and the linear relationship between census years in Illinois as well as the decrease in variability of smaller governments.

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

Figure 1: Results from Cumulative Square Root of the Frequency Method (using total pay as size variable) on Illinois City and Townships



Source: U.S. Census Bureau, 2002 and 2007 Census of Governments: Employment

Because the smaller government units do not contribute much to the overall total and are more likely to be nonrespondents, one goal of sampling is to minimize the number of smaller and non-contributory units in the overall design. In 2009 a new sample design was implemented to address this issue. It was decided that cities and townships were similar enough that they could be combined into a single stratum, called cities/townships. We developed a new two-stage cut-off sample in order to reduce the overall sample size, thereby reducing respondent burden and improving data quality. After certainties are removed, the first stage of the sample design is a probability proportional-to-size (PPS) sample stratified by state and government type. Total payroll from the 2007 Census of Governments: Employment was used as the size variable. The first-stage sample yielded excessive amounts of smaller cities/townships and special districts. After the first-stage sample, a cut-off point was found for each state-by-government type stratum for cities/townships and special districts. These cut-off points were chosen based on the cumulative square root of the frequency method and represent the decision point for distinguishing small and large governmental units in the stratum. Instead of conducting a standard cut-off sample where we would ignore the units below the cut-offs, we chose to subsample the smaller cities/townships and special districts using simple random sampling. Although the smaller units do not contribute much to the overall total this method still allows us to estimate their share.

In this paper, we are attempting to find a way to divide a population into small and large strata and then subsample the strata, drawing even fewer units from the small strata into the sample. In addition, we wish to conduct more research with this method and propose new methods that might be better at defining these cut-off points.

2. Objective

Before discussing the objective of this paper, we first define some notation which will be used throughout the paper. Let S be the overall sample of size n and $S = S_1 \cup S_2$ where S_1 is the small size unit sub-stratum with size n_1 and S_2 is the large size units sub-stratum with size n_2. Cut-off point, c, is defined as the breakpoint between size-based strata. This cut-off point distinguishes small and large sub-samples. Also, we define p as the sample rate where we sub-sample S_1* from S_1. The sub-sample size n_1* is equal to pn_1.

We aim to find an acceptable range of small strata reduction which will result in a smaller overall sample size without significant gains in total mean squared error. We define total mean squared error as follows:

$$\min_{\substack{x_{\min} \le c \le x_{\max} \\ 0 \le p \le 1}} \left\{ MSE_1(c, p) + MSE_2(c) \right\}$$

Where:

$$MSE_1(c, p) = \frac{1}{n_1^* - 2} \sum_{\{j: x_j \le c,\, j \in S_1^*\}} \left(y_{1,j} - \hat{y}_{1,j} \right)^2$$

$$MSE_2(c) = \frac{1}{n_2 - 2} \sum_{\{j: x_j > c,\, j \in S_2\}} \left(y_{2,j} - \hat{y}_{2,j} \right)^2$$

The MSE formulas can be found in many standard sampling text book. $\hat{y}_j$ is the linear predictor from the sample data corresponding y_j. We are minimizing the sum of MSE over small and large sample sets to obtain the optimum of the cut-off point and sub-sample rate.

2.1 Simulated Data

Before applying this theory to historical Census Bureau data, we first wished to examine the effects of cut-off point selection and small strata reduction on simulated data. The simulated data are piecewise and based on the following distributions and equations:

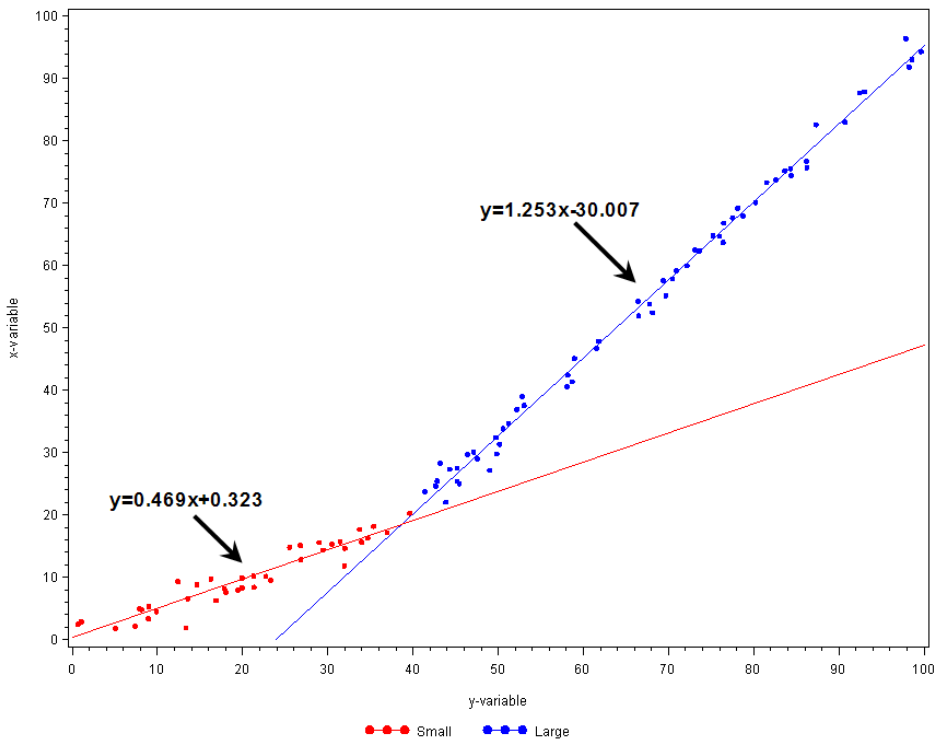
$$y = \begin{cases} 0.5x + e & x \le 40 \\ 1.25x - 30 + e & x > 40 \end{cases}$$

Where:

$x \sim$ uniform(0, 100) and $e \sim$ normal(0, 1.5)

The corresponding data are plotted in Figure 2. Ignoring the smaller strata reduction by setting p equal to 1, we first wished to examine the effect of cut-off point selection on total mean squared error. All possible cut-off points were examined, meaning the first cut-off point would result in the smallest observation in the small stratum and the remaining ninety-nine observations in the larger sized strata. The second cut-off point would result in the smallest two observations in the smaller stratum and the remaining ninety eight observations in the larger stratum and so on. This process resulted in ninety-eight possible cut-off points with ninety-eight total mean squared errors. The cut-off points and total mean squared errors were plotted in Figure 3 to find the optimal cut-off point.

Figure 2: Simulated Data with Noise and Linear Fits

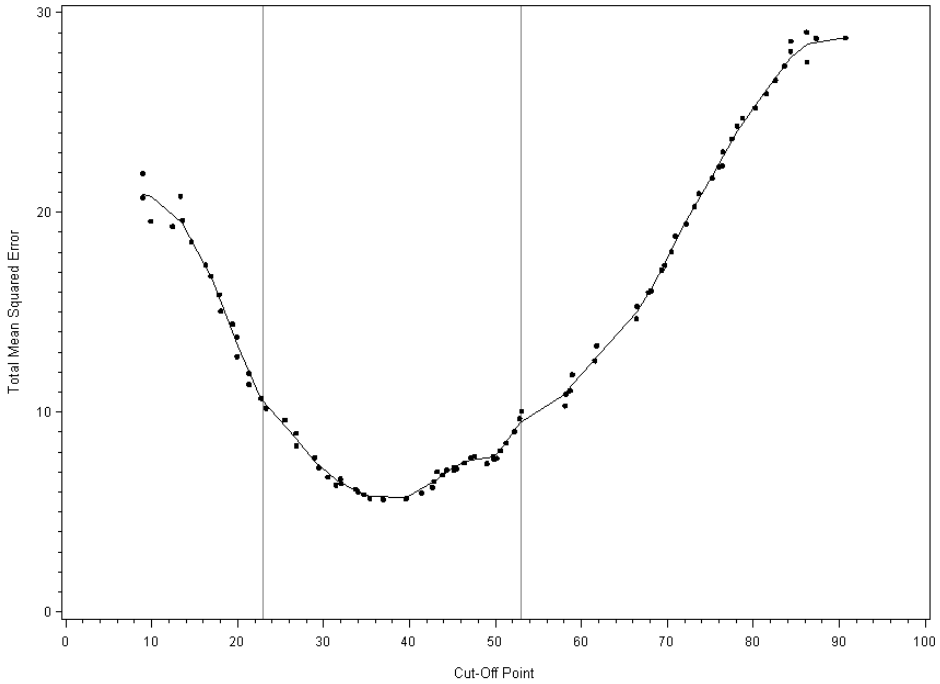


Source: Simulated data for illustrative purposes only

The vertical lines show an arbitrarily chosen acceptable range of total MSEs we can tolerate. Naturally we would want to choose the minimum point in this curve, around 40, however the cutpoint can be moved about 10 units to either side without a great loss in total MSE.

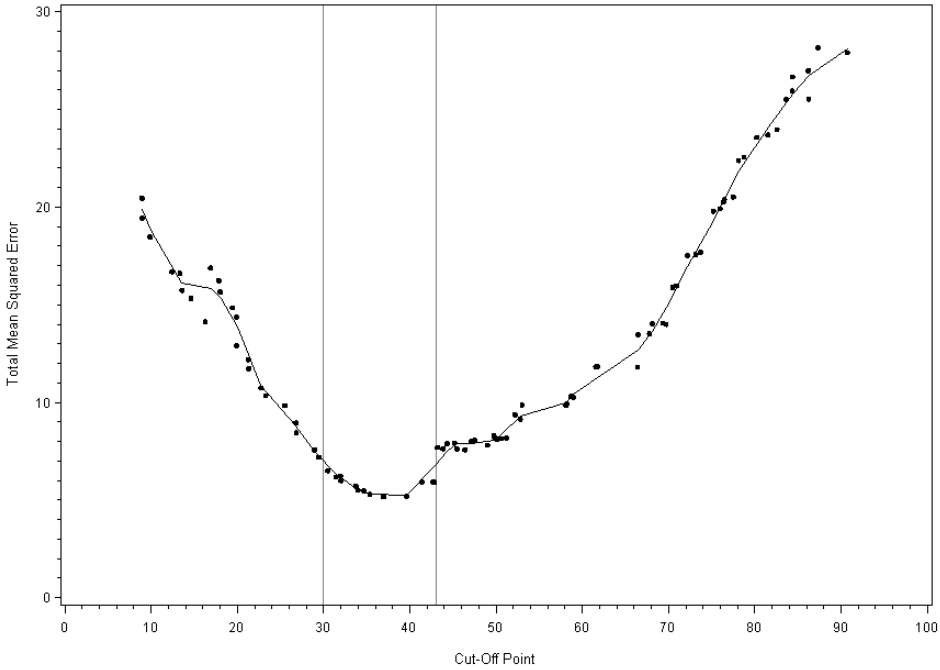
Figure 4, shown below, plots the possible cut-off point choices versus the total mean squared error, but 60 percent of the smaller stratum is reduced systematically each time. For example, if we chose the cut-off point, 40, we would have 40 points in the smaller stratum and 60 points in the larger stratum. We would then reduce the smaller stratum by 60 percent, leaving 16 points, and calculate a total MSE. The resulting plot displays a more variable total mean squared error. This result is expected as mean squared error is dependent on the total sample size. The minimum mean squared error is still in relatively the same place, around 40, but the acceptable MSE range is narrower.

Figure 3: Graph of Simulated X-Value Cut-Off Points Versus Total Mean Squared Error



Source: Simulated data for illustrative purposes only

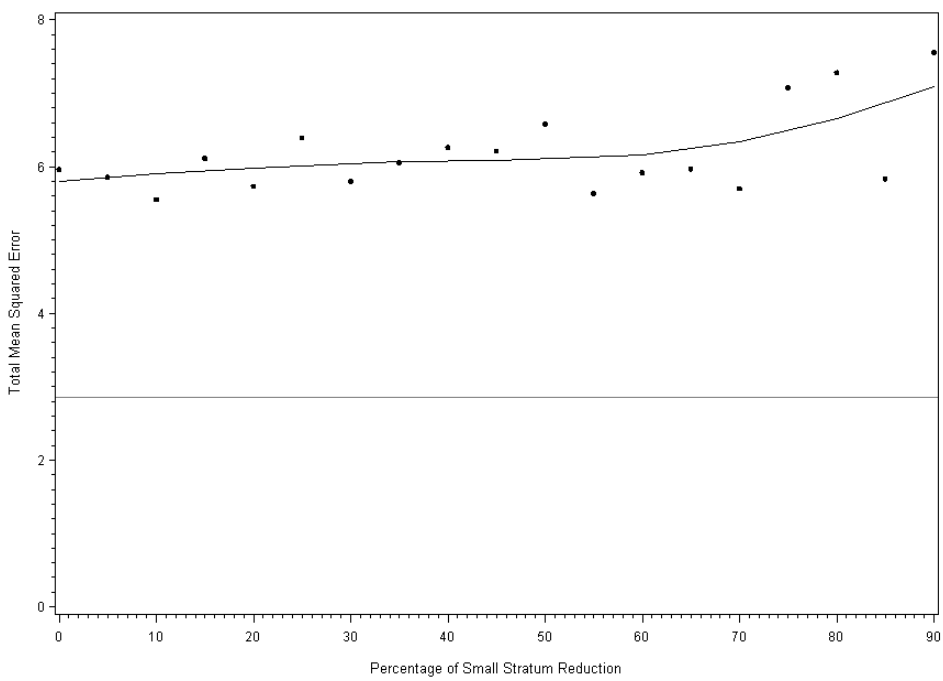
Figure 4: Graph of Simulated X-Value Cut-Off Points Versus Total Mean Squared Error – Reducing the Smaller Stratum by 60 Percent



Source: Simulated data for illustrative purposes only

For a given cut-off point, we wish to randomly sample the smaller size stratum. When keeping the cut-off point constant we can try varying levels of r, the sampling rate. In Figure 5, we have fixed the cut-off point at 40 and randomly selected smaller stratum samples, varying the level or r by 5 percent increments. The results show the total mean squared error stays roughly constant with small stratum reduction to a certain extent. However, when you remove over 70 percent of the data in the smaller stratum, leaving only 12 points to run regression, the total mean squared error appears to suffer and grow more variable. Again, the acceptable range of Total MSE is arbitrarily chosen in this case. This suggests the mean squared error is controlled much more by the cut-off point we select versus the sampling rate of reduction. This theory becomes even more apparent in Figure 6 where we held the cut-off point at 60, instead of 40, and systematically selected smaller stratum samples. The variability increased much faster as the percentage of small stratum reduction increased.

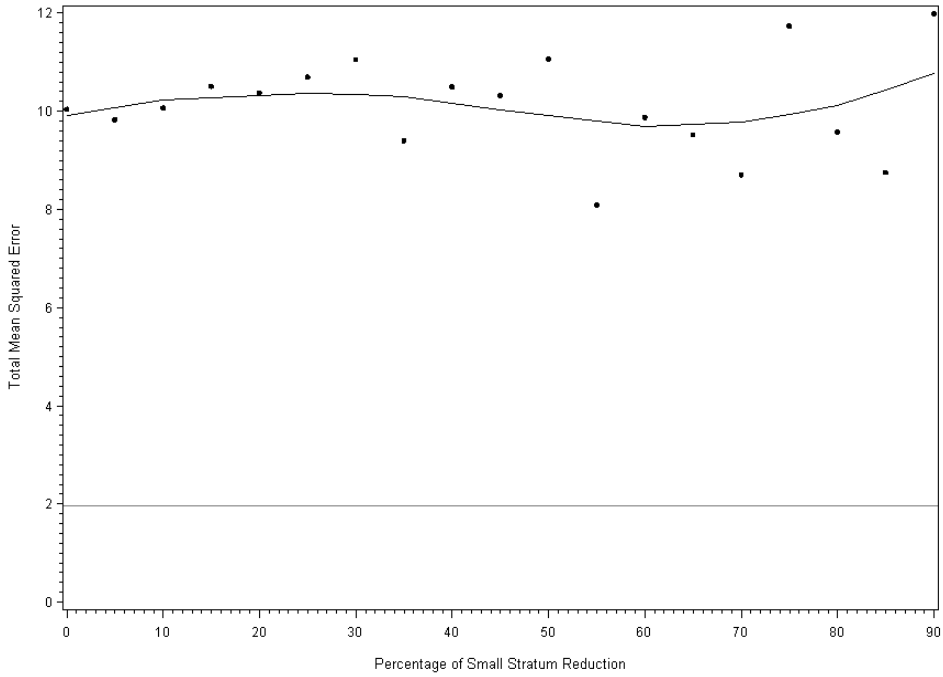
Figure 5: Holding the Cut-Off Point at 40 and Fluctuating the Percentage of Small Strata Reduction



Source: Simulated data for illustrative purposes only

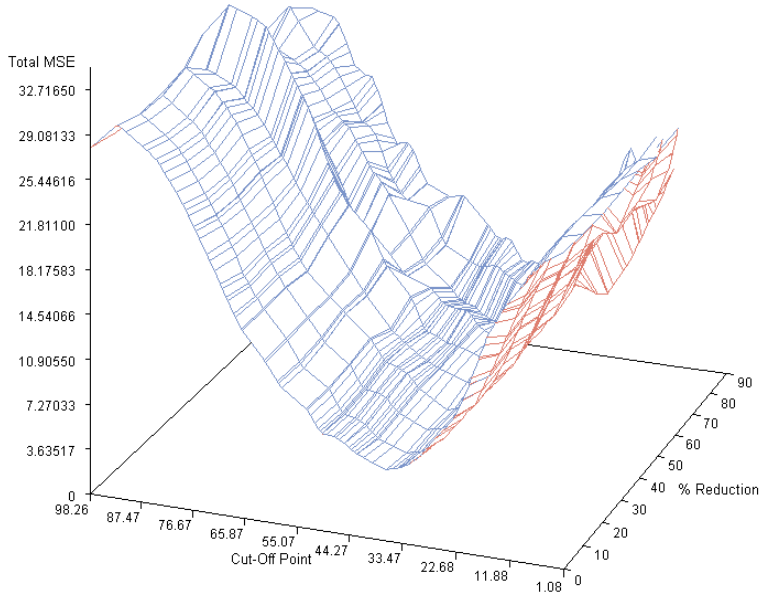
Lastly, we wish to combine all three variables (cut-off point, percent reduction, total mean squared error) together to make a three-dimensional plot. Again we notice that the cutpoint is the more influential variable in the total mean squared error calculation. As the reduction rate fluctuates for a given cut-off point the total mean square error stays relatively consistent.

Figure 6: Holding the Cut-Off Point at 60 and Fluctuating the Percentage of Small Strata Reduction



Source: Simulated data for illustrative purposes only

Figure 7: All Possible Combinations of Cut-Off Point, Small Strata Reduction and Resulting Total Mean Squared Error – Simulated Data



Source: Simulated data for illustrative purposes only

2.2 Census of Governments: Employment Data

For this paper, we chose two state-by-government types to try this process on, namely California special districts and Pennsylvania cities/townships. California was chosen because it is a large state with many special district governments and Pennsylvania was

chosen because it has many smaller, fully functioning township governments. The data are plotted in Figure 8 and show, after certainties are removed, the data remains skewed right. This means most of the non-certainty governments are considered to be small, however, there are a few non-certainty governments that are still large. Table 2 below summarizes the overall frame-by-government type for these two states.

Table 2: Government Organization for Studied States in 2002

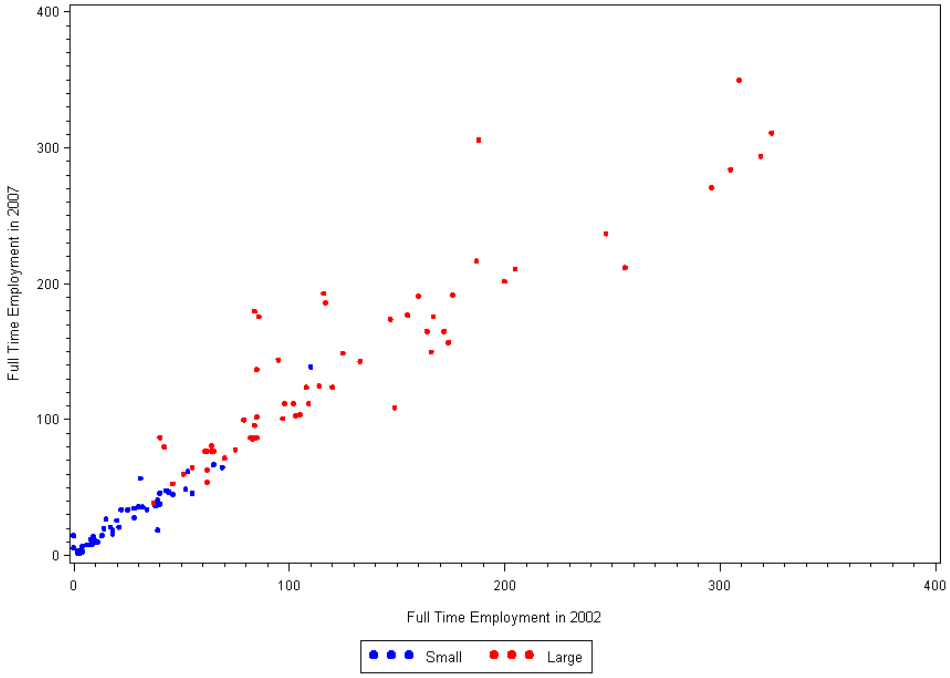
State	Counties	Cities and Townships	Special Districts	School Districts	Total
California	57	478	2,765	1,044	4,344
Pennsylvania	66	2,562	1,728	515	4,871

Source: U.S. Census Bureau, 2002 Census of Government Organization

4.1 California Special Districts

The California special districts are skewed right as a result of about 7 "outlier" data points. These points are large but were not considered to be certainties. It is still possible to find an optimal total mean squared error in Figure 9, but the acceptable range of total mean squared errors is much wider than the simulated data. This is attributed to the fact that the simulated data had a clear cut-off point and our data appear to have the same slope throughout.

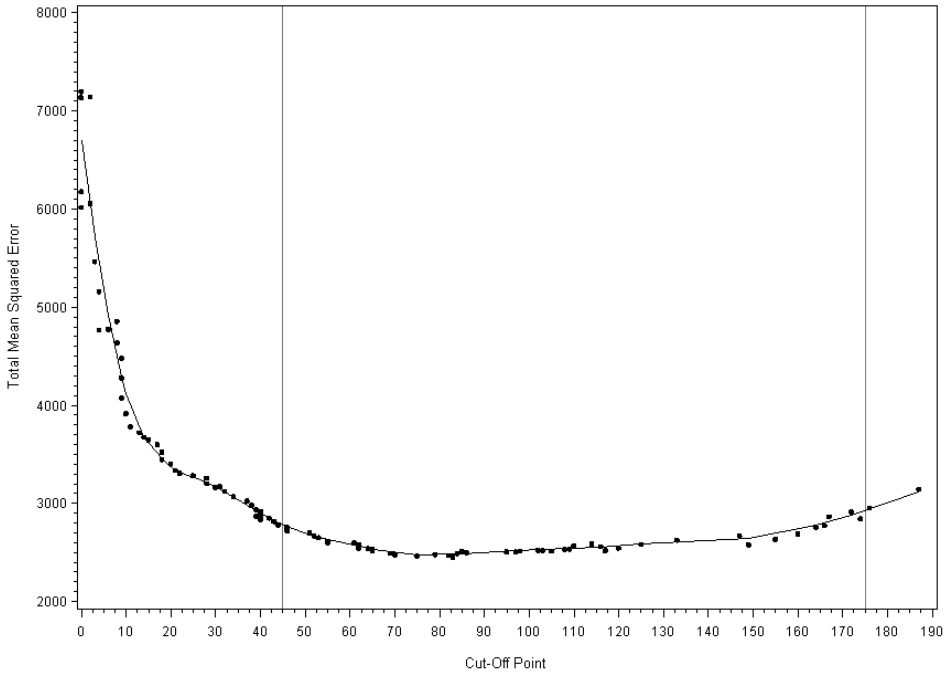
Figure 8: California Special Districts Full-Time Employment Values from 2007 Regressed on 2002



Source: U.S. Census Bureau, 2002 and 2007 Census of Governments: Employment

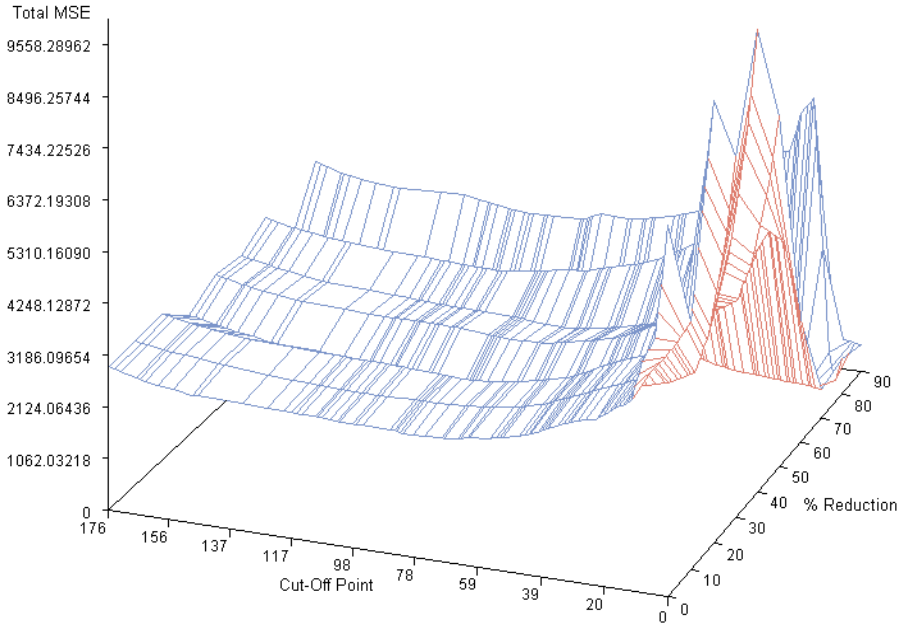
The three-dimensional graph below suggests if the data do not have an obvious cutoff point, it becomes much more difficult to find an optimal total mean squared error. There is no one clear cut-off point to select but it is clear what points you would not want to select. Once a cut-off point is selected it does not matter what the percentage of small strata reduction is, it does not have an effect on the total mean squared error.

Figure 9: California Special Districts Graph of 2002 Full-Time Employment Values as Cut-Off Points Versus Total Mean Squared Error



Source: U.S. Census Bureau, 2002 and 2007 Census of Governments: Employment

Figure 10: All Possible Combinations of Cut-Off Point, Small Stratum Reduction and Resulting Total Mean Squared Error. California Special Districts – Smoothed.

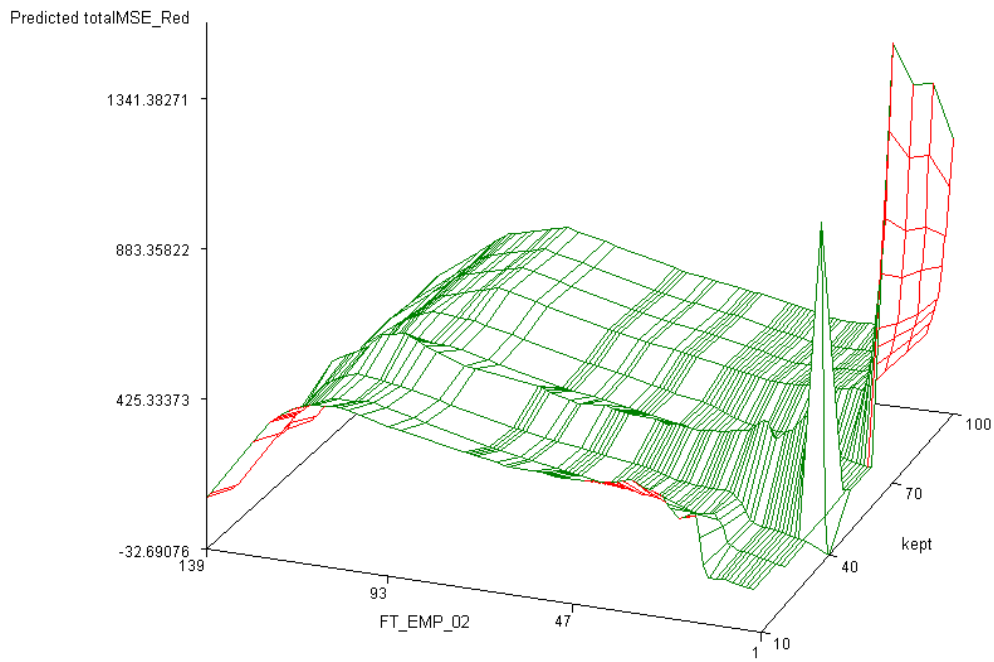


Source: U.S. Census Bureau, 2002 and 2007 Census of Governments: Employment

The results for Pennsylvania cities and townships were similar to that of California special districts. To save space, only the three-dimensional graph is shown as Figure 11.

The graph suggests if no clear cut-point is present the optimal solution is not obvious. Although, it is still clear what points you would not want to choose for cut-off points.

Figure 11: All Possible Combinations of Cut-Off Point, Small Stratum Reduction and Resulting Total Mean Squared Error. Pennsylvania Cities and Townships – Smoothed.



Source: U.S. Census Bureau, 2002 and 2007 Census of Governments: Employment

3. Future Work

In the future we would like to add a penalization variable to the total MSE equation. This will allow us to take into consideration a cost or burden factor a government would experience on account of our survey. Non-response follow-up takes resources from both the Census Bureau and the government unit. If the government unit never responds, we use imputation techniques. If we could minimize nonresponse, we would save resources. The penalization function will be as follows:

$$MSE_{penalized} = MSE_1(c, p) + MSE_2(c) + dn_1^k p^k$$

Where: d=cost

k=½, 1, 2

After including the penalization variable it will be possible to minimize the total mean squared error equation with respect to the cut off point and proportion of small stratum reduction if given a specified cost. If we try to minimize the mean squared error with respect to just the cut-off point and rate of reduction, the algorithm will almost always choose to keep 100 percent of the sample. We are hoping we will gain more insight into how the mean squared error relates to all three variables simultaneously.

Lastly, we would like to examine using other measures of variability in our analyses. The California sample we used is just one sample from the California frame and the

corresponding total MSE is subject to error. It is clear from the figures that using total MSE makes it difficult to identify one optimal cut-off point. By using the variance instead of the total MSE and using re-sampling methods, such as bootstrap, we might be able to find a more optimal range of cut-off points.

4. Conclusions

The simulated data showed that as long as there was a defined cut-off point, an optimal total mean squared error could be found. The mean squared error was controlled much more by the selection of the cut-off point versus the amount of small strata reduction, thus suggesting that we should be focusing on identifying the cut-off point. Total mean squared error was not changed when the smaller stratum was reduced but at this point the weights have not been considered. Adding the weights to the mean squared error should have an effect on small stratum reduction. Future research will look into adding a penalization variable to the total mean squared error function that will account for a cost or burden a government will be faced with when they are selected into the sample.

We would also like to see how changing the rate of small stratum reduction affects the overall estimates of total. Our research has shown that the total mean squared error will stay roughly constant with small stratum reduction but we do not know how the totals will be affected. Future research will address these issues.

Lastly, it would be interesting to research how the size of the government affects the likelihood of response. We assume larger governments are more likely to respond than smaller governments but no official research has been conducted in this area.

Acknowledgements

We would like to thank Lisa Blumerman and Carma Hogue, both of the Governments Division at the U.S. Census Bureau for serving as reviewers as well as Dr. Eric Slud from the University of Maryland and Dr. Jun Shao from the University of Wisconsin-Madison.

References

Barth, J., Cheng, Y., Hogue, C. (2009), Reducing the Public Employment Survey Sample Size, JSM 2009

Cochran, W. (1977). *Sampling Techniques* (3rd ed.). New York, NY: Wiley.

Lohr, S. (2009). *Sampling Design and Analysis* (2nd ed.). Boston, MA: Brooks Cole.

Shao, J., Tu, D. (1995). *The Jackknife and Bootstrap*. New York, NY: Springer-Verlag.