Considerations before Pooling Data from Two Different Cycles of the General Social Survey

Michael Wendt[1]
February 27, 2007

Summary: The General Social Survey Program has been gathering data for some time now in a series of independent annual household surveys. A rich collection of data now exists for social science and other researchers. One interesting potential tool is the integration of data across two or more cycles. This means using the data from two or more cycles to estimate quantities of interest. There are two ways of doing this (which, in general, give different answers). One is to compute separate estimates by cycle and combine them. Another way is by simply pooling the data sets from different cycles together and computing estimates using the pooled data. This document provides a brief description of situations when the second method, pooling, is appropriate as well as some things to take into account for the researcher interested in pooling. Additionally, some basic, practical recipes are provided to aid the researcher in the integrating exercise. Perusing the contents immediately below will provide a broad overview of the considerations and recipes. In essence, the researcher must decide whether or not the two data sets *should* be pooled and *can* be pooled in order to conduct their analysis project.

Contents:

[1] Social and Aboriginal Statistics Division, Statistics Canada, michael.wendt@statcan.ca

Introduction: Statistics Canada's General Social Survey (GSS) is an annual cross-sectional household survey that has been gathering social data on Canadian adults since 1985. The most recent data release, for GSS-20, occurred in June, 2007.

With twenty cycles of complete data collected, a wealth of information exists that can be analyzed in various ways that take advantage of these different cycles. For example, a project is underway to harmonize the variables across the 20 cycles of data. The idea is to create a collection of twenty independent datasets with a consistent file format, in which variable names and formats are the same and variable definitions are comparable. Additionally, weight variable names will be harmonized and bootstrap weights for variance estimation will be provided. Such a collection would, in particular, help researchers to follow characteristics of Canadians over time by computing individual (annual) estimates and observing the time series.

Apart from such large-scale projects, researchers are increasingly comparing information from two or more different cycles for various types of analyses. Two types of situations come immediately to mind. The first is that researchers are using cycles with similar themes (usually repeated every five years) to create *sparser* time series of separate estimates or using a few adjacent cycles to create *shorter* time series. It is important to note that different cycles of the GSS represent different instances of a changing Canadian population. For that reason, in many circumstances, computing individual-cycle based estimates and following a time series or comparing "then" with "now" would be the only plausible thing to do. This type of analysis will not be considered in this paper.

A second situation occurs when researchers wish to directly combine information from two or more cycles to create some sort of *integrated* estimate based on the cycles. The most common reason for researchers to integrate separate cycles is to increase sample sizes of small domains. For two cycles, for example, the idea is that small + small = big enough for a meaningful analysis. If it is big enough, then researchers would normally study characteristics within the integrated domain. Consider, for example, people identifying themselves as "belonging to a visible minority." In the samples for two separate cycles, there may not be many such respondents and so, for confidentiality or high variability reasons, characteristics about the group could not be published separately. However, in combining the data, it is hoped that a large enough sample is obtained so that some analyses could be performed. At issue is whether or not it makes sense to combine. If the population of visible minorities has changed in size between the two years or if it has changed composition between the two years, then what does a combined estimate even mean? Indeed, if a large change has occurred, to what target domain population does the estimate even refer? These are important considerations for researchers. Since, despite its limitations, pooling is now commonly being used by researchers, this document is meant to give advice on what to do before using that particular method.

In this document, our focus is on combining information from separate cycles. Although we caution that there are many instances when integrating should not be done (as will be detailed below), we will assume that the researcher wishes to try to integrate information

from two[2] different cycles to increase the sample size of a small domain. We give a list of considerations for researchers to take into account to decide when a particular method of integrating is appropriate, specific to the GSS context. We give some recipes on how to integrate if it is appropriate, and give some suggestions if it is not. The key term is in the previous paragraph: *meaningful analysis*. It is that aspect that will guide the advice given here about integrating data.

Much literature already exists about integrating different surveys, notably [Binder and Roberts], [Korn and Graubard] and [Thomas]. A series of health analysis related examples is given in [Schenker and Raghunathan]. In the GSS context, an overview of the methodology for a specific case of combining GSS-13 and GSS-18 to establish (crime) victimization rates among visible minorities was provided in [Marchand]. Our document here is not intended to be a survey of the methodology of integrating rather it is hoped to be of practical use to researchers with a specific task in mind. The idea is to provide some simple recipes that will work in many useful cases in the GSS context.

Separate approach versus pooled approach: There are two ways of computing estimates for a combination of two cycles when each target population is assumed to be finite[3] (as is the case for descriptive statistics such as means and proportions, for example). First, one could compute separate estimates for each cycle and combine them afterwards by some sort of weighted average (this is called the *separate approach*). The second way to combine cycles is to pool the data, adjust the survey weights and continue as if the combined sample was simply one larger sample (this is called the *pooled approach*). Under certain circumstances, either approach will lead to an unbiased population estimate but, in general, the separate and pooled approaches will lead to different estimates with potentially different interpretations. For this reason, it is advisable that researchers explore the possibility of the separate approach *first*. The pooled approach should be used *only* if it can be assumed that the characteristics as well as the domains of interest are similar from one cycle to the next.[4] Indeed much of this document provides considerations to check if these assumptions are true for the specific project a researcher has in mind.

There are many things to consider *before* pooling to compute pooled estimates. The considerations form, in some sense, an iterative process: think about why integrating is desired, check that it makes sense, integrate the data, compute estimates or perform analyses, check the results, does it still make sense to integrate?, repeat.

The researcher is advised that, if possible, they should consult a survey methodologist on the subject as each pooling project may involve different methodological issues. What to do first in a pooling project really depends on one's point of view. An analyst might begin with questions like "are my samples the same?" and "are the questionnaires the same so that my variables mean the same thing?" whereas a methodologist might begin

[2] The considerations in combining three or more cycles are similar though more care needs to be taken.
[3] Another situation arises when parameters of a model are to be estimated and this is treated in Consideration 1 below.
[4] Furthermore, if a composite estimate is desired, then the separate approach is the route to take.

with "we would like to estimate the population mean for this given variable; to what population will the estimate refer?" We attempt to combine both points of view and the the list of considerations and recipes is couched in that context.

Consideration 1: What type of estimate is desired? We noted that a *meaningful* estimate or set of estimates is desired. In some sense, the whole point of this document is that the researcher has to decide why they wish to integrate different cycles and whether there is an interpretation of any estimates computed from the combined information.

When integrating data from two or more cycles to increase sample sizes for small domains, researchers are generally interested in doing two things: estimating population means or proportions[5] for characteristics and estimating model parameters like coefficients in a linear or a logistic regression. We shall refer to these, respectively, as *descriptive analysis* and *modelling*. The type of analysis desired determines some methodological aspects like the population of interest but the considerations below could apply to both types.

Often, an analysis project will desire many descriptive estimates as well as many model parameter estimates at the same time. However, the researcher is cautioned that they may have to do different things for different estimates[6]. Furthermore, it is recommended to limit the number of variables and the scope of the project, especially in view of consideration 4 below, in which a detailed attempt at harmonizing the two sets of variables must be made. The harmonization process usually takes the largest amount of time.

Adjacent cycles of the GSS are non-overlapping and not completely statistically independent. To reduce individual respondent burden, telephone numbers selected in one cycle are excluded from future cycles for two years. Thus, the second sample depends on the first. However, the chance of overlap is extremely rare and proceeding as if the two cycles were simply two independent samples from the Canadian population in different years would be acceptable. For cycles several years apart, it is conceivable that one respondent could be in two cycles but, again, this is extremely unlikely.[7] The assumption of non-overlapping data allows for a much easier method of combining two cycles and the assumption of independent data sets allows for a much easier method of computing combined estimates. It should be noted, however, that GSS-20 and GSS-21 will form an exception as some of the respondents of GSS-21 were also (purposely chosen) respondents to GSS-20.

When descriptive estimates are desired, the target population is considered finite. As noted above, the pooled and separate approaches will give different answers. The

[5] The estimation of a domain total based on two cycles doesn't really have a meaningful interpretation.
[6] This as part of the iterative process alluded to above.
[7] The author once performed a test to establish if any telephone numbers were in two samples several years apart and could not find any. The test was not exhaustive but empirically confirmed the extremely low probability of a respondent being in two different cycles.

separate approach should be explored if the characteristics have changed in value between the two cycles or if the domains of interest have change much in size.

If the pooled approach is used, however, then pooled estimates will be formed by first concatenating the two separate data sets, adjusting the weights and computing as if one large sample represented "the" population. More details will be given below.

Another situation arises when the researcher is interested in estimating parameters of a model (the coefficients in a linear regression model, for example). In this situation, the statistical model describes an infinite population and one can assume that the model has generated the values (for the variables in question) of each of the finite populations targeted in the two cycles. When using pooled data to estimate parameters, it is a good idea to add a "cycle effect" into the model. One can then test for inequalities in parameters among the two finite populations assumed to have been generated by the model. An example will clarify:

> Example: The linear model $Y = \beta_0 + \beta_1 X + \varepsilon$ describes a (theoretical) relationship between X and Y and suppose one wishes to estimate β_0 and β_1. We assume, for the first cycle, the model applied to X_i, for individuals in the first population, would generate Y_i (and we can estimate using the first sample). Likewise, for the second cycle, we estimate the parameters for the second population using the second sample. When the data is pooled, we can estimate the same model using the larger sample but should first consider the model $Y = \beta_0 + \beta_1 X + \beta_2 I_{\text{cycle 1}} + \varepsilon$, where I is an indicator for the first cycle, say, and look for *cycle effect*.

The main advantage of the pooled approach is that once a set of suitable weights are found for the pooled sample, they can apply to many different estimates.[8] The disadvantage is that a minimum estimated variance for all estimates may not be achieved.

Consideration 2: What is the target population? When working with any survey data, a researcher must define their target population. Often, details are provided in the survey documentation itself. GSS generally targets Canadians in the ten provinces over the age of 15, not living in institutions, on reserves, or on military bases. When combining or pooling data from two different surveys, the first question that arises is: were the samples supposed to represent the same target populations? In the GSS, this is almost always the case.[9] The next question that a researcher must ask is a *conceptual* one: what target population does the pooled sample represent? We use the term "conceptual" because there is no fixed *statistical* or *methodological* answer. Furthermore, the question cannot really be answered before some of the other considerations below are taken into account. At this point, the researcher is cautioned that if it is found that pooling is not statistically

[8] As noted above, most studies don't just compute one estimate or even one type of estimate. For example, a researcher might be interested in estimating totals, proportions, and regression coefficients *within* the same analysis.

[9] In GSS-16 (and in GSS-21, now in the final stages of collection), the target population consisted of people 45 or over, living in the provinces, etc.

appropriate, then the pooled sample doesn't really represent any concrete population and perhaps some other method or other source of data is warranted.

Notwithstanding such a cautionary note, one could think of the samples as simply two instances of the process of sampling taken at two different time points. Indeed, the whole issue in pooling of different GSS cycles is about whether time matters or not. When combining two adjacent cycles, it seems appropriate to assume that the *real* population of Canadian adults hadn't changed much from one year to the next.[10] However, in terms of the characteristics being studied *or* in terms of unknown factors related to cycles that are five years apart, more care needs to be taken to assure oneself that what is being studied hasn't changed in any critical way. Indeed, as we have mentioned, this *assumption needs to be thoroughly tested* and the details are given in Consideration 5 below.

Consideration 3: Are the samples comparable? Next, the researcher should ask the question: does it make sense to *try* to combine these data sets? For a GSS pooling project, we mean: have the samples been gathered in the same way? Do they really represent the same population in the same way, just at different time points?

The answer for GSS, typically, is that samples for different cycles are gathered in the same way using the same collection methods and the same (or similar) sample design. However, methods have changed incrementally over the years. For example, the researcher is cautioned about the various Labour Force Survey supplementary samples used in some earlier cycles. Another example is that data for GSS-16 was collected in a different manner than for other cycles, which typically use Random Digit Dialing methodology. Additionally, GSS sample sizes went from around 10,000 up to GSS-12 to around 25,000 since GSS-13.

Having noted these caveats, the sampling designs of different cycles do not differ that much in a way that would affect most estimates (indeed, our experience tells us that other factors[11] far outweigh any impact of design on estimates). In particular, the survey weights and bootstrap weights are often fairly comparable in that they represent a respondent in the same way in the respective populations. For some details on weights and sample sizes, the reader is referred to the Appendix. Furthermore, a careful reading of the respective Public Use Microdata File User Guides' methodology section is warranted.

Another question to address is: does pooling make sense from an analytic point of view? This means: were the same concepts globally measured in the same way. For example, the General Social Survey sometimes gathers information on themes from different individuals at five year intervals. Great care is taken to measure comparable themes in similar ways BUT the researcher is cautioned that changes have occurred. Again a careful

[10] In fact, in recent GSS cycles, sampling has been performed in monthly waves over a most of the year so there may be as little as a few weeks between the end of collection of one cycle and the beginning of collection for the next. Indeed, the (survey) weighting process assumes that the GSS is collected as 12 (or so) independent monthly surveys.

[11] For example, consideration 4 below looks at how concepts are measured by variables.

reading of the two PUMF User Guides is important; in this case, details about the concepts should be taken into account.

A third question for global comparability of data sets is: does pooling make sense from a practical point of view? Will there be enough comparable variables in common to do the desired analysis? Pooling is often done to increase sample sizes for small domains. This can work for basic analysis but running complex multivariate models may defeat the purpose of pooling because the more parameters to be estimated, the more degrees of freedom are needed.

The analyst should use their common sense and their experience for Consideration 3. As we have noted, the questions raised here may form an iterative process with the following steps.

Consideration 4: Are the two sets of variables similar? Possibly the most time-consuming task in any data pooling project is the harmonization of variables. Indeed, all variables to be used directly or indirectly[12] in the analysis need to be checked to determine if they measure the same thing and were measured in the same way. This should be done variable by variable.

The analyst should first make sure that the names and formats on both files are the same. If not, these need be reconciled. A good place to begin is the respective codebooks. Each categorical variable needs to have the same categories or ensure that both sets can be collapsed to similar categories. The analyst is particularly cautioned about "not stated" and "don't know" categories as these have sometimes changed over the years. Likewise, for each continuous variable, the analysts should pay particular attention to codes outside the normal range of values that mean things like "not stated", "don't know," or even "10 or more."

Sometimes, there were subtle changes in the way a question was asked. This should be explored to make sure that the concept being measured is comparable across the two data sets. This can get quite involved. For example, the definition of what constitutes a violent crime has changed over the years to reflect policy changes. An analyst studying violent crime would have to make sure that the definitions were appropriate for their work.

The GSS questionnaire is quite complex (though that is transparent to the respondent as data are now gathered via a Computer Aided Telephone Interview application). From time to time, it is necessary to change the flow of the questionnaire or the positioning of the question within the questionnaire. This may affect which respondent is asked which question. In the codebook for each cycle of GGS, at the bottom of each variable is a brief description of which respondents were asked the particular question.

Finally, analysts should be aware that the theme of a particular cycle may affect the way in which respondents answer. For example, if the theme is "health" and a respondent has just been asked a number of detailed questions about health and is thinking about their

[12] This means the weight variables, for example.

health, they may answer "How would describe your general health?" differently than if it was a "time use" cycle.

For some cycles of the same theme, a detailed concordance table of variables between the cycles exists.

Once variables have been harmonized, a good *quality control* check is to compute a cross-tabulation of the variable by cycle, for categorical variables, or a five number summary[13] by cycle, for continuous variables. Similar category counts between cycles do not guarantee similarity of the two variables but dissimilar category counts can flag possible problems and should be investigated.

The discussion above may be summarized as:

> **Recipe 1: A checklist for variable harmonization:** The following is a checklist of things that need to be checked or modified to make the two variables comparable (or to gage if some factor in collection may affect the analysis):
>
> -are the names the same?
> -are the formats the same?
> -are the categories the same or collapsible to similar categories?
> -are the "not applicable", "not stated" and "don't know" values the same?
> -are the two questions the same?
> -are the two questionnaire flows the same?
> -is the question positioning the same or similar?
> -are there other considerations like the type of theme?
> -perform a (weighted and un-weighted) cross-tabulation or five-number summary by cycle as a final check

Consideration 5: Are the two sets of estimates similar? In order to combine variables in a pooled dataset, the variables must be similar, as in consideration 4 above. BUT, in order to compute a meaningful combined estimate, the *estimates* should be similar between the two cycles.

The weighted and un-weighted frequency tables or five number summaries at the end of recipe 1 above can serve also as a quick check to see if the estimated frequencies (means, etc.) are in the same "ball park."

In addition, the researcher should perform a formal hypothesis test using each pair[14] of estimates to see if the two *population* parameters of interest are statistically significantly different or not. At this point, it is necessary to actually pool the data.

---

[13] In fact, one could compare mean, and standard deviations as well as the five number summary: minimum, first quartile, median, third quartile, maximum.

[14] Strictly speaking, one could also compare the bivariate distribution of two variables in the first cycle with the bivariate distribution of the corresponding two variables in the second cycle (and, more generally,

Recipe 2: How to combine the data? Pooling data from two different GSS cycles is straightforward, once comparable variables have been designated. The pooled data set is simply the two concatenated, one *on top* of the other. For example, in SAS code, this could be

```
data pooled_data;
        set data_first_year(in = ina) data_second_year(in = inb);
        indicator_first = 0;
        indicator_second = 0;
        if ina then indicator_first = 1;
        if inb then indicator_second = 1;

run;
```

It is a good idea to create two indicator variables, one for the source of each data set. When weights are multiplied by the respective indicator, the individual estimates can be obtained.

We assume that variable names are exactly the same. SAS will not complain if the variable names are different from one cycle to another. It will simply create two variables with missing values in the opposite parts. SAS will, however, complain if the variables have different formats between the two cycles.

The two input data sets should contain all the variables of interest and the weights, including the bootstrap weights for both. Weights are generally comparable (be careful of person-level as opposed to incident-level or household-level weights; the correct weight depends on the type of analysis) as are bootstrap weights so they often only need re-naming. However, a careful reading of the weighting process in the PUMF documentation is required (in particular, to see which weights go with what concept).

Once the data have been pooled, hypothesis tests can be performed to see if the respective population values are significantly different or not. This requires variance estimation, which may be done via the bootstrap weights. Bootvar, Stata, and SUDAAN all are able to compute meaningful variance estimates using bootstrap weights for many types of estimates. There are many ways of doing pair-wise tests in SUDAAN. The following recipe provides one possible method. For large projects, the researcher is urged to optimize their code as run-time might be long.

Recipe 3: How to compare estimates? For the purposes of illustration, variable_1 is assumed to be categorical with three categories, value_1, value_2, and value_3. The person-level weight is assumed to be wght_per and the bootstrap weights are assumed to be wtbs_001 to wtbs_200. We will test whether

---

multivariate distributions). At some point, however, the researcher must decide what is practical in an analysis in view of smaller and smaller cell-sizes for multi-dimensional cross-tabulations.

or not value_1 is (significantly) different based on the first data set from that based on the second data set.

```
/* create an indicator variable for values of variable_1 = value_1 */
/* create a variable called one, which is constantly = 1, for denominator */
/* of ratio estimates */

data pooled_data;
        set pooled_data;
        indicator_value_1 = 0;
        if variable_1 = value_1 then indicator_value_1 = 1;

/* this computes two ratio estimates value_1 in data set 1 / one and */
/* value_1 in data set 2 / one */

proc ratio data = pooled_data design = brr;
        class indicator_first;
        weight wght_per;
        denom one;
        numer indicator_value_1;
        repwgt wtbs_001 - wtbs_200 / adjfay = 25;

run;

/* this produces a test statistic for */
/*  H0: ratio in population 1 = ratio in population 2 */

proc ratio data = pooled_data design = brr;
        class indicator_first;
        weight wght_per;
        denom one;
        numer indicator_value_1;
        repwgt wtbs_001 - wtbs_200 / adjfay = 25;
        contrast indicator_first = (-1 1);

run;
```

Proc ratio works well with categorical (binary) variables. To test a numeric variable using SUDAAN, one can use the descript procedure (be careful to remove any “not stated” and don’t know” values).

After each test, the researcher must decide what impact this will have on the pooling project. If the two respective population parameters are not significantly different, based on the estimates, then a useful pooled estimate can be computed. If, however, the test fails, then the researcher has two options: remove the estimate from consideration in the pooled data analysis project or use the variable (and value) but with caution (by adding a

caveat to the final documentation, for example). The latter choice depends upon the p-value of the test statistic and the subject matter meaningfulness of a failed test.

Consideration 6: Pooling, bootstrap weighting, and variance estimation As we noted above, there are several ways of computing estimates from pooled data. The most widely applicable method is to simply adjust the weights and compute as if one had only one sample.[15] How to adjust the weights depends upon the desired final population. As notation, let w_{1i} = the weight for the i^{th} record in the first sample and w_{2j} = the weight for the j^{th} record in the second sample. Suppose further that n_1 and n_2 are the respective sample sizes and that $N_1 = \sum w_{1i}$ and $N_2 = \sum w_{2j}$ are the respective estimated population sizes.

To get the average of N_1 and N_2, perform the adjustment:

$$w'_{1i} = w_{1i} \times 1/2 \text{ and } w'_{2j} = w_{2j} \times 1/2.$$

To get N_2 as the total *pooled population* (here, the paradigm is that the first cycle was simply an earlier collection from the second population), perform the adjustment:

$$w'_{1i} = w_{1i} \times N_2/(N_1 + N_2) \text{ and } w'_{2j} = w_{2j} \times N_2/(N_1 + N_2).$$

In general, multiplying the first weights by α and the second weights by β yields a total estimated population of $\alpha \times N_1 + \beta \times N_2$. In theory, one could make a weight adjustment for each[16] pooled estimate to be computed and for each domain, or subpopulation, of interest. For that matter, differing values of item non-response can be treated in this way by using different adjustments. In practice, however, it is probably good to choose one adjustment and use that for all the analyses. The two above seem to be the most appropriate for a wide range of types of analyses.

Whatever adjustment is chosen, it must be performed on the final weights *and* each of the sets of bootstrap weights.

> **Recipe 4: How to compute a pooled estimate and its associated variance?** Using the first paradigm above (averaging the two populations), the researcher basically only needs to compute new weights. We assume variables as in recipe 3 above. In SUDAAN, we could use:
>
> /* create new weights that are half the old ones */

[15] Recall, we assume that two different cycles are independent.

[16] Indeed, simply averaging the weights does not necessarily produce the most efficient estimates in terms of variance. Inasmuch as most studies contain many variables, more efficient pooling methods for one variable may not work for another. The most general compromise is as suggested in the text.

```
data pooled_data;
        set pooled_data;
        wght_per_new = wght_per / 2;

        array w_old wtbs_001 – wtbs_200;
        array w_new wtbs_new_001 – wtbs_new_200;

        do i = 1 to 200;
                w_new(i) = w_old(i) / 2;
        end;

run;

/* compute ratio value_1 / one and its associated variance for the */
/* whole data set */

proc ratio data = pooled_data design = brr;
        weight wght_per_new;
        denom one;
        numer indicator_value_1;
        repwgt wtbs__new_001 - wtbs__new_200 / adjfay = 25;

run;
```

Consideration 7: Stating conclusions and adding caveats When the statistical work is done, the researcher must decide on the validity and applicability of the results. In this section, we provide a list of possible caveats. One or more of these items may be added to the text of any research project or may be used as a sort of "checklist" for the final conclusions.

The first comment is a general one: for statistical tests about population parameters, when p-values are near 0.05, the typical critical value, some caution must be taken. That is, there are many factors in a pooling project that could cause an increase or decrease in the true variance of an estimator, which has been estimated by the pooled variance, say. It is the author's belief that most of the assumptions made above would have little incremental effect on variances but it is often difficult or even impossible to predict the impact of any group of assumptions on the p-value of a test. In short, the researcher is cautioned that p-values near 0.05 do not show "strong evidence" for rejecting or failing to reject.

The list of caveats more or less follows the text of the document:

Recipe 5: A list of caveats to choose from: The following is a basic list:

-usual statistical caveats and distribution assumptions for models: this is independent of the use of the pooling method

-usual GSS caveats: only adults 15+ are interviewed, living in provinces, not living in institutions, random digit dialing survey excludes persons without telephones, etc.

-two cycles one year apart[17]: "although the surveys were taken in two adjacent years, we can consider the two samples as representing the same population of Canadians, as not much change would have been observed"

-two cycles more than one year apart[18]: "although the two surveys were taken x years apart, and some changes in the Canadian population would have occurred between the two time points, extensive testing of the estimates we used in our work showed that, with regard to our analyses, the Canadian population was stable"

-non-overlapping, adjacent cycles: "to reduce individual response burden, if a respondent is one cycle, they are excluded from the next; thus, the samples are non-overlapping so we pooled the data from the two cycles by simply concatenating the two data sets"

-non-overlapping, non-adjacent cycles: "the chance of one respondent being in two different cycles of the GSS is extremely rare so we pooled the data from the two cycles by simply concatenating the two data sets and consider them as one larger data set"

-slightly dependent adjacent cycles (unnecessary for non-adjacent cycles): "to reduce individual response burden, if a respondent is one cycle, they are excluded from the next; thus, the samples are not exactly statistically independent; however, the chance of overlap is extremely rare so that impact of this is minimal so we may assume that the two data sets were drawn independently from the same population"

-same methodology, variables, estimates: it is good to allude to the extensive research and testing needed to ensure that the data sets were comparable, the variables were pair-wise the same or could be made comparable, and that the estimates were pair-wise not statistically significantly different; for those issues that were found to be different, detailed caveats can be added that suggest something like: "the impact of x difference on y would not be great"

[17] In this case, simply averaging the weights would probably work.

[18] In this case, a choice of α and β as in Consideration 6 that placed more emphasis on the newer sample could be more appropriate. If the researcher has time, different choices could be explored BUT the choice should be made *a priori* and not to fit any desired set of conclusions. That is, any choice of α and β would itself form a caveat.

-small samples sizes for various domains: finally, it should be noted that pooling may not be a panacea; the sum of two small domains may still be small; if this is the case, the usual caveats about small sample sizes (such as increased variances, confidentiality concerns, etc.) need to be added

Conclusions: We have provided an overview of some things to consider when integrating data from two or more cycles of the General Social Survey. In addition, some "recipes" are provided on how to actually carry out such a project. Pooling data can be a very fruitful exercise for researchers interested in following social trends over time or in increasing the size of small data sets and is entirely do-able in the context of the GSS, even though it is an annual cross-sectional survey.

Appendix: Some Weight Information for Various Cycles

The following table gives a brief overview of the various cycles and weighting considerations (this table will also be included in an overview document describing how to work with GSS weights to appear in Statistics Canada's Research Data Centres Technical Bulletin).

Cycle	File	Main weight variable(s)	Current Bootstrap Names	Mean	Comments
1	main	wght	N/A	N/A	wght = 10000 x weight; post weighting bootstrap project underway; 200 weights expected soon
2	main	fwght_os	N/A	N/A	post weighting bootstrap project underway; 200 weights expected soon
2	summary	fwght_ms		N/A	This is the time use summary file and contains one record per respondent
2	episode	fwght_ms	N/A	N/A	This contains the one record per time use episode; for how to use weights, refer to the PUMF User's Guide
3	main	weight32, weight33, weight34	N/A	N/A	weight32 = 10000 x weight, etc.; flag32 indicates person-level information; accident episodes may be analyzed separately using flag33, crime incidents may be analyzed separately using flag34; the weights correspond to these three types of records; post weighting bootstrap project underway; 200 weights expected soon;
4	main	pweight	N/A	N/A	pweight = 10000 x weight; post weighting bootstrap project underway; 200 weights expected soon
5	main	pweight	N/A	N/A	pweight = 10000 x weight; post weighting bootstrap project underway; 200 weights expected soon
6	main	finalwt	N/A	N/A	post weighting bootstrap project underway; 200 weights expected soon
7	main	fwght	N/A	N/A	post weighting bootstrap project underway; 200 weights expected soon
7	summary	timewgt	N/A	N/A	This is the time use summary file and contains one record per respondent
7	episode	timewgt	N/A	N/A	This contains the one record per time use episode; for how to use weights, refer to the PUMF User's Guide
8	main	wght_per	wpebs_001 - wpebs_200	25	

Cycle	File	Main weight variable(s)	Current Bootstrap Names	Mean	Comments
9	main	perwght	N/A	N/A	post weighting bootstrap project underway; 200 weights expected soon
10	main	wghtfnl	bsw1 - bsw200	25	
10	child			N/A	no child-level weight
10	union			N/A	no union-level weight
11	main	wght_fnl	wfin_001 - wfin_200	25	
12	main	wghtfin	wfin_001 - wfin_200	25	
12	episode			N/A	
13	main	wght_per	wpebs001 - wpebs200	25	
13	incident		wvcbs001 - wvcbs200	25	
14	main	wght_per	wfin_001 - wfin_200	25	
15	main	wght_per	wtbs_001 - wtbs_200	25	
15	child			N/A	no child-level weight
15	union			N/A	no union-level weight
16	main	wght_per	wtbs_001 - wtbs_200	25	
16	care receiving			N/A	only contains people 65+ who received care; no episode weight given
16	care giving 45-64			N/A	only contains people 45-64 who provided care; no episode weight given
16	care giving 65+			N/A	only contains people 65+ who provided care; no episode weight given
17	main	wght_per	wtbs_001 - wtbs_200	25	
18	main	wght_per	wtbs_001 - wtbs_200	25	
18	incident	adjwtvic, wght_vic	wvcbs001 - wvcbs200	25	see User's Guide for difference between two final weights; bootstrap weights correspond to wght_vic
19	main	wght_per	wtbs_001 - wtbs_500	25	
19	csp	wght_csp	wtcbs_001 - wtcbs_500	25	applicable to questions from section 10a
19	snt	wght_snt	wtsbs_001 - wtsbs_500	25	applicable to questions from section 10b, 11
19	episode	wght_epi	wtbs_epi_001 - wtbs_epi_500	25	
20	main	wght_per	wtbs_001 - wtbs_500	25	

Bibliography:

Binder, D., Roberts, G. (2007) *Approaches for Analyzing Survey Data: a Discussion*, preprint, Statistics Canada.

Korn, E. L., Graubard, B.I. (1998), *Analysis of Health Surveys*, Wiley.

Marchand, I. (2007) *Combinaison des cycle 13 (1999) et cycle 18 (2004) de l'Enquête sociale générale pour dériver un profil de victimisation*, internal document, Statistics Canada.

Schenker, N., Raghunathan, T. (2007) *Combining information from multiple surveys to enhance estimation of measures of health*, Statistics in Medicine 2007; 26:1802-1811

Thomas, S. (2006) *Combining Cycles of the Canadian Community Health Survey*, Proceedings of the Statistics Canada Symposium, 2006.