Generating aggregate statistics from National Survey for Wales data

A report to the Welsh Government

Susan Purdon

April 2012

Contact Details:
Susan Purdon
Bryson Purdon Social Research
10 Etherow Street
London SE22 0JY
+447966 286568
susan.purdon@bpsr.co.uk

Contents

1 Summary

The National Survey for Wales has been designed so as to give an annual sample size per Local Authority (LA) of around 600. It is designed in this way so that statistics from the survey can be generated at LA and Local Health Board level. The intention is that, to improve precision, most local level statistics will not be generated on single years of data. Instead, two or more years of data will be aggregated and combined-year estimates published.

The issue of aggregation is more complex that it may appear to be on the surface. There are a number of decisions to be made. These include:

- An assessment of the types of statistics that are suitable for aggregation, with inherently unstable statistics and those on a very steep upward or downward trajectory *perhaps* being unsuitable;
- A decision on the number of years to aggregate. This is a trade-off between a number of factors:
  - Precision: more years of aggregation gives greater precision
  - Representativeness: with more years of aggregation the aggregate statistics can become very poor representatives of the underlying single years
  - Timeliness: given the large sample size of the National Survey it is likely to be seen as unreasonable for many years to pass before local-level statistics are produced.
- How frequently to produce the aggregate statistics.
  - The different departments of the UK govt have made decisions around this on what appears to be a survey-by-survey basis. For example, Welsh Health Survey results are produced every year; Scottish Household Survey results are produced every two years.

We have not aimed in this paper to give firm recommendations. Rather we have tried to identify and articulate the decisions to be made and the options available, and to illustrate these using other surveys where appropriate. There appear to be no published guidelines on aggregation and there is relatively little consistency in how survey researchers use and publish aggregate statistics. It seems that the National Survey team will need to take their own decisions on how to proceed, with no accepted 'best practice' to guide them.

There are, however, a number of conclusions that we have tentatively reached in the paper:

- Two, or at most three, years of aggregation would seem to be appropriate. The sample size per year is large and there is no case to be made for aggregating more years for most local level statistics[1]. The sample size would, in our view, support two year aggregation.
- Most National Survey statistics would be suitable for two-year aggregation, even very fast-moving statistics, such as internet use. But the publication of statistics such as these may need accompanying text so as to aid their interpretation.

[1] The possible exception being some sub-group statistics.

On other aspects of aggregation we have not been able to reach a firm conclusion:

- Whether to publish aggregate data every two years or annually (using moving averages). Annual reporting with moving averages provides more timely data at local level. But moving averages can be harder to interpret and work with. And there producing tables every year will be more resource-intensive.

The paper includes consideration of some of the more technical aspects of aggregation, such as whether to use aggregation weights or not, and statistical testing with aggregate data. These sections (sections 6, 7, 10 and 11) can easily be skipped by those who dislike statistical theory, although they should be comprehensible to anyone with some basic statistics training and a familiarity with survey data.

2 Introduction

2.1 Generating combined-year estimates from survey data

This report aims to give an understanding of, and practical advice on, the use of aggregate (combined-year) statistics for surveys. The focus is on issues for the National Survey for Wales (National Survey), but we hope that the report will prove useful to survey researchers more generally. Most of the topics that are covered apply to many population surveys.

By 'aggregate statistics' here we mean any statistics that are generated by combining two or more years (or other time periods) of survey data. We do not cover aggregate statistics that are generated by combining two or more different surveys from the same time period. Combining different, but parallel-in-time, surveys raises a different set of technical issues, such as how to ensure the surveys are comparable in terms of their coverage of the population, and what weight to give to each survey. Although there are considerable overlaps with the issues discussed in this report, there are some crucial differences.

The aggregate statistics we discuss in this report are those that are based on continuous, or repeated, surveys where the same survey is run over a number of years. There are essentially two types of aggregate statistics that we will discuss:

- Simple, **combined year, statistics** that are generated periodically once sufficient data has accumulated; and
- **Moving, or rolling averages**, where a time series of aggregated statistics is built up over a number of years.

The table below shows six years of data from ONS's General Lifestyle Survey on smoking habits. The figures shown are the percentage of men who are current smokers, per year, from 2003 to 2008.

Year	2003	2004	2005	2006	2007	2008
Percentage current smokers	28	26	25	23	22	22

A simple, combined year, statistic based on three years of aggregate data would generate two statistics over this six year period. Taking 2003 as the start of the series, the statistics would be:

Year	2003-05	2006-08
Percentage current smokers	26	22

The 26% in the 2003-05 column is calculated as the average of the percentages for 2003, 2004, and 2005. That is (28+26+25)/3. And the 22% in the 2006-08 column is, likewise, the average of the percentages for 2006, 2007 and 2008[2].

A three-year moving average would, in contrast, generate four statistics over the six year period:

Year	2003-05	2004-06	2005-07	2006-08
Percentage current smokers	26	25	23	22

With a 3-year moving average, the 2003-05 statistic is the simple, combined-year, statistic for 2003, 2004, and 2005; the 2004-06 statistic is the simple, combined-year statistic for 2004, 2005, and 2006; and the 2005-07 statistic is the simple, combined-year statistic for 2005, 2006 and 2007. Finally, the 2006-08 statistic is the simple combined-year statistic for 2006, 2007 and 2008.

Tracking how the 2005 percentage is used in these moving averages, it is used three times: for the 2003-05, 2004-6, and 2005-7 figures. Were a five-year moving average to be adopted it would feature five times.

It will be seen from the description above that the differences between the two types of aggregate statistics (combined years, and moving averages) are somewhat artificial. The moving average statistic is, in essence just a series of simple, combined year, estimates. Nevertheless, the two tend to be used in different circumstances, and their uses, presentation, and interpretation are somewhat different. These two types of statistic are covered in some depth in Section 9.

2.2 The National Survey for Wales

The primary purpose of this report is to give guidance for when, if, and how aggregate statistics can be computed from the National Survey for Wales. The focus is on aggregate statistics for Local Authorities (LAs) and Local Health Boards (LHBs) although aggregation for small sub-groups is touched on.

The National Survey, launched in January 2012, is a large, annual, general population survey of adults aged 16 and over in Wales. The achieved sample size each year is expected to be around 14,500 (around 600 per Local Authority[3]). The questionnaire covers a range of topics, including perception of public services, and wellbeing.

Although the survey is large, there is a need for statistics at both local authority and health board level. The target sample sizes at these levels per year are inevitably much smaller,

[2] In practice there are alternative methods of calculating simple combined year estimates – these are discussed in Section 7.
[3] There are 22 LAs in Wales.

being set at just over 600 per year for LAs, and with the sample sizes per Local Health Board (LHB)[4] varying across LHBs from 600 to 3,600.

The key questions that we cover in this report are:

- What type of statistics from the National Survey are suitable for aggregation? Are there some sets of statistics for which aggregation isn't suitable?
- How many years of data should be aggregated?
- Should the data be presented periodically as simple, combined-year estimates, or annually using moving averages?
- What help and guidance should be given to data users on the use and interpretation of the aggregate statistics?

2.3 The rationale for a paper on aggregate statistics

Aggregate statistics are used across a large number of the UK's general population surveys. However there does not appear to be any written guidance for researchers on the issues around aggregation and when it is appropriate. In our review of other surveys (see Section 4) we did not find any examples of researchers documenting their decisions around aggregation, especially in terms of the number of years being aggregated.

Given what we believe is fairly ad hoc decision making across researchers, we believe this paper may prove useful to researchers outside of the Welsh Government in assessing the case for aggregation on their own surveys.

2.4 Outline of the paper

The paper divides from here into 10 sections:

- We begin (Section 3) with a summary of the challenges of aggregation;
- Section 4 describes how aggregation has been used in other large-scale population surveys in the UK
- Section 5 unpicks the implications of aggregation for different types of statistics
- Section 6 covers the gains in precision that aggregation brings;
- Section 7 looks at the options for how aggregated statistics are calculated;
- Section 8 addresses the issues around deciding on the number of years to aggregate;
- Section 9 then looks at the options around publishing the results;
- Section 10 considers aggregation under non-standard survey situations, in particular under longitudinal or panel designs;
- Section 11 covers the statistical testing of differences between two aggregate statistics;

[4] The seven LHBs are aggregates of LAs. The smallest LHB covers just one LA (Powys Teaching LHB) and has a target sample size of 600;, the largest LHB is Betsi Cadwaladr University LHB which covers six LAs, and has a target sample size of 3,600.

- And finally, Section 12 summarises the decision process for individual statistics.

3 The challenge of aggregation for small sub-groups

For continuous or repeating surveys aggregation is an extremely useful tool. It has two main purposes:

- To increase sample size. Firstly, and probably foremostly, in surveys that generate official statistics, aggregation gives a larger sample size and allows for statistics about small sub-groups to be published. These small sub-groups are often geographical areas, but they can be other groups, such as children or other particular age-groups where the sample size each year is fairly small.
- To reveal underlying trends. For surveys that report long time trends of data, aggregation (using moving averages) is a means of smoothing out 'noisy' data so that the underlying trends are more visible. This is an analytic use of aggregation and much less commonly used in reports of official statistics, which tend to be descriptive rather than analytic. Moving averages are however, sometimes used in line charts of data showing trends[5].

In this report we concentrate primarily on the first use of aggregation – that is aggregation to increase sample size. We mention the use of aggregation in time series analysis, but it is not a focus of the report. Furthermore, by 'aggregation' we mean aggregation across two or more years of data rather than shorter intervals, such as aggregation across months or quarters[6].

The questions that a researcher who wishes to generate aggregate data for a particular statistic will need to address are:

1. Is the statistic of interest suitable for aggregation? What would the aggregated statistic 'mean'? How would, and should, it be interpreted?
2. What gains in precision can be achieved? And at what cost in terms of misrepresentation of the underlying annual data?
3. How many years of data should be aggregated? What are the implications of less or more years?
4. How often should the aggregate statistics be produced?
5. What information should be added to the published statistics to help minimise any risk of misinterpretation or incorrect use of the figures?

A discussion of each of these questions is included in the sections that follow. But we start with a summary of the issues underlying each of the questions here.

[5] The Health Survey for England reports present time trend data in tables as single year statistics, but associated line charts are often based on three-year moving averages.
[6] Most continuous surveys, including the National Survey, are designed to give population-representative data every quarter. It is standard practice to aggregate quarters into years.

1. Is the statistic of interest suitable for aggregation? What would the aggregated statistic 'mean'? How would, and should, it be interpreted?

The 'standard' conditions under which aggregating across years for sub-group estimates is usually considered acceptable are:

- That there are no *very* marked, monotonic, time-trends in the data over a small number of years; and
- The underlying statistics are inherently stable (even if on a trend) and do not show big swings year on year.

In Section 5 of the report we devote space to the issue of how to assess whether an underlying time trend allows for sensible and meaningful aggregate statistics. Examples of aggregate statistics derived against different underlying time trends are included in that section.

2. What gains in precision can be achieved? And at what cost in terms of misrepresentation of the individual years?

The primary reason for aggregation is to generate a larger sample size – which means gains in precision. In section 6 we discuss what those gains are, and how to calculate them.

The gains in precision may however come at the cost of a loss of survey 'representation'. That is, after aggregation, the aggregated statistic may not give meaningful or useful information about any of the underlying years. We discuss how 'representation' of the underlying years might be assessed and a judgement made on how to proceed.

3. How many years of data should be aggregated? What are the implications of less or more years?

The number of years to aggregate is partly a technical issue and partly a practical one. The practical issue is that there may be pressure on survey teams to generate aggregate statistics as 'soon as possible' after the launch of a new survey.

The technical issues are the trade-offs to be made between gains in precision and the lack of a sensible interpretation if the aggregate statistics are generated against the background of a strong time trend.

The issue of the number of years is dealt with in Section 8.

4. How often should the aggregate statistics be produced?

For a continuous survey, once the first set of aggregate statistics have been produced then these can, in principle, be updated at very regular intervals. This would be achieved by generating 'moving averages'.

The alternative is to report only at intervals of a few years.

We discuss the implications of moving averages versus periodic reporting in Section 9.

5. What information should be added to the published statistics to help minimise any risk of misinterpretation or incorrect use of the figures?

Aggregate statistics are not as easy for users to handle as ordinary point-in-time statistics. Section 9 covers some of the issues. And in Section 11 we include a discussion of some of the statistical testing implications of having overlap years in consecutive statistics.

4 Examples of aggregation in other large population surveys in the UK

Many continuous surveys in the UK use aggregation for some of their regularly reported statistics, although, notably, some never do. DCLG's Citizenship Survey falls into this latter category[7].

For surveys that do use aggregation, the number of years aggregated is almost always either two or three. Only in one survey, the Department of Health's Public Perceptions of the NHS Tracking Study, have we found an example of more years being aggregated. In the regular reporting of that survey, a regional analysis of satisfaction levels with the NHS uses data aggregated from 2002 to the most current date: a time period that is now close to 10 years. The same report series looks at satisfaction with social care services by gender, age and social grade, using data aggregated since 2006.

Examples of aggregation used in other surveys include:

- The Scottish Household Survey, which collects data on the composition, characteristics and behaviours of Scottish households, generally uses two years of aggregate data to generate local authority level statistics[8]. The LA-level tables are published once every two years in line with this.
- The Welsh Health Survey produces tables every year at Local Authority and Health Board level that are based on two years of aggregate data. In 2010 results based on combined 2008 and 2009 data were published, and in 2009 results based on combined 2007 and 2008 data were published. This is an example of reporting using two-year moving averages.
- The National Travel Survey takes a similar moving-average approach for its regional analysis. Regional level statistics in the annual reports are based on two years of survey data. These are published every year, with the 2010 survey report combining 2010 and 2009 data, and the 2009 report combining 2009 and 2008 data.
- A similar approach is taken in the reporting of the Living Costs and Food Survey, which, each year, reports household expenditure by region based on three years of aggregate data. In this instance a three-year moving average is being used.

Other examples include:

- The Family Resources Survey, the reports of which present most statistics on just one year of data, but tables split by ethic group of the household head use three years of combined data.
- The Scottish Health Survey produces irregular themed reports for sub-groups (for example, older people or particular areas) that are based on two or three years of aggregate data.

[7] An email exchange with the survey lead at DCLG suggested that aggregation had never been considered but that there did not seem to be any obvious objections to it.
[8] For larger LAs, with larger sample sizes, information is available annually

Other surveys incorporate aggregation into their design: both the ONS' Wealth and Assets longitudinal survey and the ESRC's Understanding Society longitudinal survey ran their Wave 1 survey fieldwork over a two year period, with the intention of reporting all statistics on two years of combined data.

It is possible to generate small area/group statistics from surveys using methods other than aggregation. In 2007 the Health Survey for England team used survey data from 2003 to 2005 to generate local-area statistics for the 352 Local Authorities in England, and the 6,781 Middle Super-Output Areas. However, these were not aggregate statistics: they were, instead, modelled, or synthetic, estimates of the *expected* local statistics based on the socio-demographic profile of the area, and the known national relationship between socio-demographics and health. The reason for using model-based statistics in this instance was because, even with aggregation, the sample size per local area was still very small (less than 100 for many local authorities). Model-based estimation is however outside of the scope of this report.

The examples we have highlighted above reveal widespread, but varied, use of aggregation, with some survey reports using two years of aggregated data and some three; and with some surveys reporting periodically once sufficient data for aggregation has built up and others reporting every year using moving averages (although they are not described as such in the reports). What we did not find, however, was an explanation in any of the survey reports on why a particular approach was taken. In some reports aggregation is used without any mention in the narrative text: the table titles and footnotes state that aggregated data is used, but the associated text makes no mention of it. But in most reports it is simply stated that aggregation is used to increase sample sizes for key sub-groups. We have not found any discussion in any report of the decision on the number of years to aggregate.

4.1 The Welsh Health Survey and Scottish Household Survey contrasted

The local area estimates generated from the Welsh Health Survey and the Scottish Household Survey provide a very interesting comparison. Both provide models that the National Survey could potentially follow.

The Welsh Health Survey (WHS)

The Welsh Health Survey has a very similar sample size per year per LA to the National Survey, with around 600 interviews per LA per year. Results for individual LAs and LHBs are generated every year, but based on two years of aggregated data. This, as noted above, is an example of the use of moving averages.

The data is published online in downloadable spreadsheets every year, with just a single two-year period being covered per spreadsheet. No time series are included in the spreadsheets.

Even though the sample size per area is fairly large (at around 1,200 per LA), the only sub-group splits included are by gender.

The Scottish Household Survey (SHS)

The Scottish Household Survey has a similar annual sample size to the National Survey (and Welsh Health Survey), at around 12,500 but the sample is not evenly distributed across all LAs. Instead the sample size follows the population size, with the qualifier that there is a target *minimum* sample per LA of 500 interviews every two years. So, in many Scottish LAs, the sample size per year is only around 250.

LA-level estimates are published every two years, and based on two years of aggregated data. Moving averages are not used.

In contrast to the Welsh Health Survey spreadsheets the SHS spreadsheets include time-trend data and a very considerable amount of sub-group estimation (key sub-groups being gender, age-group, and income). Sub-group statistics are shown unless the sample base for the sub-group falls below 70.

Some implications for the National Survey

These very different approaches to LA reporting perhaps set the boundaries between which the National Survey could position itself. It should be noted that some of the sample sizes in sub-groups in the SHS are very low. But with sample sizes of 1,200 per LA over two years it would seem perfectly possible to include more sub-group analysis from the National Survey than in the reporting of the WHS.

It is very worthy of note that if the SHS sample sizes are considered to be sufficient for LA-level reporting then there would be no requirement to do any aggregation of National Survey results at all: the sample sizes per year are large enough.

5 Assessing whether aggregation is appropriate

The first question to address when considering aggregation is whether the statistic being considered for aggregation is suitable, and whether, after aggregation the combined-year statistics can be sensibly interpreted.

The issues tend to be different depending on whether the underlying, year-on-year, statistics are of one of four types:

(1) No evidence of a trend over time, and relatively stable figures each year

Examples would be the percentage of the population describing themselves as being in good or very good health, which, according to the Health Survey for England, stayed at between 74% and 76% for English adults for the whole of the period 1993 to 2010. Or, the percentage of adults in the UK who describe themselves as '(somewhat, mostly or completely) satisfied with life' which has hovered around 77% for the whole of the period 2002/03 to 2009/10[9]. Over the last decade satisfaction with the NHS in England would fit this description too[10].

(2) Evidence of slow trends over time

Many health behaviour statistics fit this description: such as trends in obesity, or smoking, which tend to show no more than a one or two percentage point change year on year. Other statistics, such as trust in local councils in England, also fit this description, increasing from 52% in 2001 to 64% in 2010/11[11].

Other statistics, such as the unemployment rate, do not show a *monotonic* trend but tend to move up and down over a period of years. The change over time is still generally *fairly* slow however, and over a small number of years the trend will be slow-moving.

---

[9] British Household Panel Survey/Understanding Society (http://www.ons.gov.uk/ons/rel/wellbeing/measuring-national-well-being/discussion-paper-on-domains-and-measures/rft-individual-well-being.xls)

[10] See for example http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_126705.pdf

[11] Citizenship Survey

(3) Evidence of a rapid upward or downward trend

It is relatively difficult to find examples of statistics that change extremely rapidly over time, but the penetration of new technology gives one example. The Office for National Statistics reported that the percentage of adults in the UK using the internet every day increased from 35% in 2006 to 60% in 2010: a five percentage point increase each year.

(4) Instability year on year (for reasons other than random variation in the survey data)

The examples given above are all inherently stable statistics, whether they show trends or not. Other statistics are inherently unstable: namely public attitudes that change because of short-term social, political or economic events. An example here would be trends in consumer confidence, which can go up or down very rapidly. A more ambiguous example is given by the statistics on trust in Parliament (from the Citizenship Survey), which were relatively stable between 2001 and 2009, trust levels then fell sharply in 2009/10 when the press were covering stories around MP's expense claims, but then subsequently rose again. Whether this is an example of an inherently unstable statistic, or an example of a stable statistic that was affected by a very particular set of circumstances, is difficult to judge.

Note that we are making a distinction here between statistics that are unstable for genuine reasons (because what they are measuring is changing over time), and statistics that are unstable because of the survey sample size.

There are a number of other possible patterns for statistics over time. Many statistics show no, or slow, trends on a year by year basis but strong seasonal trends. Given the focus of this paper on aggregating whole years of survey data, we have not covered the issues surrounding seasonality. Those issues are not, however, inherently different to the issues covered in this report[12].

Taking each of the four types of statistic in turn we describe the implications for aggregation.

5.1 Statistic Type 1: No evidence of a trend over time, and relatively stable figures each year

For statistics that change very little year on year, and with no evidence of a trend, aggregation is a simple and uncontroversial option. For time-series such as these the

[12] For statistics that show strong seasonal variation standard practice would be either to publish data by season (usually quarter) which reveals the seasonality, or to aggregate data across four quarters to generate annual statistics (which smoothes out the seasonality). It would almost certainly be sensible to avoid any other levels of aggregation, such as 9 months or 18 months, where the seasons covered in successive aggregated statistics vary, because doing so would make interpretation of the change in the aggregate statistics very difficult. Our recommendation is that the *default* 'time unit' of aggregation should be a year rather than sub-sets of a year.

aggregated statistics will be a reasonably good representation of any of the years being combined. For example, the percentage of adults in the Health Survey for England describing their health as good or very good was 75% in 2007, 76% in 2008, and 76% in 2009. A three-year combined estimate would give an estimate of 76%, which would be a perfectly good estimate for any of the three contributing years.

For statistics of Type 1 it would be possible to generate aggregate statistics on a large number of years of data, which would increase precision (because of the increased sample size,) but with the aggregate statistic still being 'representative'. There are, however, practical reasons why aggregating many years of data is rare – these are discussed in Section 8.

5.2 Statistic Type 2: Evidence of slow-moving trends over time

For statistics that change over time, but only slowly, aggregation is also relatively uncontroversial *as long as the number of years aggregated is small.*

As an example here, the percentage of adults in the Citizenship Survey saying they trust their local council 'a lot' or 'a fair amount' was as follows:

Year	2001	2003	2005	2007/8	2008/9	2009/10	2010/11
% who trust their council	52	54	57	60	61	62	64

The Citizenship Survey only took place every two years prior to 2007, hence the gaps in the data. But focussing on the period from 2007, if a four-year combined estimate for 2007/8-2010/11 was to be created it would equal 62%. This is, arguably, only a good representation of the underlying statistics for the middle two years: 2008/09 and 2009/10. For the first and last years the aggregate statistic is two percentage points off.

Whereas, were two-year combined estimates to be used the combined estimate for 2007/8 and 2008/9 would be 61.5%, which is a close match to the two year-specific percentages; and the combined estimate for 2009/10 and 2010/11 would be 63% which is just one percentage point away from the two year-specific percentages.

Of course, whether the two-percentage point difference between the four-year combined estimate and the underlying statistics for the two end years is important is a matter of judgement. But, as a general rule, if aggregation is only over a small number of years users of those statistics will be more confident about using those statistics as being representative of each year in the period.

5.3 Statistic Type 3: Evidence of a rapid upward or downward trend

For statistics that are changing very quickly over time, such as access to the internet, aggregate statistics are only likely to be useful if the number of years aggregated is very small.

The trend in access to the internet amongst households in Wales between 2004 and 2008 is shown in the table below (the figures being taken from the Living In Wales 2008 report).

	2004	2005	2006	2007	2008
Percentage of households with internet access at home	43%	49%	52%	54%	60%
Percentage of households with a broadband connection	12%	25%	37%	45%	54%

The five-year combined statistic (2004-08) for the internet access percentage would be 52%. This is a good match to the middle year of 2006, but an extremely poor match for the first and last years. The two-year combined estimate for 2004-05 would, however, be 46%. And the two-year combined estimate for 2007-08 would be 57%. Although these are still a poor match to each underlying year, the mismatch is clearly much smaller.

The same applies to broadband connections, which, over the period 2004 to 2005 was on an even steeper trajectory than internet access. The five-year combined percentage would be 35% which is almost 20 percentage points away from the final year (2008), at 54%. A two-year combined estimate would give a figure of 50% for 2007-08 which is just five percentage points away from each of the contributing years.

In making the case for aggregation of statistics such as these, that are on a fast-moving trend, it is important that the uses that are to made of the data are taken into account. On the National Survey for Wales aggregate statistics will have three uses:

- Over time they will create a time-series for trends in Wales;
- They will allow for more detailed analysis of sub-groups; but, far more pertinently for this paper,
- Aggregate statistics are to be created for local areas (LAs and Health Boards) to give planners and policy makers reasonably <u>current</u> statistics on local populations. That is

they will be used to allow for statements to be made about the situation 'now', or at most, in the very recent past.

Given this need for 'close to current' statistics then, for the example of internet access given above, if 2012/13 and 2013/14 National Survey data was to be combined for local areas then the aggregate statistic would, most probably, best represent a point in time covering approximately the first half of 2013. Given that the combined figures for local areas will not be available until sometime in mid-2014 (once data collection is complete and the data cleaned), the 2012/13 combined estimates would then be seen as about 15 months out of date. Which, on recent trends, might imply that the figures under-represented the 'current' (i.e. mid-2014) rate of internet access by about six percentage points, and broadband connections by around 15 percentage points. (Note that if three years of data was to be combined, the percentages would be over two years out of date – so potentially up to 25 percentage points away from the mid-2014 figure.)

We have not identified any surveys where this problem of aggregate statistics being automatically out-of-date has been tackled. But it is clear that a possible way forward would be to generate the aggregate statistics using two years of data per local area, but then to use the national data to estimate the average change over time since the mid-point of the aggregate estimate and the most recent few months of the survey. Assuming the national change over time holds approximately true for all local areas (which would have to be assumed), local area aggregate statistics could then be modified using an 'adjustment factor' to make them approximately 'current'. To reiterate, we do not know of other surveys where this is done, and the approach would need careful testing (especially for extremely fast-moving statistics such as broadband access), but on the face of it this would seem to be preferable to the two alternatives of either simply not producing aggregate figures at local level for these fast-moving statistics, or producing aggregate figures and leaving it to data users to generate their own adjustment factor.

5.4 Statistic type 4: Statistics that are inherently unstable year on year

The final set of our four types of statistic are those that are inherently unstable. Examples here might be sharp changes in public opinion because of stories in the press. For instance fear of crime may rise very rapidly in the aftermath of a highly-discussed crime.

Statistics that show rapid fluctuations tend to be fairly rare in continuous national surveys that have fieldwork spread throughout the year and that report on an annual basis, because a short term change in public opinion will only be seen in a small number of months of fieldwork. But there have been some partial examples. For instance, the Citizenship Survey consistently showed the percentage of people saying they had trust in Parliament at between 35% and 38% between 2001 and 2008. But in 2009 the percentage fell to 29%, before rising to its previous level in 2011. The fall was almost certainly related to the press stories around MPs' expenses. (Arguably, this is not an inherently unstable statistic at all – the time series would suggest that it is a stable statistic that only moved because of a particular set of circumstances.)

For statistics that are genuinely inherently unstable year on year, the case for aggregation is different depending on how the aggregate data is to be used.

Firstly, if the data is to be used to demonstrate trends in Wales over a fairly large number of years aggregation may prove useful. Once a time-series of National Survey data has built up then tables and charts that show the trends over time might use aggregation (probably moving averages) to smooth out the 'noise' in the data and reveal any underlying trends.

If, however, aggregation is to be used to generate local area statistics then inherently unstable data is more problematic, since the combined-year figures may be a very poor match to any of the underlying years and the inherent instability may make it very difficult to give guidance on how to use and interpret combined-year figures for local areas. (With unstable statistics generating an adjustment factor would be difficult.) For statistics that are known to be inherently unstable we would advise caution in using aggregation.

5.5 Assessing which type of statistic you have

In planning for aggregation some assessment of which of the above four types your data and statistics conform to is needed. For retrospective aggregation (i.e. when looking backwards over a time series) then an assessment can be made based on the actual data. But the National Survey position is that aggregation is being planned for in advance of the data being collected. In this instance an assessment will have to be made using similar data from other surveys – together with an assumption that past trends in particular statistics will continue broadly in the same way. If a statistic has been stable or slow-moving in the past, then it will be reasonable to assume it will remain stable or slow-moving in the near future in most circumstances.

6 The gains in precision associated with aggregation

Our review of other surveys suggests very strongly that the primary reason for using aggregate statistics is to increase standard sizes. Clearly, for surveys that are of approximately equal sample size each year, two-year combined estimates will be based on double the sample size of single year estimates, and three-year combined estimates will be based on three times the sample size. (We deal with surveys that have varying sample sizes per year in Section 10.)

The implication of the increased sample size is that standard errors, and confidence intervals, will be reduced. Doubling of the sample size tends to lead to a 30% reduction in standard errors, whereas tripling leads to a 42% reduction. Taking an example of a statistic measured as 50% on a survey with a sample size of 600 per year (about the size of each LA's sample per year in the National Survey), the standard errors and 95% confidence intervals around this statistic will be as follows[13]:

Number of survey years aggregated	Sample size	Standard error for a percentage of 50%	95% confidence interval
1	600	2.0%	+/-4.0%
2	1200	1.4%	+/-2.8%
3	1800	1.2%	+/-2.3%
4	2400	1.0%	+/-2.0%

It is notable that the gains in precision decrease with each extra year. So there are considerable advantages in precision terms in using two-year combined estimates (rather than single years). But there is less advantage to be gained in using three-year estimates rather than two-year ones.

The standard error formulae are included in Appendix 1.

[13] In practice the standard errors will be slightly larger than this because of the effect of selection weights in particular.

7 The calculation of combined-year statistics

In all the discussion above we have implied that the calculation of combined-year statistics is a simple matter of taking the average of the statistics from each of the year. In practice there are decisions to be taken on exactly how to do this. There are (at least) three possible approaches:

1. Simple aggregation of the survey data
2. Equal aggregation weights per year
3. Unequal aggregation weights per year

7.1 Simple aggregation of the survey data

The review of other surveys that use aggregation suggest that what we here term 'simple aggregation' is the norm. Under 'simple aggregation' the survey data for each year that is to be aggregated is combined into one file and treated as if from a single sample. The survey data will have standard 'survey weights' (that is, selection and non-response weights) but no additional weights will be applied in the generation of the aggregate statistics.

In practice this approach means that the contribution each year of data makes to the aggregate statistics is in proportion to its survey-weighted sample size. If, for instance, the first year in a three-year combined statistic has a larger weighted sample size than the other two years, then the aggregate statistic will be slightly skewed towards this first year.

For surveys that have close to the same weighted sample size each year the risk of a significant skew towards any one year is minimal and, in reality, the simple aggregation approach will give close to equal weight to each year. It is only when a weighted survey sample size differs year on year that problems can arise.

7.2 Equal aggregation weights per year

An alternative method of aggregation is to force each of the years in a combined-year statistic to contribute exactly equally. Thus a three-year combined estimate would be calculated exactly as the sum of the statistics from the individual three years divided by three. In other words, each year has an aggregation weight of one-third.

Under the scenario where the survey weighted sample sizes per year do not vary (or vary only slightly) there will be very little difference between this equal aggregation approach and the simple aggregation. But if there is variation in sample size per year, then equal aggregation would almost certainly be preferable even though there may be a small loss in precision as a result of the aggregation weights.

Setting up survey data with aggregation weights is relatively simple. The data from each year would be combined into one file as with simple aggregation. But the survey weights per year would be scaled by an 'aggregation weight' that forces the survey-weighted sample size per year to be exactly equal.

Of the two approaches – simple aggregation and equal aggregation weights – we recommend the latter as the default.

7.3 Unequal aggregation weights per year

A final possibility is to actively choose to give unequal weights to different years in a combined-year estimate. In other words generate combined-year statistics that are more heavily weighted towards some years than others. There are two instances where such an approach *might* be useful:

- In the analysis and presentation of time-series based on moving averages, it is fairly common for analysts to give a greater weight to the 'central' year in the average rather than to the 'non-central' years. For instance, in a three-year moving average, a weight of 0.25 might be given to the first and last years and a weight of 0.5 to the middle year. This unequal weighting tends to generate a time-series that is smoother than the single year time-series, but is not as smoothed as the equal weight moving average time series. These questions will not become of relevance to the National Survey team for some years to come.
- The second instance is rather more hypothetical. But in theory at least, in generating survey estimates using combined years (for, say, LAs), more weight could be given to the most recent year. That would generate statistics that were closer to the final year in the aggregate statistic than the equal aggregation weights method. In the example of internet access, aggregating 2012/3 and 2013/4 data from the National Survey but applying more weight to the 2013/4 data would give an estimate that is closer to the probable figure in 2014 (when the data would be published).

  In practice we have never seen this done. And there are a number of disadvantages, such as difficulties in interpretation of the combined-year statistics, and a loss of precision (because of the unequal weighting). This does not look to be a sensible option for the National Survey.

What little literature there is on combined-estimates for surveys tends to focus on a third 'unequal aggregation weights' method[14]. That is, generate an 'optimal' set of aggregation weights that minimises the standard error of the combined-year statistics[15]. In practice, for surveys that have the same sample design year on year, the optimal aggregation weights will almost always be proportional to the sample size each year. This is essentially 'simple aggregation'.

7.4 Repeat cross-sectional samples with varying sample sizes per year

Some UK annual surveys vary their sample size quite considerably each year. The Health Survey for England, for example, has had 'full survey' years and 'half survey' years when the standard sample size is approximately halved. Deciding on how best to handle this in aggregate statistics is not straightforward. The natural approach would be to apply equal

---

[14] For example *www.stat.fi/isi99/proceedings/arkisto/varasto/kalt0185.pdf*

[15] The focus of the literature is on the scenario where concurrent surveys of very different designs are to be combined.

aggregation weights per survey year, so that each year, irrespective of its sample size, contributes equally to the combined-year statistics. But this means weighting up the small sample size year, which reduces precision.

To illustrate the issues, suppose there are three years of survey data to be aggregated for a statistic with a fairly marked trend, such as the percentage of men who are smokers. Assume that the sample size was 1000 in the first two years and 500 in the last, as follows:

	2005	2006	2007
% of men 'current smokers'	25%	23%	22%
Sample size	1000	1000	500

- With equal aggregation weights the three-year combined percentage would be 23%, with a standard error of 0.89%.
- With simple aggregation of the survey data, however, the three-year combined percentage would be 24%, with a standard error of 0.85%.

So the simple aggregation estimate has a lower standard error than the equal aggregation equivalent, but the aggregate statistic itself is somewhat biased towards the earlier years.

In this example avoiding bias would probably be judged to be of more importance than any small loss in precision. So a decision to use equal aggregation weights would probably be taken. Each case would, however, need to be judged on its merits.

7.5 Aggregation with grossed survey data

The discussion above is around aggregation for surveys that are weighted but not grossed to the population. However, in the circumstance where the survey data is grossed to population counts each year, so that the survey is focused on generating totals rather than means and percentages, similar issues around simple and equal aggregation arise. With simple data aggregation (that is, simply adding all the survey datasets together into a single file), this would generate a cumulative total across all the years being aggregated (for instance, with two years of data, simple aggregation would give a total that is around double the total for each individual year). In most circumstances this cumulative total would be a fairly meaningless statistic, and it would be natural to divide it by the number of years of aggregation. This division generates an aggregate statistic equal to the average of the individual single-year totals.

The alternative would be to do the aggregation before applying grossing weights. Under this approach 'aggregate survey means' would be generated using the equal aggregation weights method as described above. That is, the average of the survey means per year would be calculated. Then, to generate the grossed total the aggregate mean would be grossed to the population count. The question then is what population count to gross to: the average of the population counts for the underlying years, or the population count for a single one of those years?

Of these two approaches (simple aggregation of grossed data followed by division by the number of years; or grossing after aggregation) the first approach is probably preferable.

This is because it generates a 'grossed total' for an aggregated set of years that is easy to interpret: being the simple average of the grossed totals for each individual year.

7.6 Non-response weights in combined-year estimation

We mentioned earlier that each year of survey data is likely to have survey weights attached to it, these weights being made up of a combination of selection weights and non-response weights. Although the selection weights are essentially fixed and their calculation does not vary under aggregation, the non-response weights can potentially be calculated differently under aggregation. That is, the non-response weights can be calculated separately on each survey year prior to aggregation; or the survey data can be aggregated with just the selection weights and non-response weights applied to the aggregated dataset.

The difference is relatively subtle. For instance, if non-response weights are used to match the survey to the population in terms of age and sex, then with non-response weighting prior to aggregation each survey year will have the correct age and sex distribution. Whereas if non-response weights are applied after aggregation the aggregated data may have the correct age and sex distribution but individual years may not.

The argument for applying the non-response weights after aggregation is that the weights are more appropriate for the data being analysed, and are likely to be less variable than year-specific non-response weights (because variation in the non-response weights that does not change the aggregate statistics can be ironed out). This should lead to a *small* reduction in standard errors, although in practice the reduction is likely to be extremely modest. Furthermore, this small gain has to be traded off against the considerable amount of time non-response weights can take to calculate and test. Having to create non-response weights both for annual data and for combined-year data would be almost a doubling of the task. On balance, our recommendation would be to use the annual non-response weights that are calculated prior to aggregation.

8 The number of years to aggregate

In our review of how combined-year estimates have been used on other surveys we found just one instance of more than three years of survey data being aggregated (the Public Perceptions of the NHS Tracker Survey). And in that survey the aggregation is used to illustrate the relative differences between sub-groups (regions, age-groups, social grades) rather than as a means of generating more precise estimates. In surveys that are used to generate local area estimates that will be disseminated to local teams, we have only found examples of two or three years of aggregation.

For a survey such as the National Survey that has a wide range of statistics that would be of interest down to local level, two or three years of aggregation does appear to be appropriate. Over intervals of that time most statistics will not change a very great deal, so the aggregate statistics will give a reasonable reflection of each of the underlying years. And with, at most, three years of aggregation the time interval between the launch of the survey and the generation of the first set of aggregate statistics is *reasonably* short.

The decision on whether to use two years for aggregation or three will need to be based on a consideration of three factors:

- Whether the accumulated sample size is sufficient after two years for all the relevant sub-groups;
- Whether any of the key statistics are of our ‘Type 3’ (that is, on a very steep trend – see Section 5). If they are then two years of aggregation is preferable to three;
- How pressing the need for local/sub-group data is.

For the National Survey the sample size per Local Authority will be over 1000 after two years, so two years of aggregated data is almost certainly sufficient for LA and LHB reporting[16]. (In practice, with samples of this size, annual reporting may be acceptable for some ‘all population’ statistics, although the sampling error would clearly be fairly high.) In contrast, for minority sub-groups of interest three years of data may be needed.

It would be possible to argue that the number of years to aggregate should vary from statistic to statistic, with statistics that show a very marked trend over time, or based on the whole population (rather than on a sub-group), using fewer years. This would however be rather impractical, and a number of years will inevitably have to be chosen that is appropriate for ‘most’ statistics. Having said that, if three-years of aggregation was used as standard then it would be possible for some statistics in the tables to be based on two-years of aggregation, or even just one year if the sample size was sufficient.

[16] This is a judgement of course. Even with samples of over 1,000 short-term change would have to be very large for statistical significance. But samples of this size would normally be considered sufficient for profiling a sub-group or local area.

9 Publishing combined-year statistics (periodic reporting or annual moving averages)

There are interesting differences in how surveys generate combined-year estimates for local areas. The Scottish Household Survey and the Welsh Health Survey between them illustrate the differences. Both aggregate two years of survey data for their Local Authority level statistics.

- The Scottish Household Survey accumulates two years of data before generating the statistics, and produces its main LA reports every two years. So there are LA tables for 2007/08. And there are tables for 2009/10. There are no tables for 2008/09.
- The Welsh Health Survey generates aggregate LA statistics every year. So there are tables for 2007/08, 2008/09, and 2009/10. These are essentially moving averages, with each single survey year featuring in two sets of LA statistics.

It is not at all clear which of these is best, or which would be most appropriate for the National Survey. The advantages and disadvantages are summarised as:

	Advantages	Disadvantages
Annual reporting of moving averages	• More frequent statistics • More up-to-date statistics	• Moving averages can be misinterpreted, and clear guidance would need to be provided on how to use them • To generate statistics every year is more expensive
Non-annual reporting	• Statistics that are not based on overlapping years are easier to interpret and are less likely to be used incorrectly • The costs of producing the tables will be less	• Less frequent statistics • The statistics will become more out of date before they are replaced

Looking specifically at the first bullet point in the right hand column (‘Moving averages can be misinterpreted and clear guidance would need to be provided on how to use them’), this is a fairly crucial point. The Welsh Health Survey produces spreadsheets of LA statistics every year based on the moving averages, but (in contrast to the Scottish Household Survey) time series are not routinely published. It is quite possible that a time series will, however, be created by data users. If so, data users will almost certainly need guidance on how to do statistical testing on change over time statistics that take into account the fact that the sample sizes per year overlap. A simple 'calculator' in Excel might suffice.

The statistical issues are covered in more depth in Section 11.

10 Dealing with non-standard survey situations

All of the discussion so far has assumed that the surveys being considered for aggregation are annual cross-sectional surveys based on a fresh sample each year. This is, of course, not the only survey model. In this section we consider the implications for aggregation under three alternative scenarios: longitudinal surveys, rotating panel surveys, and mixed mode cross-sectional surveys.

10.1 Aggregation with longitudinal surveys

Aggregate statistics tend not to be used with longitudinal survey data. This is because the gains in precision that aggregation generates with fresh cross-sectional samples do not occur if the same people are in the sample in the years being aggregated. There will be some small gains in precision, but if the responses that people give in consecutive years of a longitudinal survey are highly correlated (which would be expected in most cases), the gains in precision will be very small indeed.

Being able to aggregate across years for extra precision is one reason why most of the large annual population surveys in the UK do not incorporate longitudinal data collection.

This has implications for moving the National Survey to this design. If, for analytic or cost reasons, it was decided that some respondents in a survey year would be followed-up the next year, then the proportion of the sample followed-up would contribute very little to the combined-year statistics from the survey.

10.2 Aggregation with rotating panel surveys

Rotating panel surveys are a variation on a traditional longitudinal survey. They (typically) start with a standard cross-sectional survey in Year 1. Then in Year 2 a proportion are followed-up and re-interviewed, but the rest are replaced with a fresh cross-sectional sample. In Year 3, the rules depend on the design of the rotating panel, but with a two-year rotation those followed-up in Year 2 would now be dropped and replaced with a fresh cross-sectional sample, whereas those new to the survey in Year 2 would be followed up.

Potentially, under such a design aggregation and longitudinal data collection might both be accommodated. But the aggregate statistics would benefit from the addition of the fresh cross-sectional samples each year.

10.3 Aggregation with cross-sectional mixed-mode surveys

More plausibly than either of the two designs described above, the National Survey may stay cross-sectional but with mixed modes being used each year. Under this design respondents would be ‘surveyed’ just once, but the mode of survey might differ from person to person: perhaps being web-based for some respondents, paper or telephone for others, and face-to-face for some. This design, *in principle*, creates no problems for aggregation. As long as each year of data is thought to give unbiased estimates, that are comparable year on year, then the years can be combined. However there may be a problem if a shift from standard face-to-face interviewing to mixed mode creates a break in the survey time-series: that is, if the change in mode changes the survey statistics. If this was to occur then it may then prove very difficult to generate combined-year statistics that span the break in the time-series.

11 Statistical testing with aggregate data

Section 6 dealt with the issue of standard errors and confidence intervals around combined-year statistics. In this section we touch on the issues of how to test whether change over time as measured by the difference in two aggregate statistics is statistically significant, and the implications for using statistical software.

The issue of statistical testing differs depending on whether or not moving averages are being compared with overlap years.

11.1 Comparing two combined-year estimates with no overlap years

If two combined-year estimates are being compared for which there are no overlap years the statistical testing is straightforward. The standard error of the difference in the two estimates is simply the square root of the sum of the two squared standard errors:

$$Standard\ error\ of\ difference = \sqrt{(standard\ error\ for\ statistic\ 1)^2 + (standard\ error\ for\ statistic\ 2)^2}$$

Most statistical software will handle this. The two sets of data would be added to the same file and weighted accordingly (using survey weights multiplied by the aggregation weights if used). A standard chi-squared or t- test of the difference between the two sets of data can then be applied.

11.2 Comparing two combined-year estimates with overlap years

Comparing two aggregate statistics with overlap years is slightly more complex. The issues become clearer, however, if the difference is divided into its component parts.

Suppose three-year combined estimates have been created, the first for 2008-10, and the second for 2009-11, so that 2009 and 2010 are overlap years.

Assuming equal aggregation weights the 2009-11 estimate would have been calculated as

$$\bar{x}_{09-11} = \frac{\bar{x}_{09}}{3} + \frac{\bar{x}_{10}}{3} + \frac{\bar{x}_{11}}{3}$$

And the 2008-10 estimate would have been calculated as

$$\bar{x}_{08-10} = \frac{\bar{x}_{08}}{3} + \frac{\bar{x}_{09}}{3} + \frac{\bar{x}_{10}}{3}$$

The difference between these two statistics would then be

$$\bar{x}_{09-11} - \bar{x}_{08-10} = \frac{\bar{x}_{11}}{3} - \frac{\bar{x}_{08}}{3}$$

That is, the difference between the two combined year statistics would be a third of the difference between the first and last years (2011 and 2008) and would make no use of the data from the intervening overlap years. It follows that to test the difference for significance, the test that has to be applied is whether the first and last years are significantly different.

To implement this in statistical software, the intermediate years that do not feature in the difference statistic would have to be filtered out: a standard chi-squared or t- test can then be applied to test for the difference in the non-overlap years.

This argument perhaps seems counter-intuitive at first. It is because it is easy to mis-understand how to test for change over time in the case of moving averages that in Section 9 we suggested that if local statistics were to be produced using moving averages, then those statistics should be published together with guidance (and probably an Excel calculator) to allow for the testing of differences.

12 The decision process

In this final section we summarise the formal steps that might be taken in considering whether an individual National Survey statistic is suitable for aggregation. Whether these steps need all be followed depends on whether the case for aggregation is clear-cut or not.

- Step 1: Identify an existing dataset/survey that gives a recent time series for the statistic, or a closely related statistic. Assess which type of statistic this appears to be (based on the four types in Section 5).
- Step 2: calculate aggregate statistics for this time trend based on two-year aggregates, three-year aggregates, and so on. Check the extent to which each of these leads to mis-representation of the single years included in the aggregation.
- Step 3: assess the gains in precision that might be achieved per LA/LHB (or other sub-group) under different numbers of years of aggregation.

If even two-years of aggregation do not give figures that are reasonable representations of the underlying two individual years then it is likely to be because the statistic is either of Type 3 (very steep time-trend), or of Type 4 (inherently unstable).

If the statistic is judged to be of Type 3 (steep trends):

- Step 4: consider whether an adjustment factor could be calculated that would help to generate a plausible 'current' local area estimate from the most up-to-date combined statistic.

If not:

- Step 5: Consider whether single year local area statistics would be of sufficient precision to be useful , even if it is not possible to generate the statistic for population sub-groups within the local area.

If the statistic is judged to be of Type 4 (inherently unstable):

- Step 4: Consider whether single year local area statistics would be of sufficient precision to be useful , even if it is not possible to generate the statistic for population sub-groups within the LA/LHB.

Even though this has been set out as a formal process, all of these steps are somewhat subjective and involve judgements. Step 1 of the process (deciding which 'type' the statistic conforms to) should help identify between those statistics that are judged non-controversial for aggregation and those that may not be suitable. If the 'type' that a statistic conforms to is not immediately clear then, under Step 1 the following analysis approach could be adopted:

1. Identify a recent time-series with, preferably, at least five years of data;
2. Calculate the mean across the five (or more) years;
3. Test (statistically) for a time-trend across the years[17];

[17] A simple regression model would suffice

4. If there is no time-trend then the statistic is likely to be of Type 1 (no trend) or Type 4 (inherently unstable); if there is a time-trend then the statistic is likely to be of Type 2 (slow moving trend) or Type 3 (fast moving trend).
5. To distinguish between Type 1 and Type 4, test whether individual years are significantly different to the overall average. If none are then the statistic is likely to be of Type 1. If several are then the statistic is likely to be of Type 4 and possibly unsuitable for aggregation.
6. The distinction between Type 2 and Type 3 statistics is necessarily subjective, although an average of around a three percentage point change year of year would probably be considered by most to be a 'fast-moving' statistic (so Type 4).

This analysis should not be used blindly however, and will in many instances be excessive (most statistics will be very obviously of a particular type). The results from the analysis need to be combined with a knowledge and understanding of how the statistics would be expected to change over time, how they will be used and interpreted, and a judgement made based on the combined evidence. More generally, it will be important for the judgements/decisions made about individual statistics to be transparent so that there is no suspicion that undesirable short-term changes in a time series have been purposefully 'ironed out' using aggregation.

13 Appendix 1: Standard errors for aggregate statistics

The standard errors for combined-year estimates differ depending on whether aggregation weights are used or not (see Section 7). If simple aggregation of the survey data is used then the standard error is calculated as

$$\sqrt{\frac{S^2}{\sum n_i}}$$

where S^2 is the overall population variance in the survey variable across all the survey years, and n_i is the sample size in year *i.* If the combined estimate is a percentage (*p*) then S^2 =*p(*100-*p).*

If aggregation weights are used the standard error is calculated as

$$\sqrt{\sum_i \frac{W_i^2 S_i^2}{n_i}}$$

where W_i is the aggregation weight applied in year *i.*