Methodology: Estimation

1 Estimation
The three main data sources in a S-DWH – census, sample survey and administrative – have very different origins:

- census data are usually a result of a legislative act – a national census carried out at regular intervals – and are included in a S-DWH as they represent the fullest coverage and most definitive measurements of the population of interest, albeit only at limited points in time
- survey data are only collected when there is a requirement for information which cannot be met directly or indirectly from existing data sources – and are included in a S-DWH initially for the purpose of producing specific outputs
- administrative data are uniformly collected for an alternative purpose – for example, tax collection – and are included in a S-DWH as they are freely available (subject to data sharing agreements), even though they are not always directly relevant to statistical outputs

Estimation can involve all three sources of data – in isolation, or in combination. The implications for a S-DWH are very different in each case, and need to be explained at least at a high level of detail.

2 Single source estimation
The methods used in estimation of statistical outputs based on single sources are very different, by necessity, for the three data sources.

Census
In theory, estimation is unnecessary when using census data, but in practice there is nearly always a small amount of non-response that needs to be accounted for. If adjustment is not required, or the adjustment takes place in census-specific production systems outside the S-DWH, then within the S-DWH estimation can be based on census data as a single source – otherwise combined estimation is required. The common approach to adjusting for non-response is based on “capture-recapture” methodology, requiring an additional data source (eg a census coverage survey). In an S-DWH environment it is essential to include all the additional data required for non-response adjustment, and to ensure that appropriate metadata exists linking these to the census data.

Sample survey
Using sample survey data as a single source when estimating a statistical output will be decided:

- a priori when designing the survey – only when the data are used to estimate their primary output
- as a result of testing – either when the data are used to estimate their primary output, or secondary outputs

Single source estimation is based on survey design weights for both primary outputs and secondary outputs eg derived variables or different geographical/socio-economic domains – and hence is known as design-based estimation. In a S-DWH it is essential to have comprehensive metadata accompanying the survey data in order to estimate secondary outputs, and also to ensure methodological consistency when combining the survey data with other sources. The metadata should at least include:

- Variables (collected and derived)
• Definition of statistical units
• Classification systems used
• Mode of data collection
• Sample design – target population, sampling frame, sampling method, selection probability (or design weight, which is the inverse of the selection probability)

Administrative
Although administrative data generally represent a census, there are still many issues (in common to most administrative data) when using them for estimation of statistical outputs:
• coverage issues – the target population of the administrative data collection exercise is unlikely to correspond to the target population of the statistical output – if overcoverage is the problem, the administrative source could still be used in single source estimation, but if undercoverage is the problem then combined estimation would be required
• definitional issues – the variables in the administrative source are unlikely to have exactly the same definition as those required by the statistical output – if the variables can be redefined using other variables in the administrative source, or simply transformed, it can still be used in single source estimation, otherwise combined estimation is required
• timing issues – the timing of the collection of administrative data, or the timeframe they refer to, are based on non-statistical requirements, and so often do not align with the timing required by the statistical output – to align timing this requires time series analysis (interpolation or extrapolation commonly) using the same administrative data for other time periods, in which case the estimation is still single source, or using other data source(s), which is combined estimation
• quality issues – as with census data, administrative data generally suffer from some non-response, which needs to be adjusted for during estimation – if non-responses are recorded as null entries in the dataset, then estimation can still be single source, but if other data sources are needed to estimate for non-response, it becomes combined estimation

In a S-DWH, the impact of these issues is that for administrative data to be used in single source estimation, both additional data – the same administrative source in different time periods – and thorough metadata (eg details of definitions, timing) are essential.

3 Combined source estimation
Data sources are combined for estimation for a very wide range of purposes, but these can be categorized into 2 broad groups:
• calibration – to improve quality of estimates by enforcing consistency between different sources
• modelling – to improve quality of estimates by borrowing strength from complementary sources

Methodological consistency
Any sources can be combined at a micro-level if they share the same statistical unit definition, or at a macro-level if they share domains, but using combined sources in estimation requires further effort to determine whether methodology is consistent (via analysis of metadata) as this will have quality implications for resulting estimates (eg even if variables share the same definition, differences in data collection modes and cleaning strategies could make results inconsistent and combining lead to biased estimates).

When combined sources are consistent in terms of methodology, but results differ for the same domains and same variables, then the reliability of two sources needs to be investigated:
• if the sources are combined for calibration (see below) the more reliable source is predetermined – by design – or the sources can be combined to form a new calibration total
• if the sources are combined for modelling, either the more reliable source needs to be determined via additional analysis – and identified as the priority source in processing rules – or the sources need to be combined as a composite estimator, acknowledging that neither is perfect, with weights reflecting their relative quality (e.g. a classic composite estimator uses relative standard errors to weight the components)

**Calibration**
The classic use of calibration is to use scale population estimates from a sample survey to published population estimates from a census, administrative source, or a larger equivalent sample survey. Known as *model-assisted estimation*, this adjusts the design weights to account for unrepresentative samples (e.g. due to non-response) based on the assumption that the survey variables are correlated with the variable that is used to calibrated to the population estimates (e.g. business surveys are commonly calibrated to turnover on the statistical business register). Hence this type of calibration is usually an intrinsic part of survey weighting. The assumption of correlated estimates and population variables is either made:

- *a priori* during design of the survey – when the data are used to estimate their primary output
- as a result of testing – either for the primary output, or secondary outputs

Calibration can also be an extrinsic process, such as contemporaneous benchmarking.

**Modelling**
Estimation based on modelling involving combined sources is rarely true *model-based estimation* – which assumes that a theoretical super-population model underpins observed sample data, allowing inference from small sample sizes – as the only practical application of model-based estimation is small area estimation (see below). More generally, modelling aims to replace poor quality or missing results – and is sometimes essentially *mass imputation*.

Modelling is generally used when a single source is unable to produce estimates of sufficient quality, or even at all, for domains (geographical or socio-economic) of interest. The additional source(s) either provide these estimates directly, or indirectly by specifying a model to predict them from existing data (or results) from the single source – this includes the mass imputation scenario.

A specific example of modelling is for census data, which require combined estimation to adjust for non-response. The common approach to census non-response is based on “capture-recapture” methodology, which requires a census sub-sample (e.g. a census coverage survey). In a S-DWH environment it is essential to ensure that appropriate metadata exists to link any additional source to the census data.

Small area estimation is a technique to provide survey-based estimates for small domains (often geographical), for which no survey data exist, and/or to improve the estimates for small domains where few survey data exist. The method involves a complex multilevel variance model, and borrowing strength from sources with full coverage of all domains – such as the census – selecting specific variables that explain the inter-area variance in the survey data. The chosen full coverage variables are used to estimate the domains directly, or in combination with the survey data as a
composite estimator. In a S-DWH environment, as long as the model is correctly specified in the analysis layer, the data requirements are still simply linked data sources – this time not at the micro- but the macro-level (aggregated for domains of interest) – and full and comprehensive metadata.

4 Outliers
Outliers are correct responses, as they are only identified once data cleaning is complete, which are either extreme in a distribution and/or have an undue influence on estimates. Outliers can cause distortion of estimates and/or models, so need to be identified and treated as part of estimation.

Common methods for identification and treatment are as follows:
- identification – visualisation, summary statistics, edit-type rules
- treatment – deletion, reweighting
- simultaneous identification and treatment – truncation, Winsorisation

In a S-DWH, identification and treatment both take place in the analysis layer.

Identification
Outliers can be identified qualitatively (eg visual inspection of graphs) or quantitatively (eg values above a threshold). Qualitative methods are more resource intensive, but are not necessarily of higher quality as the quantitative threshold is usually set subjectively, often to identify a desired number of outliers or a desired impact on estimates from treatment of the outliers.

Treatment
Outlier treatment fundamentally consists of weight adjustment:
- an adjustment to 0 percent (of original) equates to deleting the outlier (eg truncation)
- an adjustment to P percent (of original) equates to reducing the impact of the outlier (eg reweighting and Winsorisation)
- an adjustment to 100 percent (of original) equates to not treating the outlier (eg ignoring it)

All treatments reduce variance but introduce bias – so Winsorisation was developed to optimise this trade-off by minimising the mean squared error (the sum of the variance and the squared bias).

Outliers in a S-DWH
In a S-DWH environment there are three types of outliers – outliers in survey data, outliers in administrative data, and outliers in modelling:
- survey data – outliers are unrepresentative values, which means they only represent themselves in the population (population uniques) rather than representing (p-1) unsampled units in the population, as is assumed when weighting a unit sampled randomly with selection probability 1/p eg footballers with extreme salaries randomly selected from a general population
- administrative data – outliers are atypical values, which means they are simply extreme in the population as administrative data represent a census so do not require weighting and each unit
is treated as unique (eg similar sources are prioritised for updating a statistical business register, but if the difference between them is above a certain limit, this identifies an outlier)

- modelling\(^1\) – outliers are influential values, which means they have an undue effect on the parameters of the model they are used to fit (eg an extreme ratio when imputing & Fig Y below)

![Figure Y: Modelling outlier – regression without extreme x-value (LHS, green) and with (RHS, red)](image)

Identifying and treating outliers is complicated by the intended re-use of data in a S-DWH:

- survey data outliers are conditional on the target population (eg if the target population was footballers only, a footballer’s salary would no longer be an outlier)
- administrative data outliers are conditional on their use (eg if ratios of turnover to employment were consistent for two similar sources even though numerators and denominators were different – due to timing perhaps – the differences would no longer identify an outlier)
- modelling outliers are conditional on the model fitted (eg an outlying ratio for average of ratio’s imputation, would no longer be an outlier for ratio of averages & in Fig Y above if the model is “average y-value” the extreme response with an x-value is no longer an outlier)

In summary, any unit in a S-DWH can be an outlier (or not an outlier), conditional on the target population, the use in estimation, and the model being fitted. Hence, it is impossible to attach a meaningful outlier designation to any unit. The only statement that can be made with certainty is:

**Every unit in a data warehouse is a potential outlier**

It is not even possible to attach an outlier designation to any response by a unit – as it would have to record the use – ie the domain and period for estimation, and the fields combined and model used – and this will not be fixed given the intended re-use.

Given that neither the units in a S-DWH, nor the specific responses of units, can be identified as outliers per se, identification is domain- and context-dependent. This means that outliers are identified during processing of source data, and reported as a quality indicator of the output – if the output itself is stored in a SDWH, the outliers identified will become part of the metadata accompanying the output, but will not be identified as outliers at the micro-data level.

---

\(^1\) Modelling includes setting processing rules (for example, editing/imputation), as well as statistical modelling
5 Further estimation

Often referred to as further analysis techniques, index numbers and time series analysis methods are often an integral part of the process leading to published estimates, but these methods will have no impact on metadata at the micro-level as they are applied to macro-data only. In a S-DWH, processing should be automatic, so these further steps are assumed to be part of estimation.

Index numbers

If users are more interested in temporal change than cross-sectional estimates (*eg* growths not levels), instead of releasing estimates as counts they are often indexed – by setting a base period as 100 and calculating indices as a percentage of that. Indices are also used, sometimes in combination with survey sources to provide weighting, to combine changes in prices or quantities across disparate products or categories into a single summary value. There are also many different index number formulae (*eg* Paasche, Laspeyres) that can be used, and different approaches to presenting time series of index numbers (*eg* chained, unchained).

Interpolation and extrapolation

If estimates are required for time periods that cannot be calculated directly from data sources, time series analysis techniques can provide estimates between existing time periods – interpolation – and before the earliest (*ie* backcasting) or after the latest (*ie* forecasting) existing time periods – extrapolation. Both interpolation and extrapolation can be used in single source or combined source estimation, and for both there are a huge variety of methods available (*eg* ARIMA, splining).

Benchmarking

Benchmarking is a time series analysis technique for calibrating different estimates of the same phenomenon. This usually requires physically bringing the data (micro- or macro-) together, aligning metadata, and then choosing a linking method – but in a S-DWH the data are already in the same environment, with consistent metadata, and pre-defined linking methods: so benchmarking will be easier in a S-DWH.

The most common use of benchmarking is contemporaneous – calibration at the same point in time – but temporal benchmarking – calibration over time – is also used, especially in the context of seasonal adjustment (see below) where the annual totals of the seasonally adjusted and unadjusted estimates are constrained to be consistent.

The aim of benchmarking is constant – to ensure consistency of estimates – but can be approached in two fundamentally different ways:

- binding benchmarking – defines one estimate as the benchmark, and calibrates the other estimates to it
- non-binding benchmarking – defines the benchmark as a composite of the different estimates, and calibrates all estimates to it

Non-binding benchmarking is theoretically appealing, as no estimate – by definition – is correct, and non-binding benchmarking combines all the estimates to form an improved estimate, but it is very rarely used in practice. The main reason for this is revisions – binding benchmarking means that the more reliable estimate, which is used as the benchmark, doesn’t have to be revised. Given that the
benchmark estimate is usually a headline publication, it is understandable why producers do not want to change it – albeit possibly only by a small amount – based on a less public and less reliable estimate. Even if the headline estimate was released after non-binding benchmarking – which is feasible as being more reliable, is also likely to be less timely than the less public estimate(s) – any revisions to the less public estimate(s) would revise the non-binding benchmark, and hence cause revisions in the headline estimate.

**Seasonal adjustment**
Estimates have three (unobserved) components – long term change (*trend*), short-term sub-annual movements around the trend (*seasonal*) and random noise (*irregular*). As the seasonal component repeats annually (*eg increased retail sales at Christmas*) it can distort interpretation of short-term movements (*eg sales increases November to December do not imply an improving economy*). Hence the seasonal component is often removed from published estimates – they are *seasonally adjusted*. However, not all time series have a seasonal component (*eg sales of milk*) so seasonal adjustment is sometimes not required.

As the seasonal component is unobserved it has to be estimated – and as the nature of seasonality changes over time, the estimation parameters – and even the variables – also need to change to ensure the estimates are properly seasonally adjusted. The seasonal component can be estimated automatically, so this *moving seasonality* is not in itself a problem in a S-DWH. However, the nature of the seasonal component – a repeating annual effect – means that when the seasonal component is re-estimated, it is re-estimated for the entire time series. Hence any changes cause revisions throughout the time series. There are two common approaches to reducing these revisions – only revising the time series back to a certain point, and keeping the estimation variables for the seasonal component constant over a set time period (usually 1 year). The advantage of the first (*eg an up-to-date seasonal component for current estimates*) is the disadvantage of the second, but the advantage of the second (*eg a stable time series*) is not the disadvantage of the first, which is a discontinuity in the time series. Given that the chosen approach is usually applied to all outputs within an NSI, again this is not in itself a problem in a S-DWH.

However, a more problematic issue in a S-DWH is a *seasonal break*. This can be an abrupt change in the seasonal component (*eg in 1999 new car registrations in the UK changed from annual to biannual, and the seasonal component for new car sales immediately changed from having one annual peak to two*), or a series becoming seasonal (or non-seasonal). Although the treatment of seasonal breaks can be automated, their detection cannot be (with any degree of accuracy). As seasonal breaks can occur in *any* time series at *any* time, all seasonally adjusted estimates should be quality assured before release. Ideally, this quality assurance should be manual, but a compromise is to have an annual quality assurance supplemented by automatic checks to identify unexpected movements or differences (*eg between the unadjusted and seasonally adjusted estimates*).