Balancing and standardization

Josh Browning

October 30, 2015

Methodology

Balancing and standardization should only take place after all elements in the food balance equation have been collected/estimated/imputed. It is highly preferable to have actual measurement and data collection (surveys, censuses, administrative sources). While every effort should indeed be made to collect data, there is still the need to deal with missing data, ratios, shares and conversion factors, i.e. there is a need for imputation.

This document introduces an innovative balancing mechanism as well as the process of consistently aggregating commodity detail data into aggregates, i.e. the process generally referred to as "standardization".

The basic balance and the balancing mechanism

At the most basic level, Food Balance Sheets are, like all commodity balances, simple identities. In these identities, the sum of all supply variables is equal to the sum of all demand variables; the two most common identities set domestic supply equal to domestic demand (first equation) or total supply equal to total demand (second equation).

$$P_{ijt} + I_{ijt} - X_{ijt} - \Delta St_{ijt} = Fo_{ijt} + Fe_{ijt} + Lo_{ijt} + Se_{ijt} + IU_{ijt} + T_{ijt} + ROU_{ijt} \quad (1)$$

$$P_{ijt} + I_{ijt} - \Delta St_{ijt} = X_{ijt} + Fo_{ijt} + Fe_{ijt} + Lo_{ijt} + Se_{ijt} + IU_{ijt} + T_{ijt} + ROU_{ijt} \quad (2)$$

where $\Delta St_{i,j,t} = St_{i,j,t} - St_{i,j,t-1}$, P=Production, I=Imports, X=Exports, S=Stock level, Fo=Food, Fe=Feed, Lo=Losses & waste, Se=Seed, IU=industrial use, T=Tourist consumption, ROI=Residual Other Use. Moreover, the i index runs over all countries, the j index over all commodities, and t over years.

Ideally, as many variables as possible should be measured and measurement should take place with a maximum degree of accuracy. When and if empirically measured, measurement can and should include both an estimate for the expected value of every variable as well as its measurement error. In reality, this is seldom the case and a number of other problems can complicate the matter.

Firstly, measured values are mostly limited to variables on the supply side (production, imports and exports), but even when that is the case, measurement is typically available only for the expected values and not for the measurement errors. On the demand side, most estimates are imputed data and again, estimates are often limited to the expected values without their respective measurement errors.

Secondly, it is not possible to include values for all variables in the balance, at least if they enter the identity as point estimates. If that were the case, the balance would not have a solution unless one variable is left as a "balancing item". Inevitably, this balancing item assumes all measurement errors, implicit or explicit in all other variables. If all the estimates for the other elements are unbiased, i.e. have measurement errors with expected value 0, then the expected value for the error of the balancing item/residual will be 0. However, the variance of the error for the balancing item/residual is the sum of the variance of the measurement errors for all other elements; this will inevitably cause a large variability in the estimate of the balancing item/residual.

Thirdly, and in view of the fact that the balancing item assumes the measurement errors of all other variables, a prima facie case could be made that not all variables are equally suitable to function as balancing items. Intuitively, the variables with higher degrees of annual variability would be more suitable as balancing items.

The underlying rationale for such a choice would be the fact that the residual/balancing item assumes the sum of all measurement errors, and is therefore more likely to exhibit greater year-to-year variability.

In practice, however, the choice of a variable as a balancing item often reflects the availability of data (or the lack thereof), rather than a clear economic rationale and empirical evidence. It is therefore not surprising that different SUA compilers/SUA approaches have chosen different variables as their balancing items. USDA's balances, for instance, use feed (and residual use) as the balancing item, while the FBS often used food to balance supply and demand. Conveniently, the XCBS approach often chooses whatever variable is not explicitly available. Clearly, none of these approaches overcomes the problem of accumulating measurement errors in the balancing item. No matter which variable is used as the balancing item, this variable is fraught with the measurement errors of all other variables. In the context of the Food Balance Sheets, this means that using food as the balancing item would therefore be the least suitable solution.

The approach in detail

The new balancing approach aims to overcome the fundamental problem of identifying one single variable as the balancing item. To this end, all variables enter the balance with an expected value. Obviously, this is tantamount to creating an over-identified equation, at least if the estimates enter this equation as point estimates, i.e. only with their expected value. To overcome this problem, we use for every variable an expected value as well as its measurement error (e) or an approximation of the measurement error.

$$\begin{aligned} P + e(P) + I + e(I) - \Delta St + e(\Delta St) = X + e(X) + Fo + e(Fo) + Fe + e(Fe) + Lo + e(Lo) \\ + Se + e(Se) + IU + e(IU) + T + e(T) + ROU + e(ROU) \end{aligned} \quad (3)$$

The indices above are dropped for brevity, but the measured values and their measurement errors are still specific to each country/commodity/year.

The measurement errors used by FAO are based on information inherent in the meta-data of each variable. Overall, the term "measurement error" refers to the degree of confidence in a particular element of the food balance sheet. It should also be noted here that this approach does not change official data as we assume that official data has a measurement error of zero.

The availability of expected values and measurement errors is no guarantor for a balanced identity of supply and demand. If measurement errors are too small, and if too stringent bounds are placed on too many elements, the identity may not be solvable. Conversely, if the bounds are large, the identity has multiple or indeed infinitely many solutions. If many solutions are possible, the next step is to single out the best possible solution.

For the new balancing mechanism, the "best" solution is defined as the one that provides the combination of variables with the highest aggregate "probability". In more theoretical terms, this means that an objective function must be established that selects the combination of variables with the highest aggregate likelihood. The objective function needs to maximize the joint density of all distributions. In practice, the approach amounts to maximum likelihood estimation, i.e. a process that selects the combination where the product of all likelihoods has a maximum value.

$$Obj_{ijt}(x_1, x_2, \cdots, x_n) = \prod_{(k=1)}^{n} f_{ijtk}(x_k) \quad (4)$$

where $Obj_{ijt} is the objective function for the ith country, jth commodity and tth year, f_ijtk is the density for the kth element of the balance (i.e. production, imports, etc.) and x_k is the value for this kth element. The x_k are the parameters that are adjusted by an optimization routine; they will be set in a way

so as to maximize the objective function, and hence can be seen as the most likely values of the elements given the provided distributions.

If we allow the distributions for each of the different variables to be any arbitrary distribution, then optimizing the above objective function must be done numerically. Numerical optimization can solve difficult problems such as this, yet it has several drawbacks: it can be time-consuming, it can fail to converge to a solution, it can find a local rather than a global optimum, etc. However, if we assume all of the distributions above are normal distributions, then optimization becomes very simple:

1. Calculate the total imbalance (Imb_{ijt}) by computing

$$Imb_{ijt} = P_{ijt} + I_{ijt} - X_{ijt} - \Delta St_{ijt} - Fo_{ijt} - Fe_{ijt} - Lo_{ijt} - Se_{ijt} - IU_{ijt} - T_{ijt} - ROU_{ijt} \quad (5)$$

2. Adjust each element based on its standard deviation:

$$\begin{aligned} P_{ijt}(\text{balanced}) &= P_{ijt} - Imb_{ijt}/(e_{ijt}(P_{ijt})) \\ I_{ijt}(\text{balanced}) &= I_{ijt} - Imb_{ijt}/(e_{ijt}(I_{ijt})) \\ X_{ijt}(\text{balanced}) &= X_{ijt} + Imb_{ijt}/(e_{ijt}(X_{ijt})) \\ \Delta St_{ijt}(\text{balanced}) &= \Delta St_{ijt} + Imb_{ijt}/(e_{ijt}(\Delta St_{ijt})) \end{aligned}$$

   and so on, where the sign is positive for variables which have negative signs in the calculation of the imbalance and vice versa.
3. The above equations do not allow the enforcing of any bounds on the variables. If a bound is not met, that variable should be assigned the value of the bound, given a standard deviation of 0, and steps 1 and 2 repeated. For example, if the lower bound for feed is 100 metric tonnes but it is assigned 50 metric tonnes, we fix it to the value of 100 metric tonnes and re-allocate the difference (50 metric tonnes) to the other variables.

Constraints

The maximization/optimization process takes place under constraints. In principle, three types of constraints are possible. The first and most obvious constraint holds for every line of the balance and simply imposes that supply is equal to utilization for every commodity. This also includes the condition that $P + I - \Delta St > X$. A second constraint can be imposed to reflect priors over all rows of the food balance, i.e. a column constraint. It has been argued, for instance, that year-to-year swings in the sum of all calories in the FBS (DES) may not exceed the value of 100 kcal; larger swings may appear from a mechanical implementation of the FBS framework, but are unlikely to occur in practice; or they do, only under special circumstances (massive economic problems or natural disasters). Thirdly, there may also be the desire to impose a multi-year constraint for an individual variable. Stock changes in one direction, for instance, are unlikely to occur over many years in a row, at least not for large aggregate quantities. Finally, the feed use estimation process imposes the constraint that the sum of the respective nutritive values (i.e. energy and protein content) of any likely combination of commodities satisfies the biological requirements of livestock, poultry and aquaculture as estimated in the feed use module.

Possible problems, infeasibilities

The proposed objective function maximizes the overall probability under multiple constraints. An optimization process that takes place under multiple constraints can lead to infeasibilities; the occurrence of infeasibilities will increase with the increased number and stringency of the constraints. Such infeasibilities occur at different stages and these stages also help determine the strategy to overcome the infeasibilities. Different cases of constraints and infeasibilities can be distinguished:

- No solution with only a row constraint of supply equal to demand. If supply cannot be matched with utilization for any given row/commodity, the imbalance is larger than the aggregate measurement error and no combination of variables within their respective measurement errors would allow us to reach a balanced identity. Obviously, this means that the bounds on the distributions are too stringent to establish the balance. In this case, one or several bounds need to be relaxed to ensure that a combination of the various variables can be found to balance supply and demand. The need to respect official data may mean that only one or a few bounds can be enlarged or lifted.
- Solution at row level, infeasibility for column constraints. Once a solution for every row has been established, we may encounter a problem with the column constraints. For example, the year over year change in the DES may be larger than some reasonable threshold, and in such a case we will need to adjust the elasticities in the food model and then re-run the optimization procedure to meet column constraints. In the best case, a solution can be found with all constraints imposed. Failing this, the column constraints need to be reviewed and eventually lifted. Alternatively, single row limits on food availabilities may need to be tightened.
- Solution at row level and column level, infeasibility for multi-year constraints. The strategy here should be rather straightforward; it only involves lifting the limits of stock changes over time. Eventually all constraints should be met. If not, all other constraints would need to be lifted gradually to eventually render a solution. Lifting these additional constraints requires manual intervention and judgement as to what constraint should be considered most binding.
- All constraints met. In this case, the food balance sheet can be considered ready for an inspection by a food specialist and refined where and as necessary.

Review and evaluation

The new procedure amounts to a radical departure from all traditional balancing approaches. The previous approach, as well as all approaches implemented by other SUA compilers, is deterministic in nature. The new approach by contrast is a probabilistic one. It identifies a most likely combination of variables whereby every variable can only vary within its prior bounds (support points). It requires every variable to be identified not only with an expected value, but also with a measurement error or at least a likely range within which the element's value can vary. Ideally, these bounds/measurement errors are available from empirical measurements, e.g. from representative surveys measuring food consumption or other forms of supply and utilization.

In effect, this system means that every variable of the balance is a balancing item, at least within the bounds of its likely value. The final determination of every variable and the final combination of all variables is determined in the constrained optimization process outlined above.

Figure 1 captures the basic set-up of expected values and associated distributions. It is based on the wheat balance for the US.

Figure 1: US wheat, unbalanced (top pane), balanced (bottom pane). In the top pane, the expected value is represented by the dashed line (and the highest point of the distribution). In the bottom pane, the dashed line represents the value after balancing.

Pros and cons

The radical departure from traditional balancing provides numerous practical advantages and theoretical consistencies. But it also imposes greater computational burdens and requires a deeper understanding of statistical processes.. The key advantages and disadvantages could be summarized as follows.

Pros

- No residual necessary. The problem that haunts all traditional approaches is that the need to identify a residual means that this variable is not only the balancing item of the identity, but also that it takes

all measurement errors of the balance. As there is no a priori reason to assume that the sum of the measurement errors is equal to zero, the balancing item is in essence the least reliable estimate of the balance. If food consumption is chosen as the balancing item, the inherent year to year fluctuations are likely to be huge and its usefulness for many purposes limited. Moreover, such a wide range of fluctuations would not meet the a priori economic rationale, which would suggest that food consumption is moving over time in a rather smooth manner and that other elements notably stocks, feed use or waste would vary more significantly.

- Harnesses all information. Information about the quality of every variable, when and where available, is retained in the creation of the balance. For instance, if some variables of a balance are measured by high quality surveys, they will be maintained with their exact value; conversely, variables that come without prior knowledge of accuracy, or a low degree of confidence, will be allowed to vary widely in the balancing process.
- Rules-based and reproducible process. The identification of every variable follows a rules-based process. The same holds for the optimization process. This allows tracing decisions and reproducing results.
- Most likely outcome. The overall solution is a most likely outcome, based on a probability maximizing algorithm, not one of many possible (notionally infinite) solutions of an under-identified equation.
- Flexibility. At any stage and for every variable, both expected values and confidence intervals can be overwritten by expert knowledge. The optimization algorithm then runs within the bounds provided by experts.
- Multiple constraints. Apart from the constraints that the balance has to be in equilibrium for every row, the approach allows to impose constraints across column sums and time. For instance, a constraint can be introduced that avoids large year-to-year fluctuations in the overall calorie availability (DES). Such a constraint reflects prior knowledge about people's desire and ability to keep a stable calorie intake level, at least within a certain constraint. Likewise, limits on stock changes can be introduced to reflect multi-year accumulations or drawdowns.

Cons

- Knowledge of prior information must be formalized.
- Statistically complex procedure; data and IT intensive.
- Multiple constraints. Imposing multiple constraints is both an advantage and a disadvantage. Such constraints can result in infeasibilities in the solution, or lead to border solutions, i.e. solution along the constraints.
- The balancing process is limited to the aggregate commodity level. This may cause some inconsistency with the underlying SUA calculations; the latter are considered auxiliary calculations with the sole purpose of converting processed products back into the primary equivalent; as this process affects mainly, though not only trade, a constraint can be introduced to limit possible posterior adjustments of imports or exports in primary equivalents.
- Additional effort is now required for each element. Expected values and error distributions must be provided for each element, and this requires more advanced modelling than just estimating expected values.

Measurement errors/confidence intervals

The measurement errors that are needed for the proposed method are metrics for the reliability of the mean. They are not measures of the distribution of the variable itself (i.e. the statistical population). But as outlined above, the measurement errors can be derived from the distributions of these variables. For instance, it is not the distribution of food access in a population that provides the metrics needed for the balancing mechanism, but the measurement error of the mean of food consumption. There are different strategies to derive these measurement errors: they can be derived directly from primary data, extrapolated from the analysis of metadata coming from secondary sources or be gleaned from the overall quality of measurement and statistical information inherent in the overall quality of a statistical system. Some of these strategies are outlined in brief below:

Primary data: actual measurement of expected value and errors, confidence interval

In the ideal case, information is available for as many variables as possible from direct measurement, i.e. from primary data collection. It is important to note from the outset that the FAO Statistics Division does not collect primary data, nor is there much information available on measurement errors from other sources. It has, however, access to a limited set of primary data, mainly household income and expenditure surveys (HIES) from which the mean of food consumption and its measurement error can be derived. These data do not directly, or in any case not fully, comply with food availability as defined in FBSs. For instance, they do not fully cover the category of food consumed outside the home ; namely, food consumed in restaurants, cafeterias, street food, etc., are not included. Additionally, food consumed in public households/locations such as hospitals, military, prisons, etc. is also generally excluded.

Secondary data. Official country notifications

While the FAO Statistics Division has only limited access to primary data for any of the FBS variables, it collects regular and ample information for many variables from country questionnaires, or other official and semi-official sources. The variables covered every year and for a large number of countries and commodities include production data as well as data for imports and exports. The data on production come from country questionnaires whereas data on imports and exports come from the UN Comtrade database and thus ultimately from customs offices. Obviously, neither source provides direct measurement errors. But there is information that can be used to approximate measurement errors.

- For production estimates, officially provided data should be assumed to have no or small measurement errors. For countries with weak statistical systems, and where balances cannot be brought into equilibrium, a deviation from official data can be justified, provided that there is enough evidence to justify such a deviation. This means that official data could be overwritten in line with an allowance dependent on the quality of the statistical system. Such information can be gleaned various sources, most conveniently from the Country Assessment Questionnaires (CAQs) processed by the Global Strategy to improve Rural and Agricultural Statistics.
- For trade estimates, the FBS system requires a fully balanced global trade matrix. To this end, the FAO statistics division has established a procedure that brings world exports and imports for every commodity into equilibrium . This process uses information about the quality of data reporting by every country and based on this creates an endorsement factor. The information from the endorsement factors is a good first approximation of the reliability of trade information, and can thus be used to derive measurement errors for the FBS balancing system.
- Imputing data. The FAO Statistics Division imputes data when they are not available from primary sources or, importantly, country questionnaires. This holds for variables for which no questionnaires are dispatched, or for data not covered in questionnaires; it also includes imputed data for countries not replying to questionnaires at all. Again, there are two basic cases to be distinguished with regard to gauging the measurement error. Some imputation methods actually provide an estimate for the measurement error. This is, for instance, the case for feed use. Others do not provide any estimate for the accuracy of the estimate or such measurement errors are difficult to extract from the method (e.g. the imputation of production). In these cases, the only gauge for the measurement error can come from the reliability of a country's statistical system.

A hierarchical system of flags

The FAO Statistics Division documents the The proposed balancing mechanism requires an FBS compiler to provide an estimate for a confidence interval around an initial estimate of a variable. These confidence intervals reflect the degree of reliability of such an initial estimate. FBS compilers can revert to different sources and employ strategies to arrive at quantitative estimates for these reliability measures; they may be able to use information about the measurement errors collected with the primary data (e.g. a household

survey), assign confidence ranges based on the imputation methods or, in the absence of specific information, simply assign confidence ranges in line with their prior knowledge of the way an estimate was produced, reflecting the approach used, the quality of the statistical system in general or for a particular survey, etc.

Given the need to assign such confidence values in the absence of country-specific information at FAO, the confidence intervals assigned by FAO are based on two readily available factors.

The FAO metadata/flags system

The FAO Statistics Division collects information from different sources and assigns different reliability levels to the various sources. These reliability levels are encapsulated in the FAO "flags" system. The flags are used to approximate confidence ranges as a first step.

Five levels of confidence are distinguished as default values to arrive at confidence intervals around an expected value. The highest confidence is placed on official numbers/estimates, which would be taken as point estimates without a confidence interval, while the lowest confidence is placed on estimated data. All confidence estimates are further modified by an allowance for the quality of a country's statistical system, which results in a hierarchy created by the quality of a statistical system nested in the confidence intervals arising from the general flags. The highest trust/narrowest confidence intervals are assigned to data from countries with excellent statistical systems, while the lowest confidence/widest intervals are assigned to imputed or missing data from countries with very weak statistical systems. An example, categorizing these levels of confidence is provided below:

Table 1: Confidence levels based on FAOSTAT flags

Source	Confidence	Implied Measurement Error
Official	1.00	0%
Semi-official	0.90	10%
Imputed	0.85	15%
Estimated	0.60	40%
Missing	NA	NA

Quality of a country's statistical system

To capture cross-country differences for the same flag, information from the assessment of statistical systems from the Global Strategy has been used to fine-tune the levels of confidence and thus determine the final confidence intervals for every estimate. Most of the information was gleaned from the Country Assessment Questionnaires (CAQs) processed by the Global Strategy to improve Rural and Agricultural Statistics, which however do not provide a full global coverage.

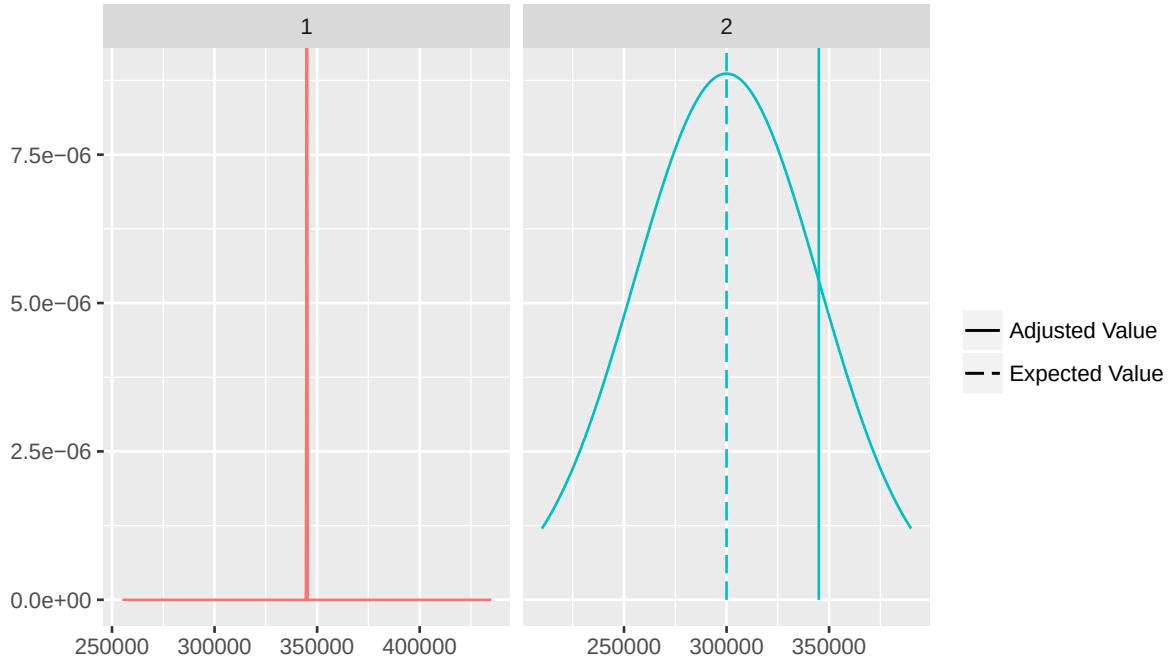
FBS compilers at country level may want to use the FAO confidence intervals as a first approximation of the quality of the estimates. Country level FBS compilers are likely to have more information available, and more reliable information, about their statistical system in general and individual FBS variables in particular. They are therefore strongly encouraged to carefully review the FAO confidence intervals, compare them with specific information available from their statistical system/experience and replace where appropriate.

Implementation

R code has been developed by FAO in the form of an R package which implements this algorithm. This section details several use cases/examples of how the function works.

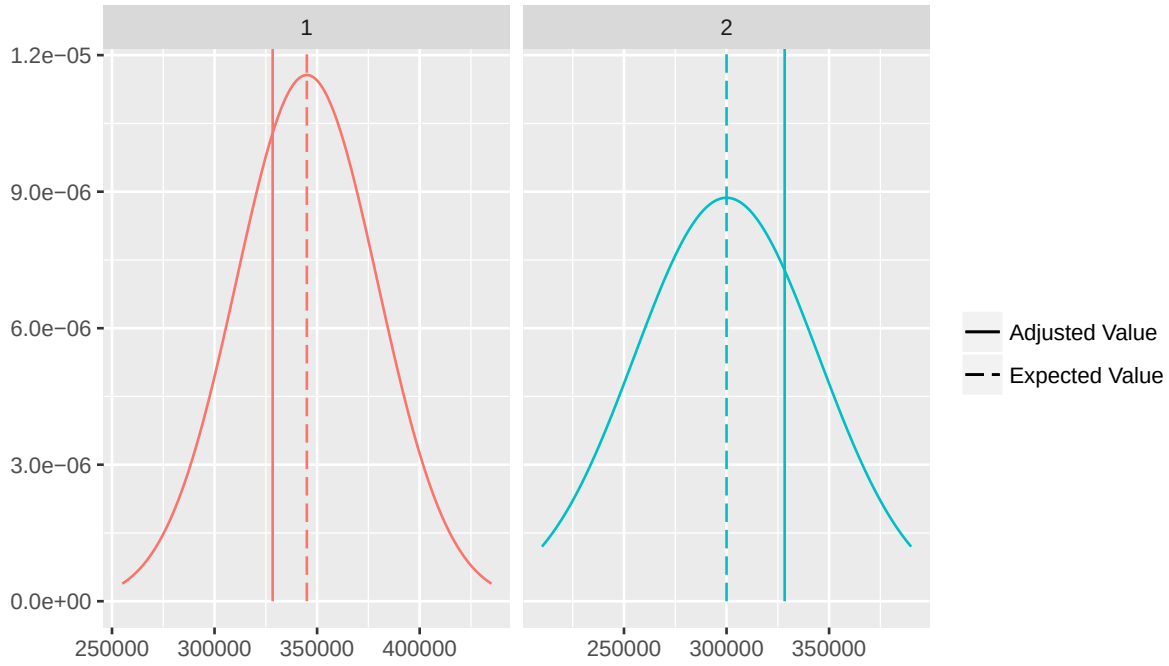
Simple Example

To illustrate how the function works, let's first consider a very simple example. Consider a simple commodity, let's say pineapples, and assume that in some particular country these pineapples are only eaten in raw form and have no other utilizations. Also, let us suppose that no trade/stock changes occur, and thus the only two non-zero elements of the food balance equation are production and food. Suppose that this country provides us with an official estimate of pineapple production of 345,000 tons, and that we estimated a food utilization of 300,000 tons via our food module. Then, assuming the official figure has an error of 0 and the estimate an error of 15%, we would have the following distributions:



The balancing algorithm would return the following estimates for production and food, respectively: 345,000, 345,000. The official estimate is not updated, as expected, while the estimate for food is updated to create a balanced equation.

Now, suppose the 345,000 value for production was not an official figure but rather a semi-official figure. In that case, the error becomes 10% of 345,000. Assuming the same value/distribution for food, we have the following figure:



In this case, the values after balancing for production and food, respectively are 328,341, 328,341. The semi-official estimate is not adjusted as much as the food estimate because it has a lower error.