FAO Food Balance Sheets: A summary of recent innovations

Zero Draft

Not for citation and/or circulation

Table of Contents

List of Figures

List of Tables

Introduction

The FAO Statistics Division (ESS) reviews and revises the methodological approaches for all its products on a regular basis. Such revisions include all databases maintained at ESS, their underlying and accompanying metadata, the approaches to impute missing data or to create analytical databases such as the Green House Gas (GHG) inventories, the Food Balance Sheets (FBS), or most recently, the System of Economic Environmental Accounts for Agriculture, Fisheries and Forestry, SEEA-AFF.

This document provides an overview of proposed methodological changes and innovations for the Food Balance Sheets. All analytical databases are, by their very nature, datasets that include a large number of imputed data or analytically derived data. Food Balance Sheets require many, and often particularly complex transformations of primary data. In undertaking these transformations, FAO always emphasizes how important it is for countries to undertake actual data collection and encourages all countries to improve and increase data collection efforts through the Global Strategy to Improve Rural Agricultural Statistics. The efforts to obtain a maximum of actual data notwithstanding, establishing Food Balance Sheets is often a process that starts with a rather limited set of hard statistics. For many countries and many commodities, actual measurement of the constituting variables is entirely absent, or where available, associated with large implicit or explicit measurement errors.

In addition to the need to impute an often large number of variables in a balance, setting up a complete set of Food Balances requires a multitude of conversion coefficients, extraction rates and nutritive factors. These too can change, albeit typically at a much lower speed. For this round of FBS revisions, the changes in conversion rates and factors were more important than on previous occasions. The main reason for additional changes lies in the fact that the underlying commodity classification systems have been revised as part of the overall reform efforts. The main change in this regard was that FAO's proprietary classification system, the so-called FAO commodity list, has been replaced by adopting the Harmonized System for all trade variables, and the Central Product Classification CPC for all other variables in the balance. Every effort was made to ensure consistency between old and new systems and across the new systems. All these efforts notwithstanding, some conversion factors had to be adjusted to reflect the product definitions of the newly adopted systems.

The focus of the revisions was, however, placed on updating the various (imputation) methods of the FBS components and, importantly, the overall approach to set up and solve the balance between all variables of supply and utilization. The motivations for these changes and the differences to existing approaches are laid out in the various chapters of this document. An important change relates to the approach taken to solving the overall balance. In essence, it constitutes a move from a deterministic approach towards establishing a process that takes into account not only the expected values but also the accuracy with the various variables ("elements") of the balance that are being measured. The approach eventually selects a combination of values for the various variables that provides the most likely outcome while taking into account the boundaries of confidence/measurement for every individual variable.

Overarching findings and principles guiding the FBS revisions

A number of overarching findings emerged from the current FBS revisions. Most importantly:

- All assumptions made are explicit and are documented. The same commitment holds for future changes and new assumptions made.
- Food Balance Sheets (FBS) are analytical data sets. They will always have to combine measured with imputed information. Imputation methods cannot replace data collection efforts, no matter how sophisticated they are.
- Every effort has to be made to collect more and better quality data at the country level, not least because the quality of the results of any imputation depends critically on the quality of measured information.
- Poor imputation methods can sometimes create vastly inappropriate results even where they are based on solid data. Every effort was made to identify sound imputation methods and base them to the extent possible on solid data.
- Imputation methods try to harness links between the various FBS variables and elements and information from outside the FBS. This allows triangulation of information and ensures overall consistency between FBS variables. The new feed use imputation method is probably the best example of how this has been implemented in practice.
- Analytical datasets are always associated with larger inaccuracies, stemming from differences in data definitions, classifications, measurement errors, imputation problems, etc. To reflect these issues in the FBS results, all estimates have expected values and an explicit measurement error. No claim is made that the estimates are point estimates. The overall philosophy guiding the revisions is to be “roughly right rather than precisely wrong”[1].
- The new approaches seek to harness innovations in both statistical approaches and new information and communications technology (ICT) to a maximum extent. However, they do not intend to replace manual inputs and quality checks. On the contrary, time saved through automatic procedures is meant to provide more time for quality assurance and quality control (QA/QC).
- QC/QA procedures are built into the system at various stages. Full compliance with the new FAO QAF will be achieved as both frameworks mature.

[1] John Maynard Keynes

The rest of this document is organized as follows. The first section lays out the details of the new balancing mechanism. It introduces the balancing mechanism, standardization and the methods to identify measurement errors.

The subsequent section introduces the innovations for all individual components. It starts with the supply side variables, i.e. production, imports, exports and stock changes, which are then followed by the description of all utilization variables, namely food, feed, seed, losses and waste, industrial use and tourist consumption. The document concludes with an overview of the changes introduced by the shift to the new commodity classification systems, their compatibility (or lack of) and their relationship with the existing classification system.

The purpose of this document

This document provides an overview of the innovations introduced with the new Food Balance Sheets methodology. The emphasis is placed on providing an overview of individual innovations, their motivation and how the various innovations are linked to each other. It uses illustrations and examples where possible, or at least where necessary, to explain the nature of the various innovations. As an overview, it can be consumed from front-to-back. It may be a starting point to explore all methodological or practical aspects of any or all of the innovations presented in greater detail.

The document is, however, not a surrogate for the detailed description of the methodologies used for every change and innovation. These are provided as separate documents and reference is made to them throughout the document. Nor is it meant to be a document to help roll out the new methodology to FAO member countries. Such a step-by-step introduction to the new Food Balance Sheets will be provided at a later stage. It will be part of an overall package that includes "worked examples", e-learning material and a multitude of practical applications, including a software package that allows practitioners to implement the methodology at country level.

Finally, many of the innovations developed will be applied to balances for non-food products such as cotton, jute or rubber. The difference to, and the analogies with the FBS system will be laid out in a separate document.

The IDWG-Statistics will be invited to review all methodological documents and to provide comments and suggestions for further improvements. At a later stage, comments will also be sought on all new capacity development materials.

The balancing mechanism

Introduction

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 or total supply equal to total demand.

Domestic Supply = Domestic Utilization

$$\boldsymbol{P_{ij} + I_{ij} - X_{ij} - dSt_{ij} = Fo_{ij} + Fe_{ij} + Lo_{ij} + Se_{ij} + IU_{ij} + T_{ij} + ROU_{ij}} \quad \text{(Equation 1)}$$

Total Supply = Total Utilization

$$P_{ij} + I_{ij} - dSt_{ij} = X_{ij} + Fo_{ij} + Fe_{ij} + Lo_{ij} + Se_{ij} + IU_{ij} + T_{ij} + ROU_{ij} \quad \text{(Equation 2)}$$

Where $dSt_t = St_t - St_{t-1}$, P=Production, I=Imports, X=Exports, Fo=Food, Fe=Feed, Lo=Losses & waste, Se=Seed, IU=industrial use, T=Tourist consumption, ROI=Residual Other Use[2]

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 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[3].

Secondly, it is not possible to include values for all variables in the balance, at least if they enter as point estimates. In this case, the balance would not have a solution unless one variable is left as a “balancing item”. This variable is the residual/unknown of the balance and the identity is solved for the residual. 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 residual will be 0. However, the variance of the residual is the sum of the variance of the measurement errors for all other elements, and herein lies the problem. Allocating such a large variability to the residual, and no measurement errors to the other elements, does not seem appropriate.

[2] All variables/elements are described in detail in their respective chapters

[3] What is more, many estimates are the result of rather complex imputation methods, involving many steps and many variables of the balance which make an empirical estimate of the measurement error difficult.

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 of data), 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 use food to balance supply and demand. Conveniently, the XCBS approach often chooses whatever variable is not explicitly available. Clearly, none of these approaches renders a satisfactory solution to the problem and, no matter what variable is used as the balancing item, this variable is fraught with the measurement errors of all other variables. Given the fact that there is no a priori reason to assume that the measurement errors cancel out, the balancing item is bound to be the most inaccurate variable of the balance. Extending the logic to the Food Balance Sheets, using food as the balancing item would therefore be the least suitable solution.

The approach in detail

The approach presented here tries 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[4] 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_i) or an approximation of the measurement error.

$$P_{ij} + e_{ij}(P_{ij}) + I_{ij} + e_{ij}(I_{ij}) - dSt_{ij} + e_{ij}(dSt_{ij}) = X_{ij} + e_{ij}(X_{ij}) + Fo_{ij} + e_{ij}(Fo_{ij}) + Fe_{ij} + e_{ij}(Fe_{ij}) + Lo_{ij} + e_{ij}(Lo_{ij}) + Se_{ij} + e_{ij}(Se_{ij}) + IU_{ij} + e_{ij}(IU_{ij}) + T_{ij} + e_{ij}(T_{ij}) + ROU_{ij} + e_{ij}(ROU_{ij}) \qquad \text{(Equation 3)}$$

Where measurement errors are not available from surveys or estimation processes, an approximation is used. A number of approximation possibilities are being proposed, including information inherent in the meta-data (flags) or the historical variability. In the absence of any quantitative information, an estimate for the measurement error can be provided by expert information (prior, support point). In this case, the measurement error would vary between two

[4] Strictly speaking it is not an over-identification, as this would require more than one equation. The problem is however the same in principle.

support points, which reflect expert knowledge of reasonable/expected minimum and maximum bounds[5].

In this context, we use the term "measurement error" to refer to the degree of confidence in a particular element of the food balance sheet. We estimate elements of the balance in terms of the total quantity per year, or in terms of the average quantity per person per day (in which case a simple conversion can be applied to convert to a quantity per year). For example, household surveys provide a measure of the distribution of food consumption per person per day. Such distributions can be very wide, representing a large spread in the calories consumed per person per day. While such information is useful in its own right, it must be translated into a different form in the FBS context. We need to convert such data, say C_j for each individual j, into an estimate for the average food consumption per person per day, say $\overline{C}$, and the distribution of this quantity is substantially different. By the property of the central limit theorem, we have the following:

$$\bar{C} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{(Equation 4)}$$

where μ is the true mean food consumption per person per day, σ^2 is the variance of an observation (estimated as the variance of C_j), and n is the number of individuals sampled. The implications of this can be seen in the Figure 1 below: the distribution of the DES for each individual is quite wide, but the distribution of the mean value is very narrow. Note that 1,000 observations were used in this example, and more (less) observations would have led to a narrower (wider) distribution of the mean.

[5] If no information for a possible distribution is available, a truncated normal distribution is assumed for all variables.

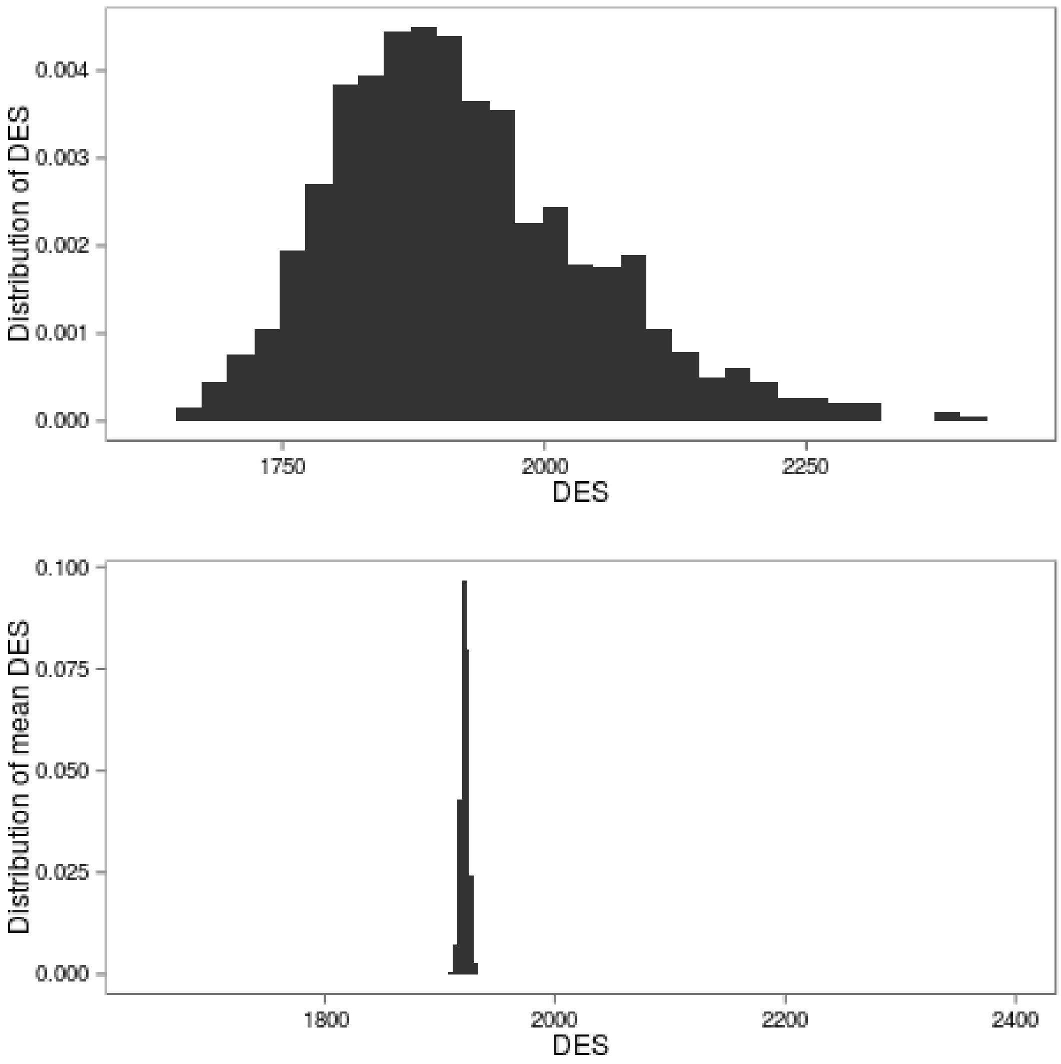


Figure 1: From distributions to measurement errors, using central limit theorem.

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 of 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.

The best possible solution requires a clear definition. 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 practice, 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 aggregate density of all distributions. In practice, the approach amounts to likelihood maximization, i.e. a process that selects the combination where the product of all likelihoods has a maximum value.

$$Obj_i(x_1, x_2, \ldots, x_n) = \prod_{j=1}^{n} f_j(x_{i,j}) \qquad \text{(Equation 5)}$$

where Obj_i is the objective function for the ith country, f_j is the density for the jth element of the balance (i.e. production, imports, etc.) and x_j is the chosen value for this jth element. The x_j are chosen 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.

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-dSt>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[6].

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 to establish an equilibrium. 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

[6] For details, see section on Feed use estimation

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 tries to build on both the expected value of a variable and the accuracy with which it is measured. It then identifies a most likely combination of variables whereby every variable can only vary within its bounds. 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[7] 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 value of every variable and the final combination of all variables is determined in the constrained optimization process outlined above.

[7] The approach here would be to assume measurement errors to be adjusted for their FBS equivalents. Where empirical measurement errors are not available, all available metadata information is being used to provide proxies for measurement errors. To this end, a procedure has been developed that provides a detailed strategy to identify the measurement errors based on historical distributions, known support points and confidence intervals based on the hierarchy in the FAOSTAT flag system. Details are available from a separate paper focusing on measurement errors and their empirical values.

Figure 2 captures the basic set-up of expected values and associated distributions. It is based on the wheat balance for the US.

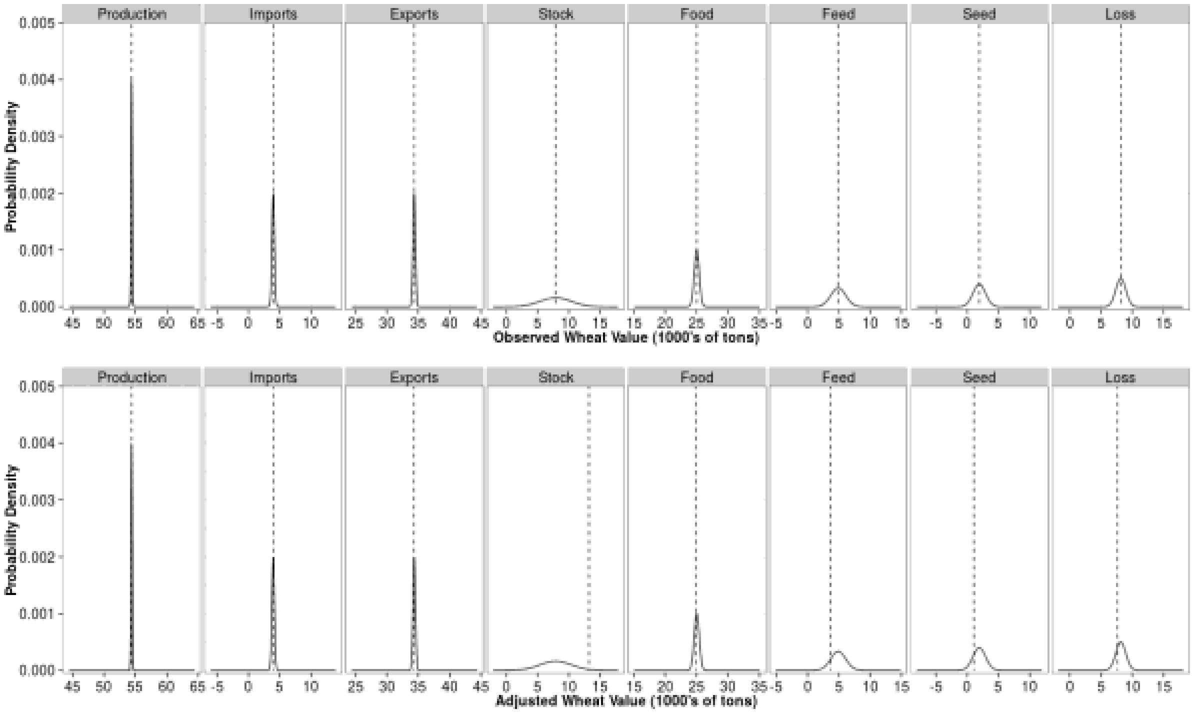


Figure 2: US wheat, unbalanced (top pane), balanced (bottom pane)

Pros and cons

The radical departure from traditional balancing provides numerous practical advantages and theoretical consistencies. But it also imposes greater computational burdens. 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, its year to year fluctuations are likely to be huge and its reliability and plausibility would be questionable. Such large fluctuations would not meet the a priori expectations 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.
- 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

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 arrive at these measurement errors: they can be derived directly from primary data, guessed from metadata that come with 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 out of home consumption of food[8]. In some cases, out of home consumption of individual households is included (food consumption in restaurants, cafeterias, street food, etc.) but in no case do these surveys cover consumption in public households/locations such as hospitals, military, prisons, etc.

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 former come from questionnaires, the latter e.g. from the UN Comtrade database[9] 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 that captures 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.

[8] For details, see section on food estimates
[9] For details, see section on trade

- For trade estimates, a fully balanced global trade matrix is provided, based on a new procedure that brings world exports and imports for every commodity into equilibrium[10]. 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 different sources and imputation methods in a complex and comprehensive metadata system and has, based on this system, developed a hierarchical system of flags. These flags indicate, for instance, whether a data point comes from an official source, whether it is a semi-official estimate, an FAO estimate (imputed data) and so on. Where measurement errors are unavailable, these flags have been used in general to establish measurement errors.

Five levels of confidence are distinguished as default values for measurement around an expected value. The highest confidence is placed on official estimates, which would be taken as point estimates without confidence interval, while the lowest confidence is placed on missing data. All confidence estimates are further modified by an allowance for the quality of a country's statistical system, which results in an overall hierarchy from official data coming from countries with excellent statistical systems 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	1.0	0.0%
Semi-official	0.9	10%
Imputed	0.85	15%
Estimated	0.6	40%
Missing	NA	NA

[10] For details, see trade section

Thus, for example, a Semi-Official estimate will have an implied measurement error of 10%, and this means that the standard deviation of the distribution for this element will be 10% of the element's estimated value.

Standardization

FBS and SUAs

In general[11], food balance sheets present all variables of a supply-utilization balance in their primary equivalents. For instance, in the balance for wheat all elements are expressed in wheat as a primary product, while the only variables that are readily available in their primary forms are production and, in principle, seed use of wheat. In many cases, information in terms of primary equivalents is not, or only partially available. Imports or exports of wheat, for instance, take place in the form of wheat, but also in the form of different wheat products such as flour (1st level of processing), bread or pastry (2nd level of processing) or even more processed forms.

Some variables of the balance may be exclusively available in their processed forms. Food of wheat, for instance, only exists in the form of flour, or in the form of flour products such as bread, noodles, pastry or biscuits. Wheat is practically never eaten as such, the same holds for all other cereals and indeed many primary products.

Given the fact that all variables of the balance (bar production) occur in practice in forms other than their primary one, all variables need to be converted back to their primary equivalents. This allows comparisons of the various variables with each other, and eventually the union of them together into a balance. Differently put, only if all elements are expressed using a common denominator can they be added up in the FBS balance. To this end, the FAO Statistics Division has developed a process known as "standardization", which is analogous to a process of creating a common denominator that allows processed products to be added up and expressed in their primary product equivalents. The different processing steps in the food chain create many processed products, which all need to be "rolled-up" into their primary equivalents. Supply and use of these processed products are also put into balances, which are here referred to as Supply Utilization Accounts (SUAs), at least by the FAO Statistics Division. Even if the naming convention reflects tradition rather than statistical or economic rationale, it has been maintained here for the sake of simplicity and continuity.

The new methodology has adopted the standardization method of the previous method in principle. It largely maintains existing processing streams and thus the same commodity tree structure. The standardization programmes have been re-written in the R language and a number of problems have been addressed in this process. Important changes have also been implemented in the parameters that link the processing levels, notably the so-called extraction rates. The same holds for other parameters, such as nutrient conversion factors and shares of uses (splitting utilization side).

[11] Some few products are specified as processed products, such as butter or vegetable oils.

Extraction rates (ERs). Extraction rates reflect the amount of primary product that sits in the next level of processed product; for instance, the extraction rate of wheat flour is, without going into the specificities of a country's milling sector, about 0.79. This simply means that 1 tonne of wheat that goes through a country's flour milling industry renders on average 790 kg of flour. Flour milling also produces a certain amount of bran, on average 180kg as well as a certain amount of wheat germ, e.g. 20kg. The remaining 10 kg are losses that occur in the milling process. A detailed account of changes in the used extraction rates is dealt with in a separate document; thus, it should suffice here to identify the two major developments that have caused changes in the ERs. The first is a shift of the definition in the product definition, brought about by the shift to the Central Product Classification (CPC). Generally, this rendered lower extraction rates. In the case of wheat, this is simply a reflection of the fact that the degree of purity in CPC is higher, i.e. wheat is defined without groats and pellets. The second is the recognition of the fact that some ERs are far from their expected values and not grounded in any empirical evidence. Such extreme values have been reviewed and, where necessary, adjusted.

Another important change in the use and calculation of extractions rates was that all extraction rates are now exogenously assumed. In the past, some extraction rates were implicit (endogenous), calculated from available information for processed production (e.g. flour) and an allocation of primary products for processing (e.g. wheat). While endogenous extraction rates are largely limited to developed countries and a limited number of products, this process often created implicit extraction rates with highly unlikely values. For Japan's cereal sector, for instance, the extraction rates were unduly low, suggesting that the country's flour milling industry is processing wheat in an inefficient way. In reality, the opposite is true. The reason for the low implicit extraction rate lies not in a low estimate for flour produced in Japan, but in the high allocation of wheat going into its flour milling industry. The new approach therefore is to keep flour production as a starting point, apply a reasonable extraction rate and calculate primary food (wheat) based on flour production and the given extraction rate. The added advantage of this process is a simplification in the standardization method and the possibility to get rid of a complicated ex-post reallocation process between processed products.

Nutrient conversion factors. Food balance sheets provide an account not only of quantities of food supplied and utilized, but they also provide an account of key nutrients, such as calories, protein and fat. Other nutrients could be added. As food products are typically eaten in processed form (rather than primary), nutrient conversion factors also apply to and are available for processed products. The new approach applies calorie conversion factors typically at the 1st level of processing. For the example of wheat, this means that the calorie, fat and protein conversion takes place at the flour level. The nutrient conversion for biscuits takes place when biscuits have been split up according to their shares, at the level of sugar, flour and vegetable oils.

Shares for different uses. In the current approach, the utilization side is constructed by applying fixed shares to availability. This applies to feed, waste, food manufacture and in part also to other variables. The new approach made an effort to determine all forms of utilization by the factors that drive them, rather than applying simple shares to availability. This means that shares in the previous methodology (e.g. for feed or waste) are no longer required. The shares of food that enter a manufacturing process outside the normal commodity trees are a notable exception.

Shares for food manufacture. For some FBS commodity trees, the underlying set of processing activities is rather diverse, resulting in many links between trees. These commodities are pruned off the main trees and crafted into the FBS as a separate balance. As separate branches, they re-enter the balances as processed rather than primary products. They include products such as sweeteners, beer, or other alcoholic beverages. During the process of food manufacturing and processing, these products often receive inputs from several FBS commodities. As an example, beer can and is being produced from many different starchy inputs, i.e. barley, maize, wheat, or sorghum, but also more exotic products such as bananas or plantains. In order to avoid double counting, they need to be identified as separate products in the FBS, and are not included in the original commodity tree.

SUAs and processed products

Every primary product is the basis for a broad variety of processed products and these are again the basis for the combination with other primary and processed products. A modern food-processing sector is indeed characterized by a vast variety of different products; the further one goes down the processing chain, the more possible combinations and varieties of food products there are. Information from the USDA database on food nutrients (ARS SR 27[12]), for instance, suggests that out of the 10 basic commodity groups (cereals, oilseeds, etc.) and the 60 primary FBS commodities, the US food industry produces more than 8000 different food products. These products can be consumed in the US but also be exported to destinations abroad. No doubt, not all 8000 processed food products are genuinely different in their food composition; many are simply different brand names with the same or similar combinations of first level food processing products. But the many different combinations suggest that the variety of food products is indeed large and that many more than 60 products are being eaten and traded. Trade information is available for about 3000 products. About 300 are genuinely distinguishable in terms of their composition and need to be converted into their primary equivalents.

SUAs and commodity trees

Standardization, i.e. the conversion and subsequent aggregation of processed products into their primary equivalents, aims to bring these products into a hierarchical order that reflects the various food processing chains in which primary products are converted into their processed equivalents. The most straightforward way to depict the hierarchy is in the form of commodity trees. The primary product, say wheat, represents the stem of such a tree. The main components of the first level processing (flour, bran and germ) form the main branches of the tree; eventually these split into ever finer twigs of higher levels of processing (bread, rolls, or dextrose). The hierarchy between various levels of processing captures the various steps of processing; the

[12] http://www.ars.usda.gov/Services/docs.htm?docid=8964

different levels and products are connected through the extraction rates (see Figure 3 below and other examples in section on classification).

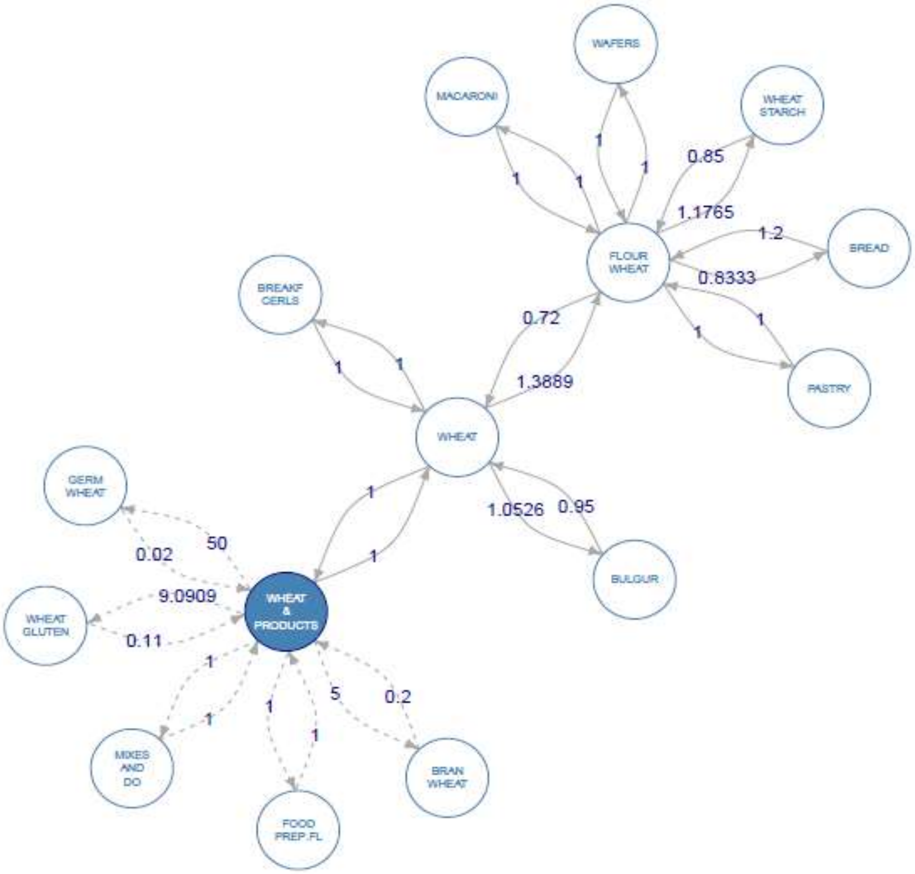


Figure 3: Standardization of wheat, processing flows

A single tree cannot capture the fact that modern food processing connects such trees by combining the processed products of one tree with those of other trees (refined sugar with flour and refined vegetable oils, see below), or processed products with primary products such as milk and eggs with flour and vegetable oils for other bakery products. This means that the trees are actually connected with one another through processed food products. The trees also provide connections at lower levels of processing, e.g. wheat or maize germ is connected to the balance of vegetable oils; likewise, bran, the other 1st level by-product of flour milling is connected to the feed balance (see section on feed use).

A review of the existing programmes revealed a number of issues. They included programming errors, breaks in the structure of the trees, hardcoded and undocumented exceptions, or as outlined above, inappropriate conversion factors. These problems become most visible when standardization leads to obviously wrong results, most notably negative utilization. Figure 4 below provides just one example of how wrong standardization has led to increasing negative utilization for Argentina's palm oil economy. Many similar cases exist. These are the visible signs of obvious problems; others are less readily identifiable, as they do not result in negative utilization. The current revisions tried to identify as many of these cases as possible and address them by using appropriate shares, identical commodity trees for different variables/elements of the balance and more reasonable extraction rates.

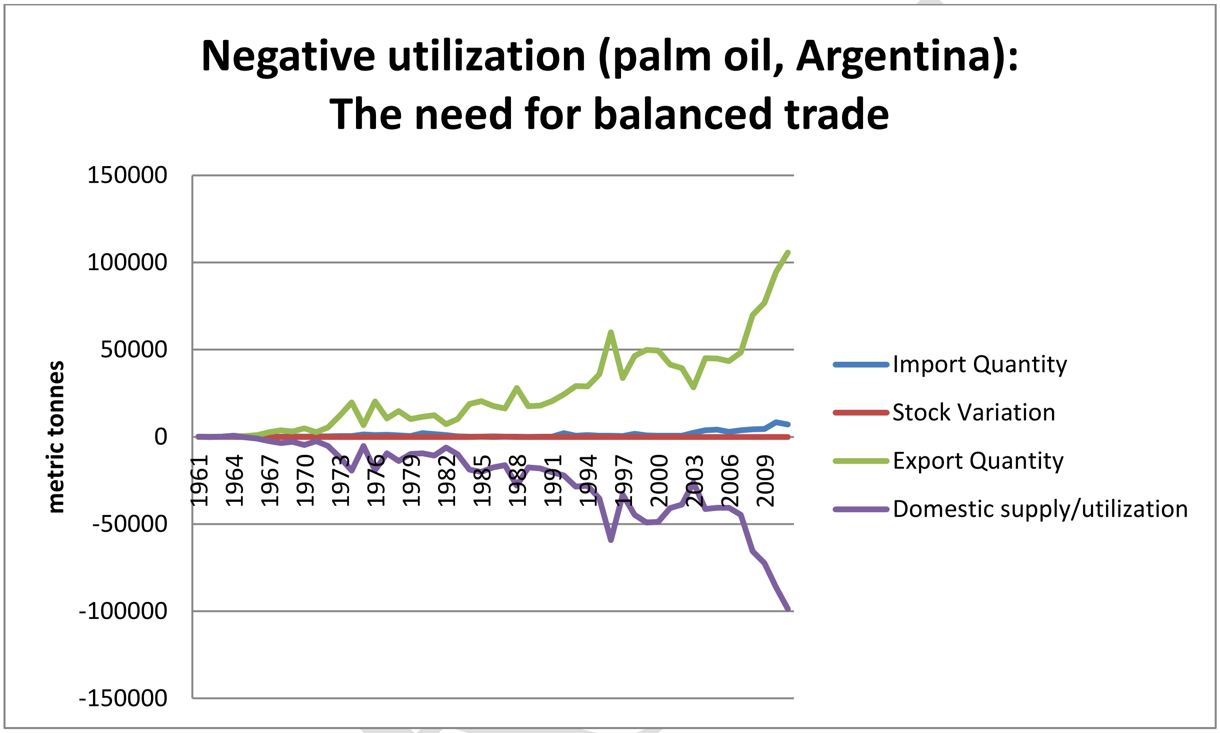


Figure 4: Standardization and negative utilization

Standardization and trade

Notionally, all variables of the balance are to be standardized, i.e. all underlying processed products are to be converted back into their primary equivalents. If, for instance, wheat is stored in terms of flour, these quantities would need to be converted back into their wheat equivalents for stocks. Indeed, where data are available, this is being implemented. In practice, however, the lion's share of information about processed products use and supply is limited to imports and exports. There are a number of aspects that deserve particular attention before trade can be standardized.

Product specificity. Products of the same name and classification but from different provenances have different characteristics in terms of nutrient contents and food values. Grass-fed beef from Argentina, for instance, is typically less caloric than corn-fed beef from the US. The same holds for many fruits and vegetables, other meats, and even cereals. There is a growing recognition of the need to account for these differences in the standardization process and in applying more specific calorie conversion factors. There are plans to account for these specificities in future versions of the SUA/FBS system. While there is awareness of the problem, these subtleties could not be implemented in the current revisions of the system.

The Standardization Algorithm

The specifics of the standardization process are outlined below:

1. We start with the primary commodities: wheat, barley, milk, etc.
    - If some elements of the balance are missing (i.e. our modules have failed to impute them), then these values should be imputed using the old FBS methodology (as a ratio of production, total supply, etc.).
    - Now, we must balance at this primary level. The details of the balancing algorithm have been provided in the preceding section.
2. After balancing, we have the amount allocated to food for processing for the primary commodity. This is being converted into production at the first processed level (flour, butter, etc.) if the values at that level are missing, otherwise the official data are used. This conversion is done in the following way:
    a. Some of the primary commodity may be consumed as such (e.g. eggs, groundnuts or soybeans for confectionary use), and hence the percent going to processing should be determined and a quantity then removed from food for processing.
    b. The shares are applied to allocate the amount processed into the various processed products.
    c. The extraction rate is then applied to convert the quantity of the primary commodity into quantity of the processed commodity. For example, suppose we want to create wheat flour production from wheat. We may have 90% of wheat being processed (and 10% left as such), 95% of processed wheat that is allocated to flour and 5% allocated to beer, and an 80% extraction rate of wheat to flour. Then, if we had 100 thousand tons of wheat, we would convert this to 100(90%)(95%)(80%) = 68.4 thousand tons of flour.
3. In the above step, all elements as specified by the commodity tree structure must be created. For example, bran and germ are also created when wheat is processed into flour. These are not elements with separate shares but should be thought of as by-products in the production of flour.
4. We now balance at the first processed level. However, we must standardize all further processed commodities back to this commodity in this balance in order to ensure we have accounted for any imbalances further down the commodity tree.
5. We can now compute nutrient information (calories, proteins, fats) from quantities.

6. After the calculation of calories, all quantities must be converted back up to their primary equivalents. We start at the lowest nodes and divide by extraction rates to compute parent quantities. Calories, on the other hand, can be added directly in the standardization process. However, there are several special cases/important notes:

   – With "by-products" (for example, wheat bran and germ) we do not standardize quantities as they are already accounted for in the main product standardization. However, standardization of calories/fats/proteins is performed for all products by adding the calorie/fat/protein values.

   – Some products (oils, juices, beers, etc.) can be created from multiple parents. In this case, the products must be rolled up into various parents, and the appropriate allocation to parents is not clear. Allocation is determined based on availability. However, in some cases we need to be able to specify that preference be given to certain parents. An example of this could be beer where preference should be given to barley over, say, bananas, wheat, etc.

   – Production is not be standardized. For primary products, production is already expressed in primary equivalents, for a processed commodity, production comes directly from food of the "parent" commodity, and so essentially it is already accounted for. All other elements need to be standardized.

Example

Consider a very simple example of the wheat tree. In this example, we would need distributions to perform the balances, but that is ignored for the sake of simplicity and balances are simply done arbitrarily to avoid complication. Also, we assume there is only production, imports, exports, food and waste. Dashes indicate unavailable data.

7. Initial Table

	Prod.	Imp.	Exp.	Food	Waste	Food for Processing
Wheat	90	20	10	-	5	100
Flour	-	30	5	-	0	0
Biscuits	-	0	10	-	0	0
Bread	-	0	10	-	0	0

8. Balance Wheat:

	Prod.	Imp.	Exp.	Food	Waste	Food for Processing
Wheat	90	20	10	-	**2**	**98**
Flour	-	30	5	-	0	0
Biscuits	-	0	10	-	0	0
Bread	-	0	10	-	0	0

9. Process to flour (assuming an extraction rate of 0.84):

	Prod.	Imp.	Exp.	Food	Waste	Food for Processing
Wheat	90	20	10	-	2	98
Flour	**82**	30	5	-	0	0
Biscuits	-	0	10	-	0	0
Bread	-	0	10	-	0	0

10. Create by-products (skipped for simplicity, but bran and germ should be created in this process).

11. Standardize bread and biscuits (using extraction rates of 0.5 and 0.65):

	Prod.	Imp.	Exp.	Food	Waste	Food for Processing
Wheat	90	20	10	-	2	98
Flour	82	30	**35**	-	0	0
Biscuits	-	0	10	-	0	0
Bread	-	0	10	-	0	0

12. Balance flour:

	Prod.	Imp.	Exp.	Food	Waste	Food for Processing
Wheat	90	20	10	-	2	98
Flour	82	30	35	**77**	0	0
Biscuits	-	0	10	-	0	0
Bread	-	0	10	-	0	0

13. Standardize to wheat (using the 0.84 extraction rate):

	Prod.	Imp.	Exp.	Food	Waste	Food for Processing
Wheat	90	20	10	**92**	2	98
Flour	82	30	35	77	0	0
Biscuits	-	0	10	-	0	0
Bread	-	0	10	-	0	0

	Prod.	Imp.	Exp.	Food	Waste	Food for Processing
Wheat	90	20	10	**92**	2	98
Flour	82	30	5	77	0	30
Biscuits	-	0	10	-	0	0
Bread	-	0	10	-	0	0

Summary of changes

Table 2 below provides an overview of the changes that have been implemented in the new standardization algorithm (as compared to the previous version):

Table 2: Overview of changes in the standardization procedure

	Previous Standardization Process	New Standardization Process
Extraction Rates	In the previous standardization routine, if production of a processed product was available then the extraction rate was computed using that production and the estimated value for food production. Also, extraction rates could be manually adjusted.	The computation of extraction rates based on processed product production and estimated input into processing is no longer performed. Instead, a reasonable extraction rate is applied and the input into processing (i.e. the estimated value) is updated. Also, manual updates of extraction rates need to be made explicit and be evidence-based.
Shares	Shares for standardization did not vary by country or commodity (Figure 4: Standardization and negative utilization depicts this problem).	Shares will now be driven based on availability of parent commodities.
Commodity Trees	Standardization is not homogeneous across all variables. Sometimes, production of processed products is standardized applying a different tree than for e.g. trade	All elements/variables are standardized applying the same underlying commodity trees.
Nutrient conversion factors	Regional averages for nutrient conversion factors	Country-specific conversion factors where available, regional conversion factors otherwise. Where no information is available, the conversion factors of USDA ARS are being employed (default)
Trade flows and product shares	No corrections made for differences in the composition of processed products in trade.	The new standardization process will implement trade-flow specific shares for the composition of processed products (not yet implemented in the current version).
Explicit assumptions	The previous standardization procedure made many assumptions that were undocumented and unclear.	In the new standardization procedure, we are explicitly stating all assumptions we make and thoroughly documenting the procedures implemented. This will provide more clarity and transparency into the process.

Production

Data collection

The FAO Statistics Division collects data on agricultural production via an annual questionnaire. The questionnaire is dispatched to 180 countries, including to EUROSTAT and EU member countries separately. Not all countries return the questionnaire. Some countries fail to compile the questionnaire but make all their data available on Internet sites of their respective/responsible statistical office (NSOs, MoAs). Others compile the questionnaire, but leave large N/A gaps, which means that they have not collected the data in most cases[13].

In order to improve availability and comparability, data are also collected from other sources such as National Official Publications (general and agricultural yearbooks, monthly bulletins) international data bases (EUROSTAT, OECD), or from FAO and UN reports based on missions to countries. In the recent past, the FAO Statistics Division also organized ad-hoc workshops on five continents to present the FAO data collection process to the countries, and to solicit their help in enhancing the overall data reliability, as well as in strengthening data timeliness and the coherence of the series. As a result, the number of questionnaires returned increased sharply, in particular in Africa, where the response rate doubled. It should be noted that the main producing countries provide data through the annual questionnaire and that, as a consequence, a large share of agricultural production in FAOSTAT is based on official figures.

Table 3 below provides an overview of the return rates of the production questionnaire over the past 10 years. It also includes an indication of the completeness of the questionnaire for those countries that return the questionnaire to FAO.

Table 3: Response rates to the FAO questionnaire.

	2005		2006		2007		2008		2009		2010		2011		2012		2013	
	RC	RR	RC	RR	RC	RR	RC	RR	RC	RR	RC	RR	RC	RR	RC	RR	RC	RR
Africa	8	15	16	30	20	38	18	34	14	27	30	58	26	50	32	60	32	60
America	15	36	17	41	12	28	15	35	15	36	16	38	17	40	16	50	17	53
Asia & Oceania	15	25	27	45	23	38	25	43	33	55	40	66	38	63	36	65	35	64
Europe	35	69	31	61	32	64	30	58	33	65	30	59	34	66	32	70	34	72
World	73	40	91	50	87	48	88	49	95	52	116	64	115	64	116	65	118	66

RC: Number of responding countries, RR response rate expressed in %

The production questionnaire includes four basic variables. The first rubric includes information about activity levels. In the case of crop production, these are area harvested and where available,

[13] Most of these countries are characterized by weak statistical systems. The Global Strategy to Improve Rural and Agricultural Statistics provides a detailed account of the state of statistical systems around the world. Alas, most countries with weak agricultural systems are also countries where other information (surveys) suggest that they are exposed in a particular way to food security issues.

area sown. In the case of livestock production, these are herd sizes, i.e. the number of animals kept. The second is output levels, i.e. production levels in the case of crops, meat, milk, eggs, etc. and production in the case of livestock. The third rubric contains productivity measures, i.e. yields and cropping intensity for crops and slaughter weights as well as offtake rates in the case of livestock. It also includes a fourth rubric, which can be described as "other, direct uses" of production, i.e. seed quantities for crops and breeding animals for livestock, as well as losses and waste.

On the basis of the replies to the production questionnaire, the FAO Statistics Division undertakes further research and data work. The focus of this work is to provide Quality Control and Quality Assurance (QC/QA). This works also includes a vast number of other operations, e.g. adjusting units of measurement to those used by FAO, weeding out outliers, transcription errors, filling in obvious gaps, and so forth. Under the SWS, all QC/QA work will be undertaken under the auspices and guidance of the corporate Statistical Quality Assurance Framework (SQAF)[14].

The review process also includes further efforts to fill data gaps and to complete the information for the production and FBS domains. For instance, many countries only provide information about two of the first three rubrics, e.g. area harvested and production[15]. FAO Statistics Division calculates the missing variables (i.e. yield) by applying the simple identity that productivity=production/activity.

Data imputation: a new approach

Once the QA/QC has been undertaken, data issues have been weeded out, and data gaps have been filled with information available from other sources[16], the remaining data gaps are filled by an imputation procedure that harnesses all information available at this point. To this end, a new imputation method has been developed, which combines a powerful econometric approach with expert knowledge and the experience of production officers and clerks.

The new imputation method not only harnesses all available data and expert knowledge, it also incorporates experience with imputation methods that have been used previously, or that are employed by other institutions. A central insight from the review of existing methods is that no single approach/model is able to render satisfactory results across all countries and commodities of the production domain. Based on this insight, an ensemble learning model was developed,

[14] http://www.fao.org/docrep/019/i3664e/i3664e.pdf

[15] The FAO production definition includes production for own consumption. While this may not be a major concern in highly developed economies (US/Europe) it is a major issue for countries where food insecurity looms large. These differences in the definition require further adjustments and are undertaken by the FAO Statistics Division in the course of quality control and assurance (QA/QC).

[16] In addition to the websites of NSOs and MoAs, the FAO Statistics Division also harvests data from authoritative sources, which often specialize on one commodity. These include Oil World, ISO, ICO, ICAC, etc.

tested and adjusted to the needs of the production domain. Tests and application have shown that the particular strength of the ensemble model is in successfully handling the great heterogeneity that characterizes the time series of the production domain.

A detailed description of the methodology is available in a separate document. Here description is limited to a few main features. First, the imputation of yield in any given country incorporates available information from notionally all countries and over multiple commodities. This enables maximization of information usage and the improved stability of the imputation. Second, rather than relying on a single method or model, the imputation process consists of a dozen candidate models, which are averaged, in the ensemble, with weights assigned in accordance with the predictability of each model. Third, an ensemble can cope well with sparse data availability. For instance, when and where only a single observation is available, the value can be carried across many years and be assigned to the missing observations successfully. On the other hand, if abundant data are available, more weight will be given to the sophisticated models, which have the ability to capture the complex pattern of the data[17].

The performance of the new methods (imputation and validation) has been tested and reviewed; both the quantitative error and qualitative assessment suggest that the new method is superior to previously employed methods and approaches taken by other institutions.

The new imputation approach in detail

There are three main steps to the imputation procedure:

1. Productivity Imputation
2. Imputation by balancing
3. Production Imputation

Productivity imputation

Productivity imputation is performed via the ensemble approach alluded to above. Currently there are 10 models implemented in the ensemble framework. A separate model is fit to each individual country and commodity pair unless otherwise noted.

1. Mean: This model computes a simple mean of all observations and imputes this value.
2. Linear: A linear regression model is fit to the available data, and predictions from this model are used to estimate missing values:

[17] In addition, a statistical learning algorithm was deployed based on the history and the current set of values. Historical values which were overwritten were perceived unreliable. In this case, the new algorithm learns a discerned pattern and validates new values. Five different models have been implemented thus far, whereby each model has a vote to ultimately decide on the severity of a given value. If a single model flags the value, then it will have a severity of 1 and requires minor attention; on the other hand, if all 5 models flag the value then a severity of 5 is assigned; thorough review is required to ensure this value is valid.

$$P_{i,j,t} = \beta_{0,i,j} + \beta_{1,i,j}\, t + \varepsilon_{i,j,t} \qquad \text{(Equation 6)}$$

3. Exponential: A regression model is fit but the exponential of time is used:

$$P_{i,j,t} = \beta_{0,i,j} + \beta_{1,i,j}\, e^{t} + \varepsilon_{i,j,t} \qquad \text{(Equation 7)}$$

4. Logistic: A non-linear regression model is fit to the data:

$$P_{i,j,t} = A_{i,j} + \frac{B_{i,j}}{1+exp(-C_{i,j}(t-D_{i,j}))} + \varepsilon_{i,j,t} \qquad \text{(Equation 8)}$$

   If the non-linear regression model fails to converge, then $A_{i,j}$ is assumed to be 0 and a new non-linear model is fit to the data. If that model also fails to converge to a solution, then $B_{i,j}$ is assumed to be the largest observed value. In this case, a logistic regression (which will always converge) is possible by performing a logit transformation, and this is model is then used.
5. Naive: Missing values between two observed values are interpolated linearly. For observations outside the range of the observed data, the nearest observation is carried forward or backward.
6. ARIMA: Several Autoregressive Integrated Moving Average time series models are fit to the data, and the best one is selected based on the AICC. Then, imputation of new values from this model is done via Kalman Filter smoothing.
7. LOESS: A local regression model is fit using linear models as the base learners. The model window varies based on sample size, and thus more flexible models are fit when more data is available.
8. Splines: A cubic spline is fit to the observed values and used for interpolation.
9. MARS: A Multivariate Adaptive Regression Spline, which is a mathematical model appropriate for piecewise linear time series, is fit to the observed values.
10. Mixed Model: This model is fit to all countries at once, but still restricted to just one commodity at a time. The linear mixed model estimates production as a smoothed function of time, and the time trend is considered as a random effect by country.

Each of the models are fit to the available data, and an estimate of the model's ability to explain the data is produced via cross-validation. In other words, each model is fit to all the available data except for a handful of observations. The performance of an individual model is evaluated by its ability to estimate these known observations. This process is repeated with different groups of observations so as to provide a good estimate of a model's ability to fit the data. This ability is captured in the form of an error term, $e_{i,j,k}$, which depends on the country i, the commodity j, and the model k. Then, each model is assigned a weight according to its ability to fit the observed data (with better models receiving larger weights) subject to the constraint that the weights sum to one:

$$w_{i,j,k} = \left(1/e_{i,j,k}\right) / \left(\sum_{l=1}^{10} 1/e_{ijl}\right) \qquad \text{(Equation 9)}$$

Then, a missing observation is imputed using a weighted mean of the estimates of the ensemble models at the missing time. One additional constraint is imposed: models which may not extrapolate well (such as an exponential fit) are removed from the weighted average for missing values that lie too far outside the range of observed data.

Imputation by balancing

Once productivity has been imputed, we proceed by balancing. If both productivity and activity values are available, we impute production so that it satisfies the identity production= productivity*activity.

Production imputation

For imputation of production, we follow the same procedure as outlined in step 1. We apply this procedure instead to the production data, and impute values using a new ensemble fit to the currently available production data.

Example

Figure 4 depicts an example of the imputation methodology applied to a sample dataset. Each box represents a different country, and each line represents a different model in the ensemble. The thickness of the line corresponds to its weight in the final ensemble, and the "x" marks indicate the ensemble (as opposed to the dots, which represent official data).

The top left graph shows a dataset with an extreme value in 2006. This value causes problems for several models (Loess and Arima), but those models don't receive much weight. Instead, the imputations are roughly imputations between the known observations (but with a bit of oscillation introduced by the extreme value). The top right graph shows that the ensemble model is able to fit simple trends well. The data don't suggest any strong patterns, and so a simple regression dominates the ensemble and takes most of the weight.

In the bottom left graph, the official data follow a very strange pattern (close to 0 in 2005 followed by a huge spike in 2006 and then a slow decline). Because of this wide variability, the ensemble gives most of the weight to the mean of the observations, but also smoothes the approach to this global mean slightly. In the bottom right graph, we see official data which follows some complicated trend. The ensemble approximates that trend well. By design of the ensemble and weighting procedure, we can be confident that the ensemble is a good approximation of the missing information across different countries and products.

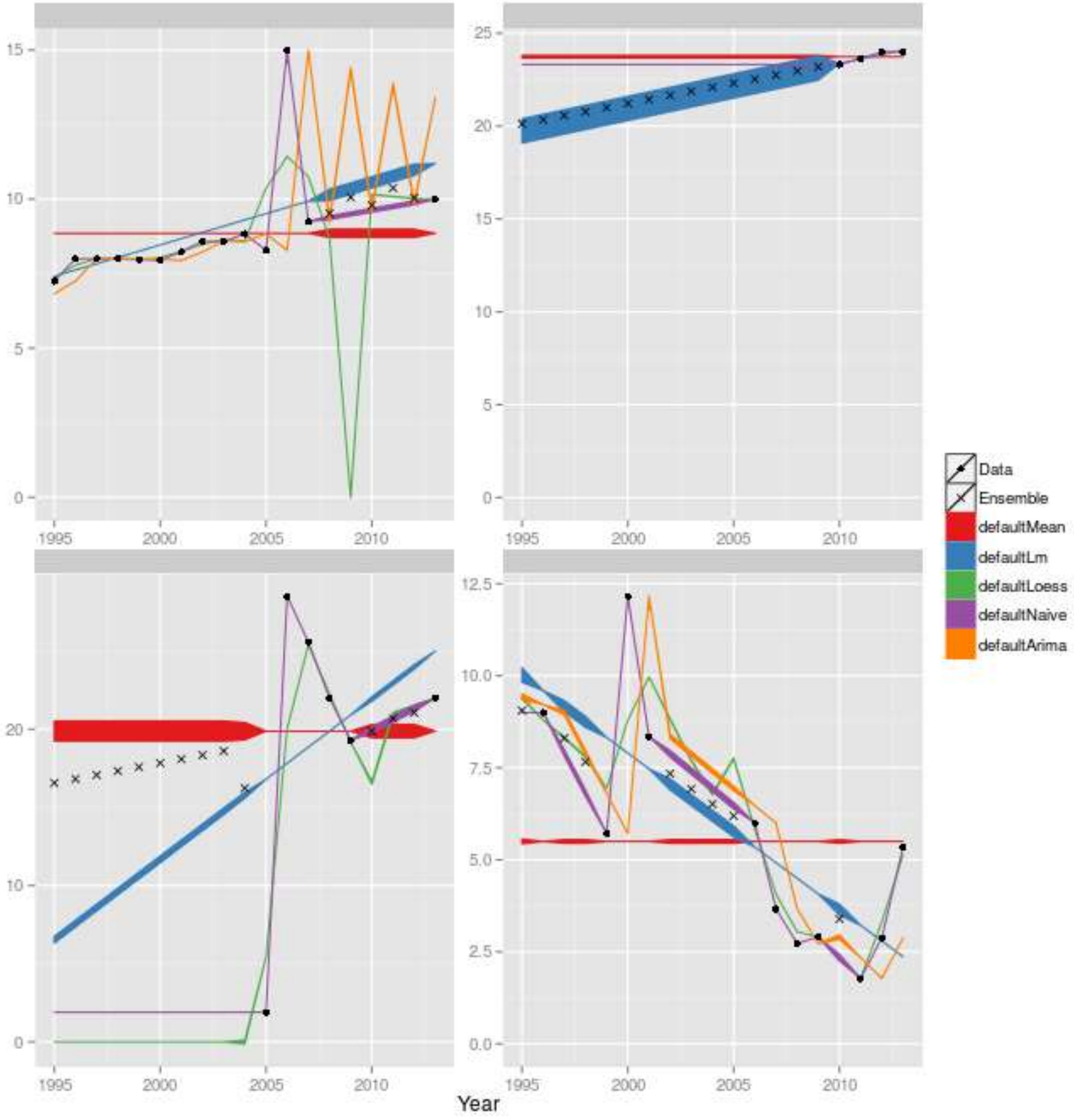


Figure 5: Imputation methods and results

Table 4: Response to the FAO production questionnaire in 2013

Country	Total Production Cells*	Production Cells* with officially reported data	Share of official Data
Europe	7873	4607	59
Africa	6674	1299	19
America	6275	2301	37
Asia	8112	2819	35
Oceania	1461	143	10
World	30395	11169	37

*A cell denotes a data point for yield, area, or production

Trade

Introduction

In the new statistical working system (including the new FBS module), all trade data come from the United Nations Commodity Trade Statistics Database (UN Comtrade). UN Comtrade provides access to standardized data from its website at http://comtrade.un.org/. It also offers a detailed description of the methodology used in compiling trade data (http://unstats.un.org/unsd/trade/methodology%20imts.htm and in http://unstats.un.org/unsd/trade/imts/imts_default.htm). Most importantly, it provides guidelines for the compilation of International Merchandise Trade Statistics (IMTS). In view of the fact that all methodological aspects are readily available from UN Comtrade, no effort has been made here to repeat or summarize any of the underlying methodological aspects relating to the compilation of trade information by UN Comtrade. All aspects presented hereafter refer to the work that is undertaken by the FAO Statistics Division, i.e. once all necessary data have been downloaded from UN Comtrade database.

UN Comtrade data are available at different levels of disaggregation, all levels of disaggregation refer to the internationally recognized standard Harmonized System (HS) classification. The HS classification is maintained by the World Customs Organization (WCO) and updated in intervals of 5 years[18]. The latest version is HS 2012, preparatory work for the next version (HS 2017) is already underway. FAO has made numerous contributions to the improvement and extension of the HS system in the past to include essential data and reflect changing patterns in agricultural and food commodities, and is actively involved in the development of the HS2017.

The HS system is a hierarchically structured classification of products and distinguishes different levels of disaggregation. The first 6 digits of the HS system are common to all countries, higher levels (that can go up to 12 digits) reflect national extensions. To ensure comparability across countries, it was decided to harvest trade data at the 6-digit level. The 6-digit level is also sufficient to ensure consistency with the classification used for other variables of the FBS system, i.e. the classification system applied to production data and all forms of utilization (Central Product Classification (CPC), version 2.1 expanded). For details of compatibility between HS and CPC, see section on classifications of this document.

Processing UN Comtrade data

UN Comtrade data are processed by the FAO Statistics Division in different ways and for different purposes. The main purpose is the creation of specialized trade information (see overview in Table 5 below) but there are also other needs, including the FBS/SUA, as well as many other datasets such as trade of fertilizer and pesticides, machinery, etc.

[18] http://www.wcoomd.org/en/topics/nomenclature/overview/what-is-the-harmonized-system.aspx.

The processing entails different steps and ultimately leads to different products. The various steps and products are described in detail in the methodology paper for the compilation of trade statistics. At this juncture, it should suffice to list the various steps. The resulting products are summarized in Table 5 below.

Processing steps

i. Direct harvesting data from UN Comtrade via an Application Processing Interface (API). For details see trade documentation

ii. Data validation and error correction.

iii. "Mirroring", i.e. completing the trade matrix with information from trading partners/reports wherever available

iv. Calculating representative import unit values (IUV) and export unit values (EUV) (from reporter and/or partner information) and based on these, calculate missing quantities from available values and missing values from available quantities.

v. Calculating total trade (total imports and exports for every commodity at HS 6 level) flows, which are potentially still imbalanced.

vi. Harvesting and/or manual input of data from specialised commodity reports

vii. Converting HS codes into CPC codes

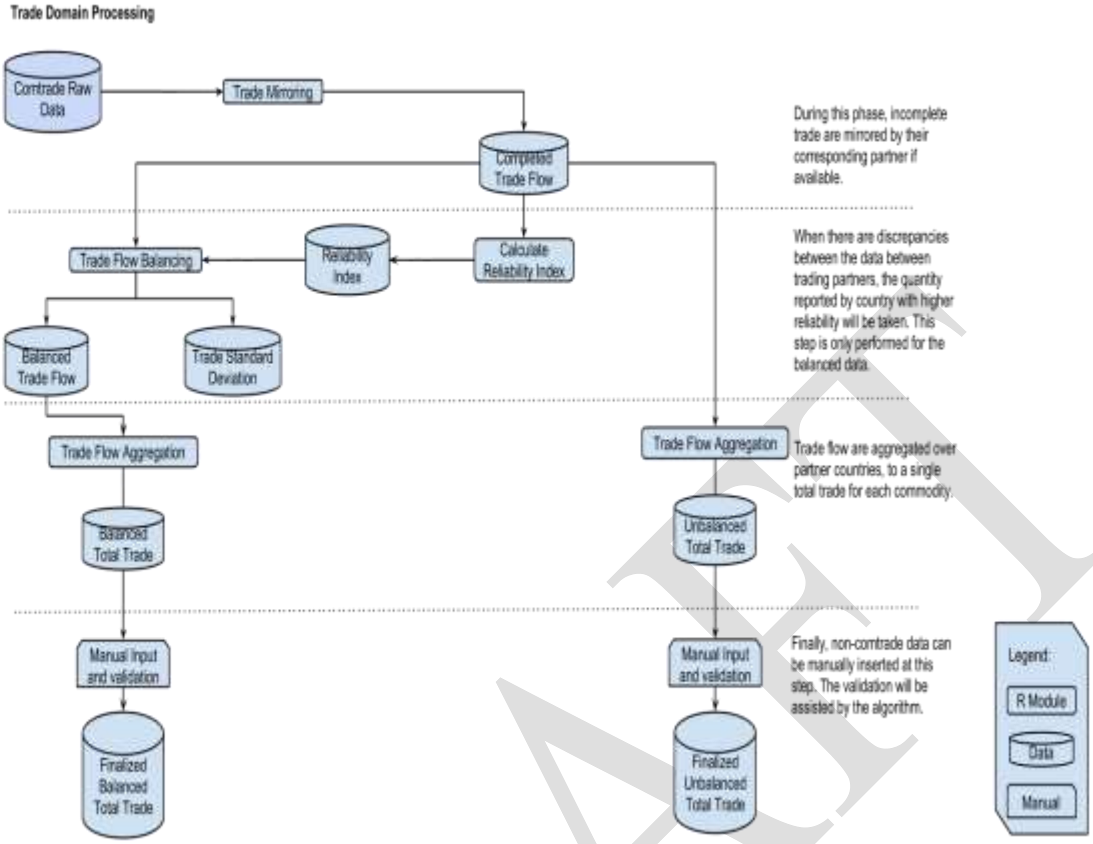
Up to this point, all imputation methods keep officially reported data intact (apart from weeding out obvious reporting errors and filling in data gaps). All imputations supplement existing data, but do not replace them. The trade information that best serves the FBS needs however is trade that is as complete, as balanced (imports vs. exports) and as consistent as possible at the global level. This requires not only imputation of missing data from non-reporting countries, but also adjustments of the trade matrix to ensure that global imports are as equal to global exports as realistically/statistically possible. Balancing the matrix may however involve overwriting official data and thus requires a separate process. This is done applying a system of endorsement factors, essentially reflecting the revealed reliability in the reporting of trade data. Again, the details of the methodology are described in the documentation of the new trade methodology. Here it shall suffice to list the key steps:

i. Calculating endorsement factors

ii. Balancing trade so that $\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} I_i$ for every commodity j and all countries i.

iii. Calculating total trade (total imports and exports for every commodity at HS 6 level), now of balanced flows

Figure 6 below provides an overview of the new trade processing flow.



During this phase, incomplete trade are mirrored by their corresponding partner if available.

When there are discrepancies between the data between trading partners, the quantity reported by country with higher reliability will be taken. This step is only performed for the balanced data.

Trade flow are aggregated over partner countries, to a single total trade for each commodity.

Finally, non-comtrade data can be manually inserted at this step. The validation will be assisted by the algorithm.

Figure 6: Trade processing, work flows

Dissemination products

The trade processing ultimately produces **six** datasets for dissemination through FAOSTAT, three of them are trade matrices, and another three are files with total trade only. The various dissemination products are summarized in Table 5 below.

Important for the FBS balancing system is the fact that every total trade flow comes with a reliability measure, i.e. a surrogate of a measurement error. This measurement error is defined as follows and is calculated during the last step of the trade processing, i.e. when calculating the balanced flows. It contains information about the completeness of the trade information as well as information from the reliability factors (endorsement, see below).

Table 5: Trade products

Dissemination products	**1. UN Comtrade**	**2. Mirrored flows**	**3. Balanced trade**
Matrixes of flows	Cleaned for errors and misreporting.	Cleaned for errors and misreporting	Cleaned for errors and misreporting
	No data overwritten	Data gaps filled thru mirroring	Data gaps filled thru mirroring
	No data added	IUV/EUV calculated and data gaps filled thru imputation (Q/V and V/Q)	IUV/EUV calculated and data gaps filled thru imputation (Q/V and V/Q
		Official data kept intact	Official over-written, where necessary
			Balanced trade, endorsement factors
Total trade	**4. Imbalanced official**	**5. Mirrored and imputed, but still imbalanced**	**6. Balanced trade**
	Total exports and imports are reported by country, but	Total exports and imports are reported by country, but	Total exports and imports are reported by country, and
	$\sum_{i=1}^{n} X_i \neq \sum_{i=1}^{n} I_i$	$\sum_{i=1}^{n} X_i \neq \sum_{i=1}^{n} I_i$	$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} I_i$

Calculating export and import unit values (EUV/IUV)

The overview provided in Table 5 identifies export unit values and import unit values as the key factors to impute missing quantities from existing value flows, or missing value flows from existing quantity flows. The process of imputing as such is straightforward; the choice of the most appropriate unit value is not. In theory, there are competing alternatives, their pros and cons can only be evaluated in an empirical/practical context and depend on the reliability of reported trade flows, number of products lumped together into a given HS category, or the number of available trade flows over reporters/partners.

The right choice is therefore the answer to an empirical question. Exposing various theoretical alternatives to practical tests resulted in the following procedure. Where the number of trade flows for a given product is low, typically 5 or less, a global average (mode) is taken as the basis to calculate the unit values. Conversely, where and when the number of flows is high, country-specific unit values are identified to calculate unit values. Figure 7 depicts the country-specific case.

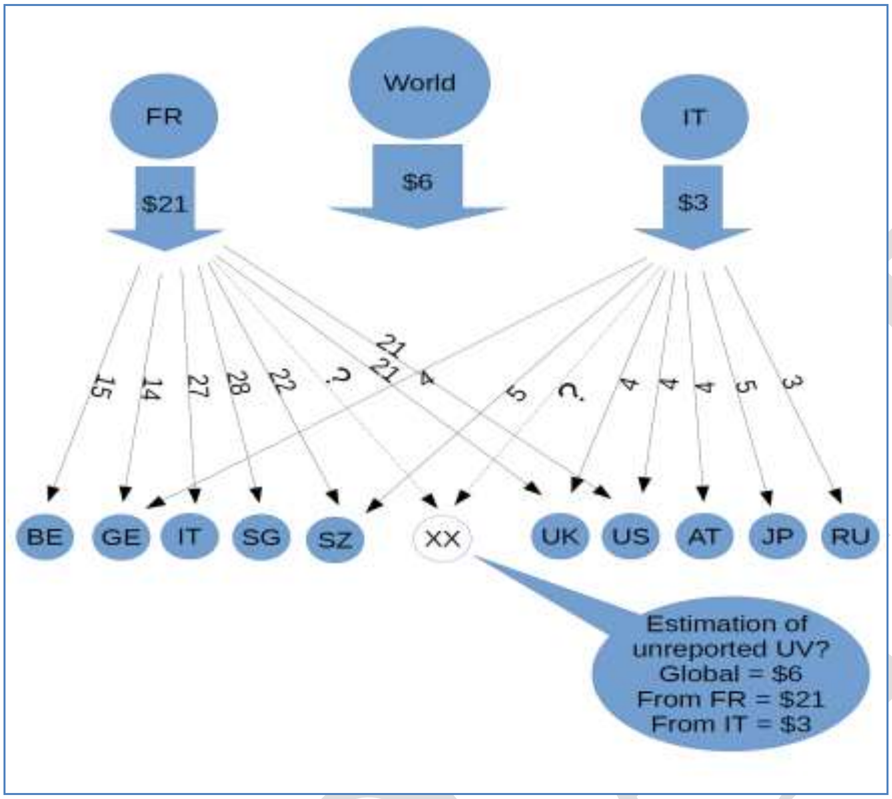


Figure 7: Calculation of unit values based on trading partners

In the concrete case captured in Figure 7, country XX importing wine from France, but without quantity or value information, is assumed to pay the same price as all other countries importing wine from France, i.e. $21/bottle on average. Likewise, if the same country XX imports wine from Italy, the underlying unit value is based on the one that other countries pay on average when importing wine from Italy, i.e. $3/bottle.

When enough flows are available, such a country-specific approach renders a more precise outcome than working with a global average, which would have been $12/bottle. In the example provided, it reflects the fact that France typically exports expensive wine and is therefore assumed to export also expensive wine to the country XX, where information is missing. Likewise, it also captures the fact that Italy typically exports “cheap” wine and thus assumed to sell the same quality to country XX, where that information is missing.

Path dependency of imputation routines

As outlined above, the process from “raw” trade data to published trade results encompasses several distinct steps. In general, these steps follow a logical and cogent sequence. For instance, cleaning data and weeding out errors is done before other processes such as imputation or balancing can take place. There is, however, also rationale for more ambiguity in the sequence of the various steps and not all steps follow a cogent and unequivocal logic. What is more, there is no a priori (conceptual) reason to prefer one step over the other in the sequence of the implementation, e.g. to undertake the mirroring step before the unit value imputations or vice versa.

Intuitively, one may be inclined to first mirror missing data so that the unit value calculations are being put on a broader footing. As outline in the section on the use of unit values, such a decision would have direct implications on the choice of the unit value, whereby more observations will eventually lead to a decision as to whether country-specific or global unit values are employed in the imputation process. The intuition is likely to be confirmed when the mirrored data for a country are at about the same level as the average unit value of a given reporter/partner. If, however, a mirrored flow is far out of the normal range, this will also affect and indeed distort the unit value used in the next step of the process and, as these unit values would be applied to all flows with missing information, many imputed data would be affected.

Without prior knowledge of the correct sequence of steps, the final results become path-dependent on the sequence of imputation steps. To gauge the magnitude of the potential effect, a number of empirical tests have been conducted. The results suggest that, at least in general, the differences that arise from different imputation sequences are negligibly small. There are, however, a few commodity specific exceptions. Particularly where flows are characterized by large product quality differences across reporting countries, poor product differentiation in reporting, etc., path dependency can influence the results in a significant way. Overall, however, such commodity-specific differences were not deemed important enough to deviate from the proposed sequence of data processing steps outlined in Table 5 above.

Reliability index and endorsement factors

Overwriting official data cannot be done on an arbitrary basis. It requires a rules-based and well informed process. The choice taken in the literature is to base this process on the reliability of reported trade data. The same approach is being pursued here in principle, however, with important modifications.

In order to quantify the reliability of a country's reported trade, we can utilize trade flow data, and in particular the level to which reported values agree with the reported values of trading partners. However, this approach can be complicated. For example, suppose country A reports different values than two of its trading partners B and C. Should A be penalized equally for both disagreements? If B is a highly reliable reporter while C is a highly unreliable reporter, than we would want to penalize disagreement with B more than disagreement with A.

1. Our solution to this problem is as follows: Compute the level of agreement between each pair of countries as the proportion of trades on which they agree (within some tolerance).

2. Initialize all countries to the same reliability score.
3. Each country will now transfer its reliability score to all its partners in proportion to the level of agreement with these partners. For example, suppose country A has a reliability score of 1 and trades with countries B and C only. Suppose also that A and B agree 90% of the time while A and C agree only 30% of the time. Then A will transfer 90% / (30% + 90%) = 75% of its score to country B and 25% to country C.
4. Step 3 is repeated until the change in the reliability scores becomes negligible.

This algorithm assigns larger reliability scores to countries that tend to agree with their partners, and the score also accounts for the reliability of the partner country. This algorithm is known to converge to a solution rapidly, and it is also used in many other applications for measuring quality within a network (most notably, perhaps, is Google's PageRank algorithm for determining quality of websites).

Example of reliability scores

The reliability index calculations are depicted in Figure 8. The nodes (circles) represent different countries, and the thickness of the edges indicates the agreement between the corresponding countries (with thicker edges indicating stronger agreement). In the first step of the algorithm (the leftmost plot) all countries have an equal reliability score. After one iteration of the algorithm (the middle plot), the top nodes/countries have a higher than average reliability and the bottom two countries/nodes have a below-average reliability. After the second iteration (the rightmost plot) the algorithm approximately converges: the top three countries all have high reliability scores (as they all agree with each other). Country 4 has a slightly higher reliability than country 5 because country 4 and country 2 agree on some trades. However, country 4 and 5 both receive low reliability scores, even though they happened to agree between themselves.

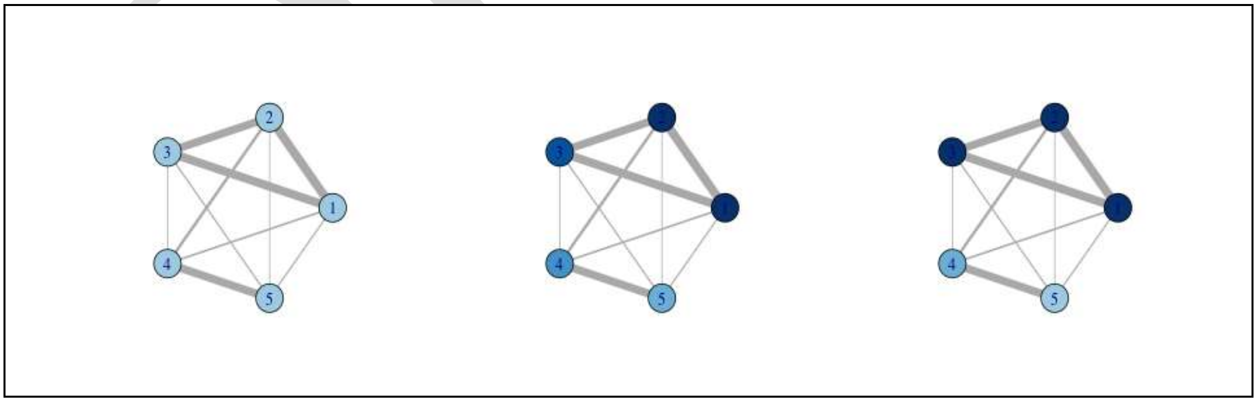


Figure 8: Reliability index calculation

Stocks and Stock Changes

Introduction

Stock estimates are currently not included in any of the FAOSTAT domains. Neither is there a separate domain that brings such estimates into one place (e.g. analogous to production or trade), nor is there a systematic inclusion of stocks in any other data domain. The only domain where stocks surface are the commodity balances, and the food balance sheets are a subset of these balances. Even there, no levels of stocks, but only year-2-year changes in stocks are available.

The main reason for low coverage lies in the fact that such data are seldom collected and where they are, data are not always available. However, as stocks or at least stock changes are an integral part of every supply-demand balance, estimates for stock changes have been included in the database. In many instances, changes in stocks functioned as a balancing item.

Using stocks as a pure balancing item, however, implies (as for any other element that is used as a balancing item) that all measurement errors are relegated to stocks (or the chosen balancing item). This also means that stock changes no longer only capture changes in stocks, but also function as a catchall for all measurement errors and would better be referred to as stock changes and residual uses. This is an undesirable outcome for any element of the balance but given the importance of stock for e.g. price volatility analysis, it would seriously diminish its value as a statistical indicator. Moreover, stock changes would "inherit" errors from previous years, resulting in steadily increasing distortions over time.

Imputation/Estimation

The need to move away from a residual approach poses the challenge of identifying an alternative method to generate an estimate or rather an expected value and a distribution. If empirical stock estimates are available (the US for instance undertakes a bi-annual survey of its cereal stocks) these would enter the balance as an observed value, ideally even with a measured distribution. Clearly, this is the first best solution and should be encouraged for a maximum number of countries and commodities.

If no information is available about stock changes, a distribution would need to be assumed. However, even assuming a distribution such as a uniform distribution (shown below in Figure 1) is problematic: we are assuming that a stock change near the assumed min is just as likely as a stock change close to 0; however, a stock change just outside of the boundaries has zero probability. Moreover, uniform distributions can be problematic in the balancing stage, as such elements essentially become residuals. Thus, it is very important to construct some distribution for stocks, even if it is very wide to indicate that there is much variability in the estimate.

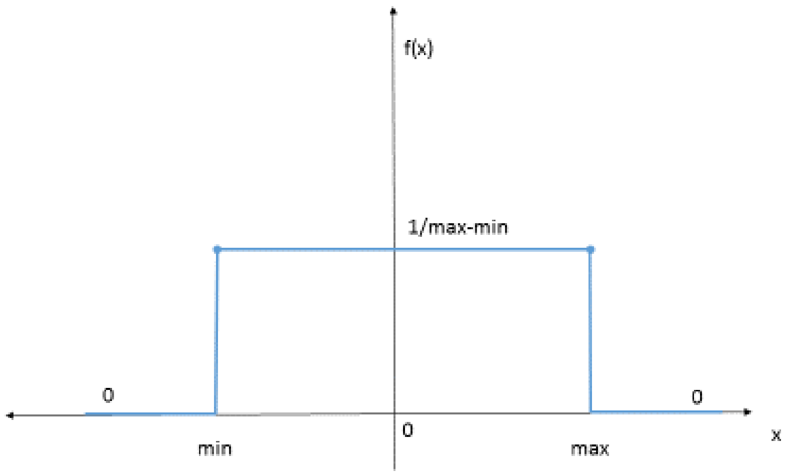


Figure 9: Bounded uniform distribution with a zero mean

The strategy here, however, is to harness additional and readily available information inherent in stock holding practice and economics. Such prior information is available e.g. from knowledge about the costs of stock accumulation and reduction over time. This prior information can be harnessed to move away from a uniform distribution in which every stock level would have the same probability between a maximum (max) and a minimum (min) to an approach that makes different levels of stocks more likely than others, still within the limits of (max) and (min).

The prior information about the economics and dynamics of storage can be used to derive both an expected value and a distribution.

a. Expected value. The expected value for stocks should be zero in the long run; empirical information fully confirms this a priori expectation, and figure x and y capture this for maize and wheat stocks in the US. The reasons for the long-run mean reversion to zero are obvious. Any long-run positive deviation from zero would amount to accumulation of stocks, which would rise the longer such positive shifts prevail and the higher the positive values are. Conversely, successive negative deviations would amount to a permanent drawdown of stocks and thus be tantamount to an eventual stock-out, or imply unlikely (very costly) high initial levels of stocks. The reasons for a non-zero short-term stock change lie in the ability of smoothing fluctuations of consumption. The desire to keep consumption stable means that stocks will function as a short term buffer to smooth surpluses and deficits, i.e. demand for stocks is, at least at high levels of stocks, very price elastic while consumption is not.

Distribution

The above implies that positive and negative stock changes are likely to be symmetrically distributed around the zero mean. The analysis for US wheat and maize suggests that the empirical distribution could be approximated by a normal distribution with a zero mean (see Figure 10 as well as Figure 14 and Figure 15, Chi-square metrics highly significant). Thus, we have already made improvements beyond assuming a uniform distribution for stock changes.

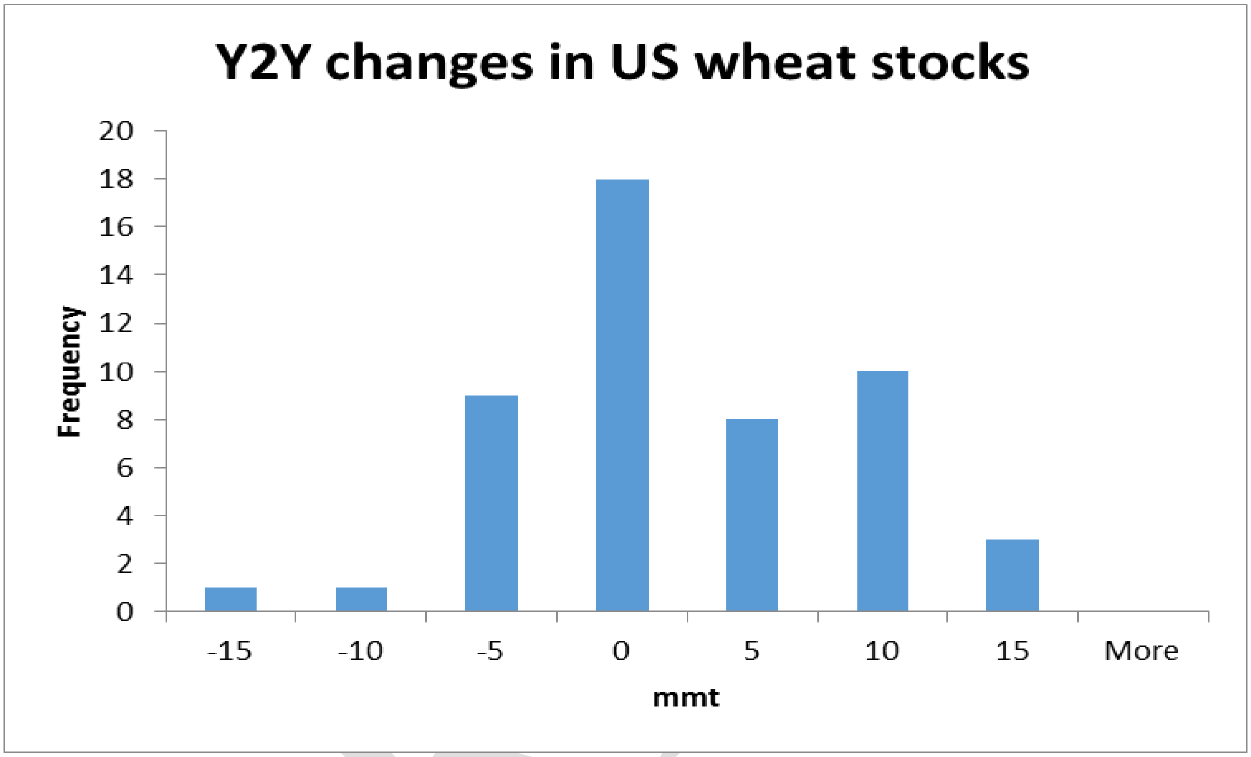


Figure 10: Distribution of year-to-year changes in US wheat stocks

Estimating stock changes in t

The analysis presented so far suggests that stock changes are likely to be normally distributed around a zero mean. However, we do not believe stock changes to be independent over time: the stock changes in previous years will likely influence the stock change in the current year. Thus, the expected value for the stock change in the current year may not have an expected value of zero, but rather some amount which depends on previous stock changes.

There is information that can be harnessed to gauge both the likely direction and even the likely amount of stock changes in t. In fact, a positive stock change in t becomes increasingly likely the longer the preceding period of stock drawdowns and the higher the amounts of drawdowns, i.e. the cumulative drawdown over time. If cereal stocks are being drawn down for say 15 years in a row, the likelihood for a drawdown in the successive year(s) becomes increasingly small[19].

[19] Such a development may lead to (near) stock-out and would be accompanied by a price spike (at least at the global level).

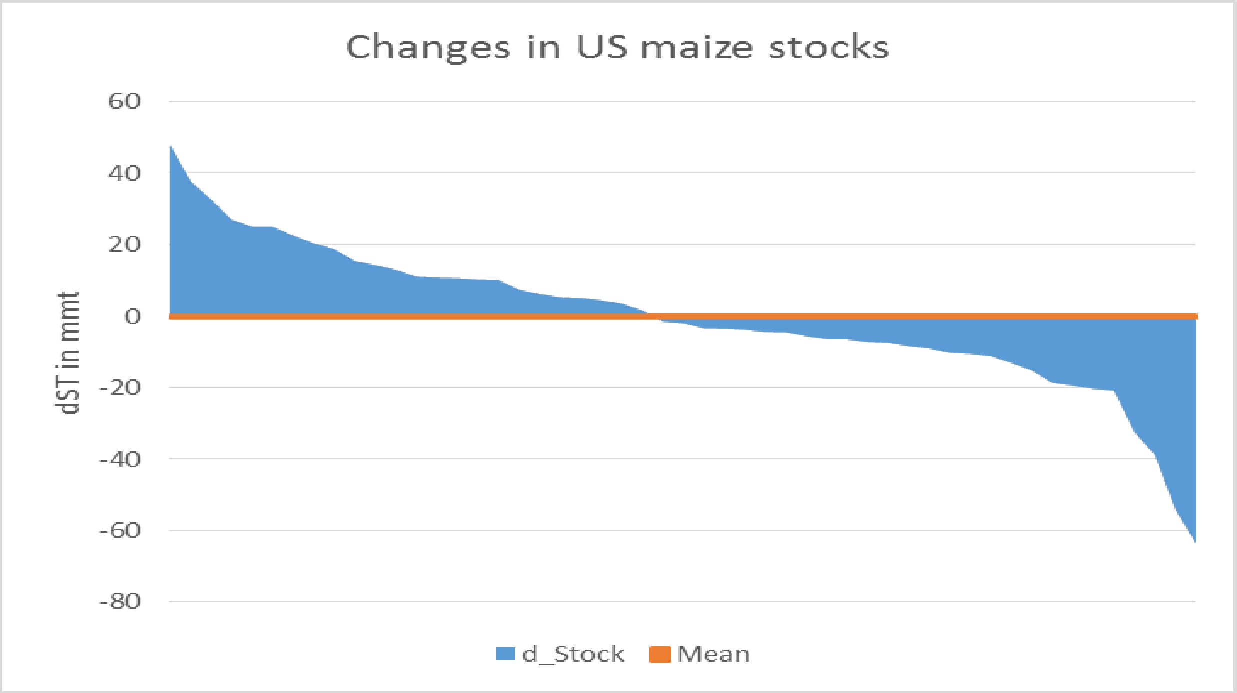


Figure 11: Changes in US maize stocks, ordered by magnitude of y-2-y change

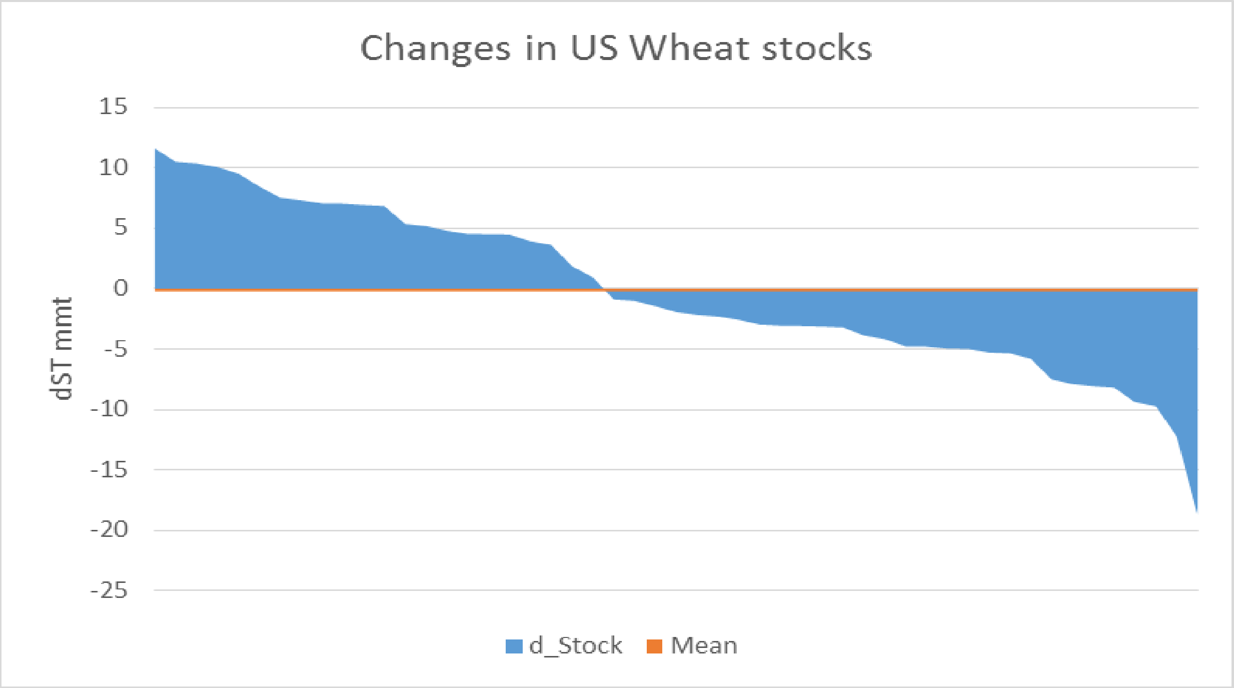


Figure 12: Changes in US wheat stocks, ordered by magnitude of y-2-y change

Conversely, if a country accumulates stocks for many years in a row, the cumulative amount of stocks would become so high that storage capacity dwindles, losses loom, and the costs of holding stocks become prohibitively high. This means that the probability for a positive change in time t increases the longer the history and larger the cumulative amounts of negative changes in the past and vice-versa.

Thus, we need to construct a model for stock changes that is informed by our knowledge of historical changes. One such model would be

$$\boldsymbol{\Delta S_t = \beta\left(\sum_{i=1}^{k} \Delta S_{t-i}\right) + \varepsilon_t} \qquad \text{(Equation 10)}$$

In other words, we assume that the stock changes at time t depend on the sum of the previous k stock changes. We add an error term to indicate that the change in stock at time t is not exactly equal to this value; rather, the distribution of the stock change at time t is adjusted given the knowledge of previous stock changes and still has variability. We could alternatively express this as

$$\mathbf{\Delta S_t \sim N\left(\boldsymbol{\beta} \sum_{i=1}^{k} \Delta S_{t-i}\, , \boldsymbol{\sigma}^2\right)} \qquad \text{(Equation 11)}$$

This provides a further improvement on the distribution of stock changes. We can assume a normal distribution and we can also adjust the mean based on the previous stock changes. This has the effect of reducing the spread of the distribution as we now know more about what the stock change at time t should be.

Returning to US cereals data, we can examine if our hypothesis holds true. Figure 13 shows the relationship between the sum of previous stock changes over 15 years and the current stock change. We see that there is generally a negative trend, thus validating our original hypothesis. The trends may not be extremely strong, but they allow us to infer a distribution for the stock change given previous changes. For example, if we know that over the previous 15 years the net stock change for "Barley and products" was 2,500 then we can say that the next stock change should be normally distributed with a mean of -400 and a variability which allows values between -1200 to 400.

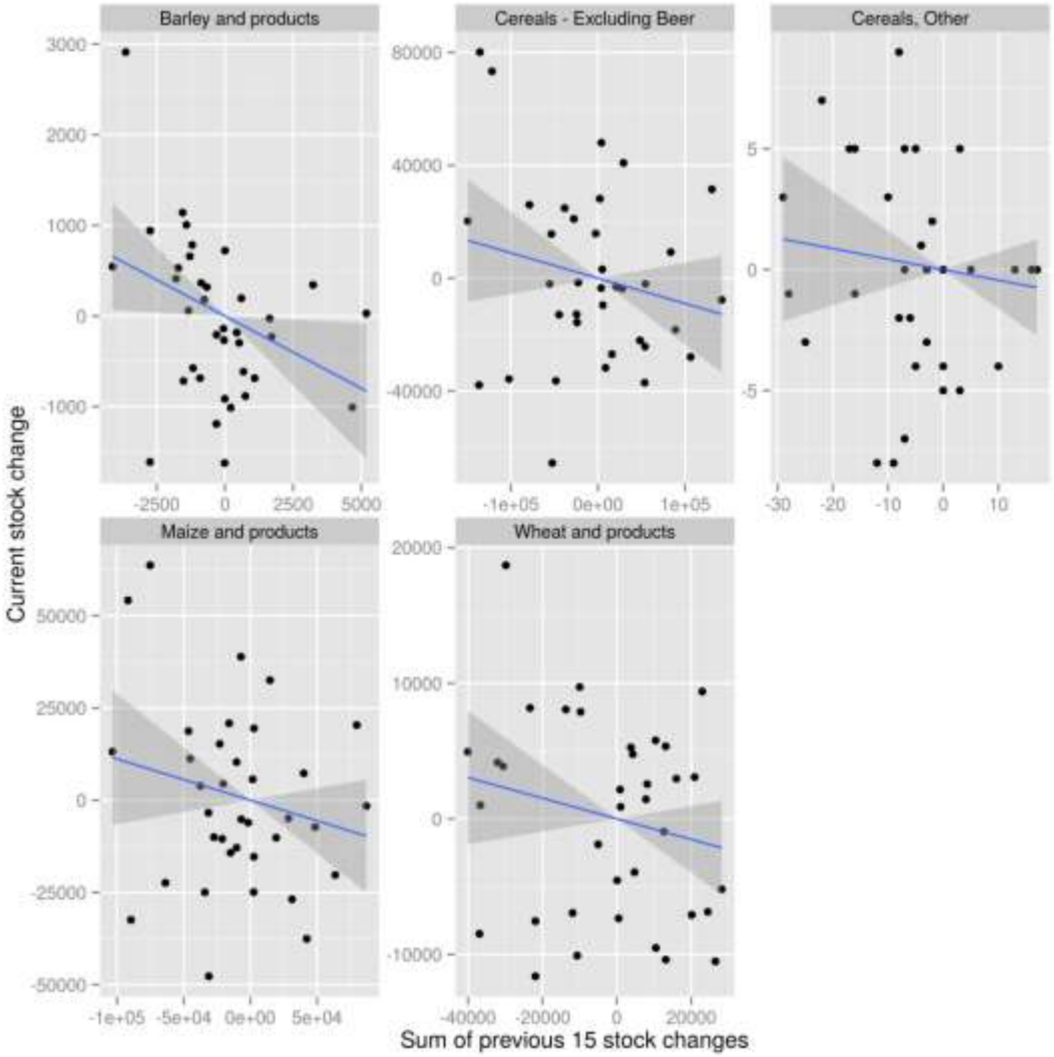


Figure 13: Stock changes in t in relation to cumulative past stock changes

Testing for normality in the distribution of stock changes

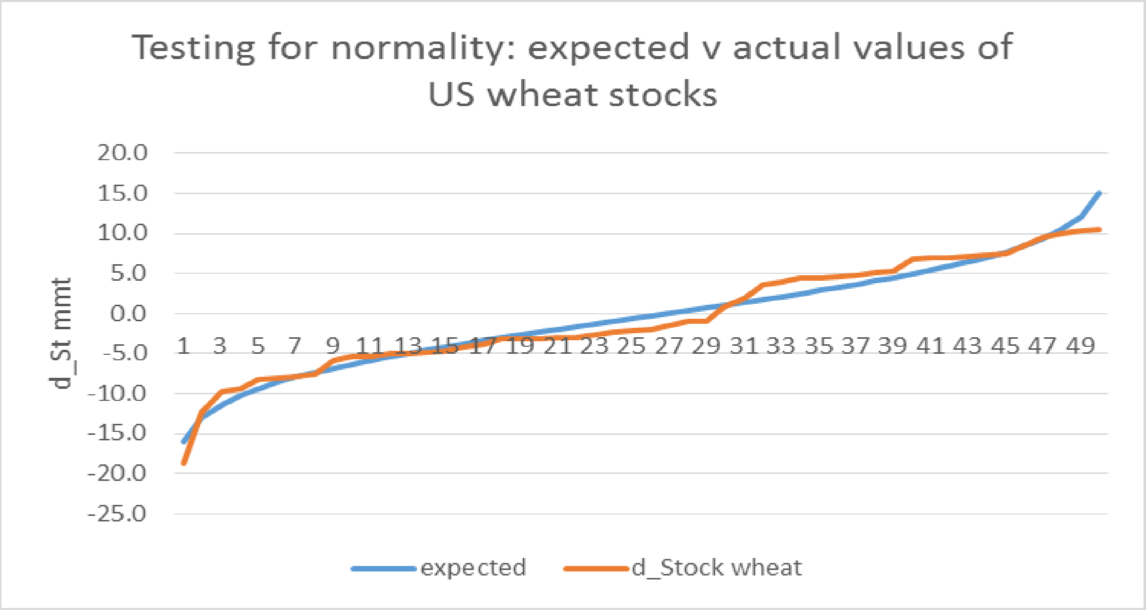


Figure 14: Testing for normality, wheat stocks

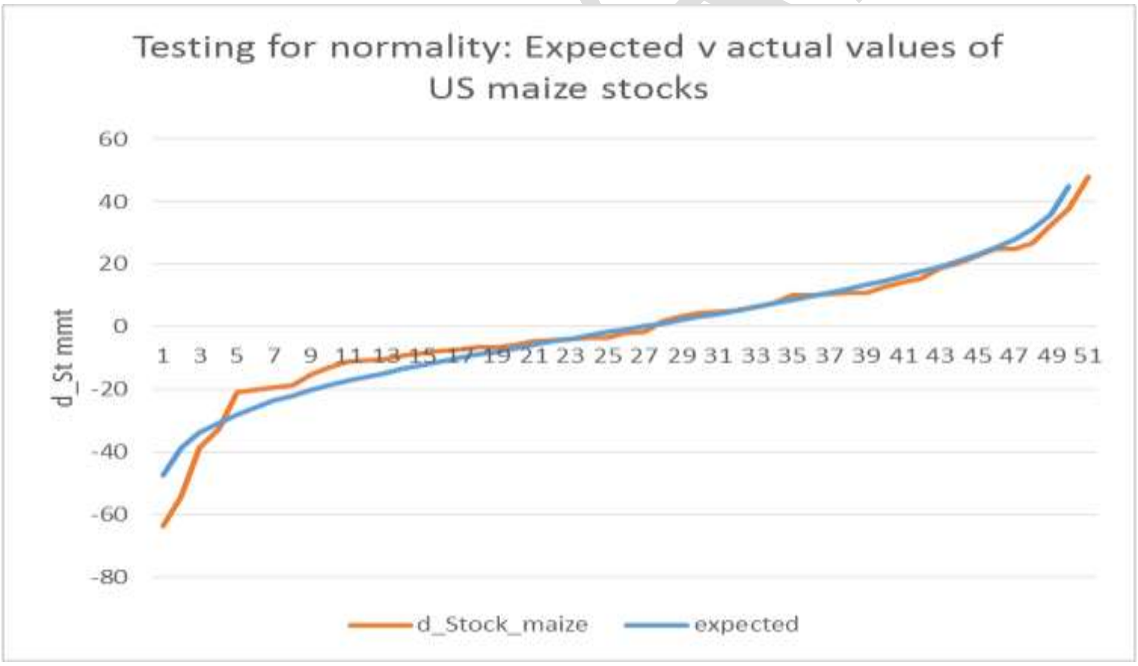


Figure 15: Testing for normality, maize stocks

Feed use

Feed use, along with many other elements of utilization, is typically not measured at the country level, no matter how developed the country's statistical system is. Even in the US balance, feed utilization serves as the balancing item and accordingly is termed "Feed and residual use". When feed demand is measured, it occurs mostly in developed countries and seldom through regular surveys. What is more, most feed surveys are limited to industrial feed sectors and are based on data only from commercial feed companies. The usefulness of such surveys is limited to countries with a highly industrialized livestock sector. Germany, Spain and China, for instance, obtain their feed estimates through a survey of feed companies. Only very few countries, e.g. Hungary and the Netherlands, have surveys that also include farm production of feedstuffs.

This lack of actual measurement appears even more surprising as feed use is a key source of disappearance in the food balance sheet (FBS) and of rising importance. Growing consumption of meat, milk, and other livestock products and ever more intensive production systems have resulted in rapidly rising demand for feed products, notably compound and concentrate feeds (CC feeds). Indirect, but rather compelling, evidence stems from the rapidly rising production of commodities, which are exclusively or at least mainly destined for feed use, notably oil meals and many coarse grain crops.

Moreover, shifts to higher consumption levels of livestock products and to more intensive feeding systems have not been limited to developed countries. Available evidence, including from the production of livestock products, suggests that the same trends now prevail in developing countries, and particularly so in many emerging economies (Brazil, China).

The absence of measured estimates and, at the same time, the rapidly growing importance of feed use means that any methodology to impute/estimate feed use deserves particular attention. Above all, it must capture intensifying livestock production systems and the rising overall importance, volume and diversity of livestock products. In addition, short term shocks such as epidemic outbreaks of diseases or other mortality-inducing events must be captured in feed use estimates. It should also capture the rapid technical progress in rearing animals in intensive livestock systems and, as part of the progress, the growing feeding efficiency with which feedstuffs are used in modern livestock systems. These trends may not always drive feed use in the same direction. For instance, information available from extension services and modern livestock operations suggests an offsetting effect of a growing feed intensity, i.e. ever high shares of CC feeds in total feed rations on the one hand and steady progress in using feedstuffs more efficiently, i.e. requiring less CC feeds to produce a given amount of livestock products. The dynamics of these changes should be reflected in an FBS imputation system.

A brief review of the existing methodology

Before conceiving any new methodology, it is useful to review and appraise the existing approach typically used to estimate feed demand, whether it adequately represents the current and prospective characteristics of feed use and livestock systems. This review can help in deciding (i) whether the old system can be used with minor changes; (ii) whether parts of the old system can be salvaged; or (iii) whether an entirely new system needs to be developed.

The basic approach currently used in the FBS system is easily stated. Feed use is simply calculated as a share of feed availability, i.e.

$$\boldsymbol{Fe_{i,j}} = \left(\boldsymbol{P_{i,j}} + \boldsymbol{I_{i,j}} - \boldsymbol{X_{ij}}\right) * \boldsymbol{r_{i,j}} \qquad \text{(Equation 12)}$$

i.e. the feed use of feedstuff (j) in a country (i) is determined as a ratio (r) of the availability of the feedstuff in this country. The share is specific to a given country and product, (largely) stable over time, and not specific to the different elements of availability, i.e. not distinguishing between production and trade.

From an ex-post perspective, the rationale behind the choice of this approach is difficult to understand. On the face of it, it may have been motivated by the need to capture the dynamics of feed use in a non-intensive feeding system, whereby feed use rises where and when supplies are ample, and shrinks where and when feedstuffs are in short supply.

Being supply-oriented in nature, the approach could reflect reality insofar as CC feed use rises in a non-intensive system after bumper crops and contracts, often sharply, after crop failures. Possibly being reflective of livestock production systems several decades ago (when the method was conceived), the method is, however, hardly representative of modern feeding systems, and seemingly fails to capture the level of feed use even in the least developed countries. It was therefore decided to radically change the imputation approach from a simple, supply-driven method to a more complex yet tractable approach that correctly respects the demand-driven nature of feed demand in most countries.

The new feed use estimation procedure

This section will provide an introduction to the new feed use methodology. Being a radical departure from the previous approach, we pay considerable attention at least to its main elements, noting that a still more detailed description is available from the specialized documentation of the new approach to feed estimation.

The new feed use estimation process can be divided into three steps or stages. To begin, a stepwise and demand-oriented approach to determining CC feed use is implemented. Total feed requirements $R_{i,t}$ in a given country i and year t are calculated as a function of the herd size N and the species, country and time specific requirement index, $r_{s,i,t}$.

- Determining total feed requirements R_i

The requirement indices are further expressed in terms of an animal unit and incorporate maintenance needs M and output requirements, such as for meat, milk or wool P, as well as the overall efficiency E (arc-efficiency over the entire herd[20]) that is characteristic of a given feeding system.

$$R_{i,t} = \sum_{s=1}^{S} N_{s,i,t} * r_{s,i,t} \qquad \textbf{(}\text{Equation 13)}$$

where

$$r_{s,i,t} = f(M_{s,i,t}, P_{s,i,t}, E_{s,i,t}) \qquad \text{(Equation 14)}$$

The requirements are calculated for metabolizable energy (MJ ME) and tonnes of crude protein (t CP) via the requirements index, which is constructed dualistically. They are inclusive of all feedstuffs, not only CC feeds. In a second step, the shares of actual CC feed in total requirements are determined.

- Determining actual use of CC feed

The first step was used to calculate total requirements based on the needs of a herd, its needs for maintenance and output, for all types of animals within a herd and across species. These needs can and indeed are being covered by both CC feeds, and other feedstuffs such as roughages or products from pastures, or even table scraps. To determine the share of CC feeds, it is necessary to multiply the Requirements R_i with an intensification factor $I_{i.}$ The intensification factor is a simple ratio that determines how much of the total requirements will be covered by CC feeds and how much by other feedstuffs.

This can be expressed as follows:

$$\boldsymbol{CC_{Feed} = R_i * I_i} \qquad \text{(Equation 15)}$$

Again, all calculations distinguish between metabolizable energy (MJ ME) and quantities of crude protein (t CP).

The intensity factors I_i vary across feeding systems and animal species A_i. Across animal species, the main dividing line is between ruminants and monogasters (pigs and poultry). For monogasters kept in modern feeding systems, the intensity factors have been rising steadily over past decades and have reached values close to unity in essentially all developed countries. This is particularly the case for poultry where values have reached high levels in all countries. Only organic agriculture leaves actual ratios for poultry below unity. For pigs, intensification factors can still be well below unity where backyard production and non-intensive family farms account

[20] E.g. in the case of pigs, the requirements are those of all "types" of pigs along the production chain, i.e. from piglets to hogs to sows, etc. Likewise, the feeding efficiency refers to the arc-efficiency across all types. This indicator lies necessarily above the better known feeding efficiency ratios which refer to rearing hogs only, i.e. growing them from 30kg to slaughter-weight of e.g. about 100kg. Lower feeding ratios signify higher efficiency, as fewer kg of feedstuff are needed to produce 1 kg of livestock output, in this example pig meat.

for a large share of output. This is still the case for very large production systems such as pig production in China; but even there intensification factors have been rising steadily and rising rates now feature prominently in most developing countries, too. The actual data for the intensification rates have been gleaned from the FAO GLEAM database (2005, 2010), namely representative feeding baskets in countries, and have been extrapolated to capture the most recent situations.

For ruminants, the spectrum of intensification rates varies much more widely than for monogasters, even in mature feeding systems. In all feeding systems, intensification rates of unity or close to unity cannot be found in practice and are unlikely to occur in the future. The physiological needs of ruminants for a certain minimum amount of crude fibre mean that even the most intensive feeding systems (e.g. milk production in Israel or meat production in US feeder lots) will have intensification rates of less than unity. On the other end of the spectrum, there are systems where roughages remain the economically most efficient source (abundant pasture in New Zealand) or opportunity costs of CC are too high for feed use (milk production in India). In these systems, the intensification rates I_i would seem only marginally above zero.

- Determining the requirements of fed/farmed fish

The methodology presented thus far has not taken into account the feeding of fish, as required in aquaculture. In fact, most available feed statistics do not take this part into account. Given the increasing importance of farmed fish as a source of protein for human consumption, it is expected that the industry absorbs rather significant amounts of agricultural as well as marine outputs. In absolute terms, the most important users of fish feeds are in Eastern and South Eastern Asia, most importantly in China. As a share of total feed, aquaculture is particularly important in small island states such as Iceland and the Faroe Islands where fed fish are often the main consumer of feedstuffs. Given their absolute and relative importance, a complete and correct assessment of feed use requires including feeds fed to farmed fish into the estimation process.

The estimation of energy and crude protein needs for aquaculture are derived in analogy to those for livestock. To arrive at a precise assessment, species-specific feed efficiency ratios (FCR) are directly applied to aquaculture production data (tonnes of fish and crustaceans) and converted into energy and crude protein equivalents, which yields the total nutrient requirements of the produced amounts of fish.

Also fish are farmed at different intensity levels. On the one hand, salmon and carnivore trout must entirely be fed with feeds from outside the ponds, while other species, e.g. carp, may retrieve nutrients from microorganisms, aquatic plants, fellow fish and aquatic wastes that are available in their surrounding waters, depending on the production system. Hence, an intensity factor is applied to the overall biological requirements in order to circumscribe the actual aqua-feed demand.

Both the parameters of FCR and feeding intensity are retrieved and extrapolated from survey data published by the FAO Fisheries Department. Eventually, the aquaculture requirements are added to those of livestock to arrive at the total feed energy and crude protein use for every year and country.

Determining use of individual feedstuffs

Thus far, the new approach has only produced total feed requirements and total feed use of CC feeds in terms of metabolizable energy and crude protein. No distinction has yet been made to identify which of the various feedstuffs contribute to cover CC feeds. The next step required to produce this information in the FBS is to break the CC feeds of energy and protein down into individual feedstuffs at the commodity level.

The most common approach used for such an allocation in practice is to set-up a linear programming system (LP) that determines the most price-efficient combination of feedstuffs, which meets the constraints of total requirements and the nutrient composition of individual feedstuffs. Theoretically, such an approach may be very appealing, but it lacks practical relevance for the FBS system. For one thing, it requires the exact knowledge and constant updating of prices for feedstuffs. Collecting this information is a tall order in developed countries and infeasible or impossible in developing ones. For another, it would render an economically optimal, but not a necessarily empirically likely solution, simply because actual use of CC feeds is determined in practice by many factors beyond relative price quotations, as supply-chain constraints and opportunity costs need also to be considered. Finally, an LP solution often produces border solutions, i.e. the constraints imposed are likely to determine the actual solution space, and ultimately the chosen combination of feedstuffs.

An alternative solution is to break the total requirements down by availability shares. This means allocating feedstuffs in accordance to overall availability, i.e. feedstuffs available in abundance are used more than those in scarcity[21]. Within the allocation of CC feeds to individual components, again a two-stage process is employed. In a first step, all those CC feeds that can only (or at least in their vast majority) be used as feeds are allocated first. In essence, these are oil meals and oilcakes as well as all brans not used for human consumption (not as breakfast cereals). Their energy and protein content is deducted from the estimated overall feed use.

$$\text{Cereals, tubers, pulses etc.} = CC_F - (OM+OC+brans)$$

In a second step, all other feeds from commodities (cereals, tubers, pulses etc.) from inside and outside the FBSs, are allocated as a simple proportion of their availabilities.

[21] This procedure retains the basic idea of the previous approach used in determining feed use. The decisive difference to the previous approach is that the availability shares are applied after requirements are determined and intensification is respected.

As requirements are expressed in energy and protein units, two factors, one for the proportion of energy availability and one for that of protein are constructed:

$$a_{f,i,t}^{n} = \frac{n_f q_{f,i,t}}{\sum_{f=1}^{F} q_{f,i,t} n_f} \qquad \textbf{(}\text{Equation 16)}$$

The proportion of feed availability (a) of feedstuff (f) in country (i) and year (t) for the respective nutritive conversion factor (n), which is either energy units or protein units per unit of feedstuff, is determined by its available nutritive quantity (q * n) divided by the total nutritive value provided by all available feedstuffs.

These factors can now be applied to the requirements and converted into quantities (tonnes). Given the different composition of feedstuffs in terms of crude protein and energy content, two slightly different allocations of feedstuffs are obtained, i.e. energy rich feedstuffs will be more important contributors to overall energy availability than protein rich feedstuffs and vice versa.

The final quantities of feedstuffs must have the property that they meet both the energy and protein requirements. Hence, it is possible that from the two allocations (i) one allocation meets both requirements, (ii) no allocation meets both requirements, (iii) both allocations meet both requirements.

In the case of (i), the allocation which meets both requirements is preferred. In the case of (ii), upper and lower bounds for each feedstuff can be established since each allocation meets its individual demand. In other words, if the protein based allocation meets the protein demand, and the energy based allocation meets energy demand, the maximum values of those two for each feedstuff must at least satisfy if not exceed both requirements. Similarly, the sum of the minimum values of each feedstuff will not satisfy any of the two requirements and thus, an optimal solution satisfying both demands lies in between. In order to avoid the allocation of excessive feed, the final quantity for each feedstuff is obtained by optimizing the allocation linearly such that both requirements are satisfied, using the least amount of additional feedstuffs in accordance with their respective nutritive values. This means that preference is given to energy and protein rich feedstuffs after the minimum quantities have been allocated. For case (iii) the same procedure as for (ii) is applied, but naturally in the opposite way, i.e. quantities are subtracted in order to arrive at the minimum solution that meets both requirements. For instance, the final allocation of wheat for feed use is determined as a share of total availability is retrieved as

$$W_{Feed} = a_W R_{Fi} c_n \square \qquad \text{(Equation 17)}$$

where c is the respective conversion factor from either protein or energy into wheat quantities and depends on the derivation of a.

Results and consistency checks

Figure 16 depicts global energy and protein feed demands respectively from 1992, grouped in the World Bank income level classification. The shift towards pig and poultry production, especially in emerging economies becomes apparent as the protein demand of these countries exceeds the levels in high income countries at an earlier stage than the energy demand. This reflects the fact that monogasters require relatively more protein than ruminants.

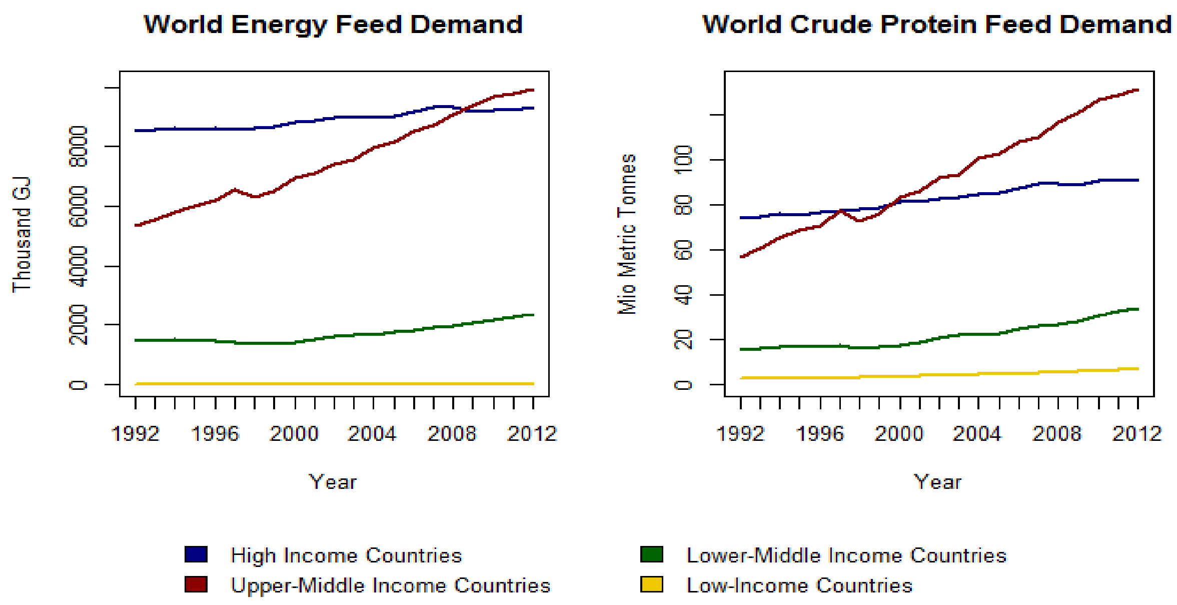


Figure 16: Results of the feed estimates at the global level

The validation of feed estimates is a challenging task, not least in view of the fact that reliable feed statistics are seldom available. The first stage of results, namely the feed requirements expressed in crude protein and metabolizable energy, are being compared to reported feed figures. As some countries officially report on feed in the annual FAO questionnaire, these have been taken as the starting point for comparisons with the results of the new method. Figure 17 depicts the results of this comparison.

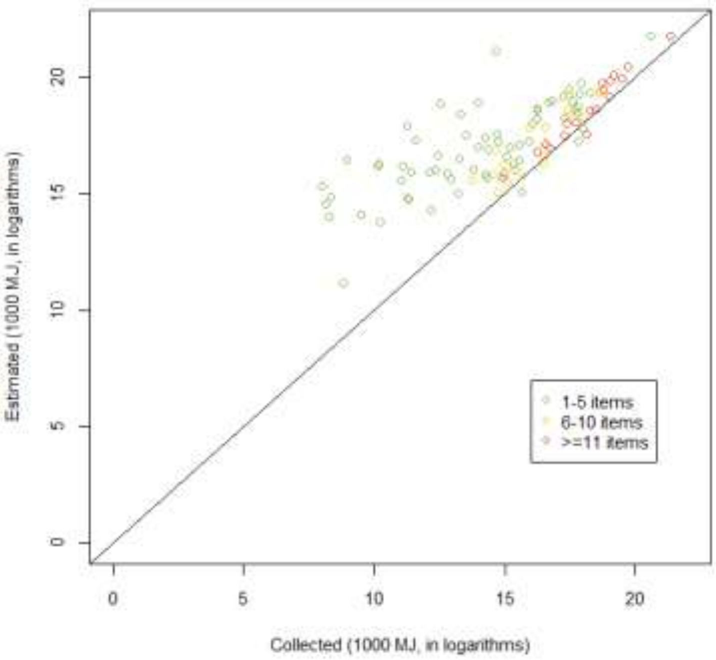


Figure 17: Estimated vs expected value for metabolizable energy

In Figure 17 each dot represents a pair of estimated feed demand and collected feed use in logarithms, and the straight line represents the scenarios in which collected energy feed fully coincides with estimated energy feed in a given country. It is worth noting, that most countries report only sporadically on individual feeds, which means that the whole range of feedstuffs is seldom covered. As the information on feed use is not complete, feed reported should not exceed estimated feed. Graphically this implies that the dots should never fall below the diagonal line, for which only five cases can be spotted. (Saudi Arabia, Cyprus, Kazakhstan, Azerbaijan and Egypt). This may be due to either underreporting of feed or overestimation of demand[22].

The difference between reported feed and estimated feed can represent the feeds that have not been reported. From Figure 17 it is apparent that, the more extensive countries report on feed, the closer the estimate to the collected aggregate. Assuming that countries that have a higher item coverage in reporting feed are also more confident about their real feed usage levels, this provides strong evidence that the estimated requirements are, indeed, not far from reality. Indeed the correlation coefficients between collected feed use and estimated demand increase steadily, when the item coverage threshold is raised. Starting with .93 for an item coverage of greater than six, for which 36 observations are available, .99 for an item coverage of 20+, which however only applies to 4 cases. In addition to that, the discrepancy of absolute feed required and feed used in these cases has shrunk to 13%, encouraging the methodology for estimating feed demand.

[22] In this round of estimation camels and horses have not yet been taken into account. Given the importance of these animals in these countries, this could well have led to underestimation in these particular cases.

The next step is to check how the final feed estimates, after allocating particular feedstuffs, compare to other prominent data sources. In Figure 18, global cereal feed data from the USDA, and FAOSTAT (the old estimates derived under the old methodology) from 1992 to 2011 are depicted. Besides the newly modelled feed, all other data sources use entirely supply-driven estimates, except for FAOSTAT figures that are officially reported by countries. However, reported figures are not modified by the new feed estimation procedure.

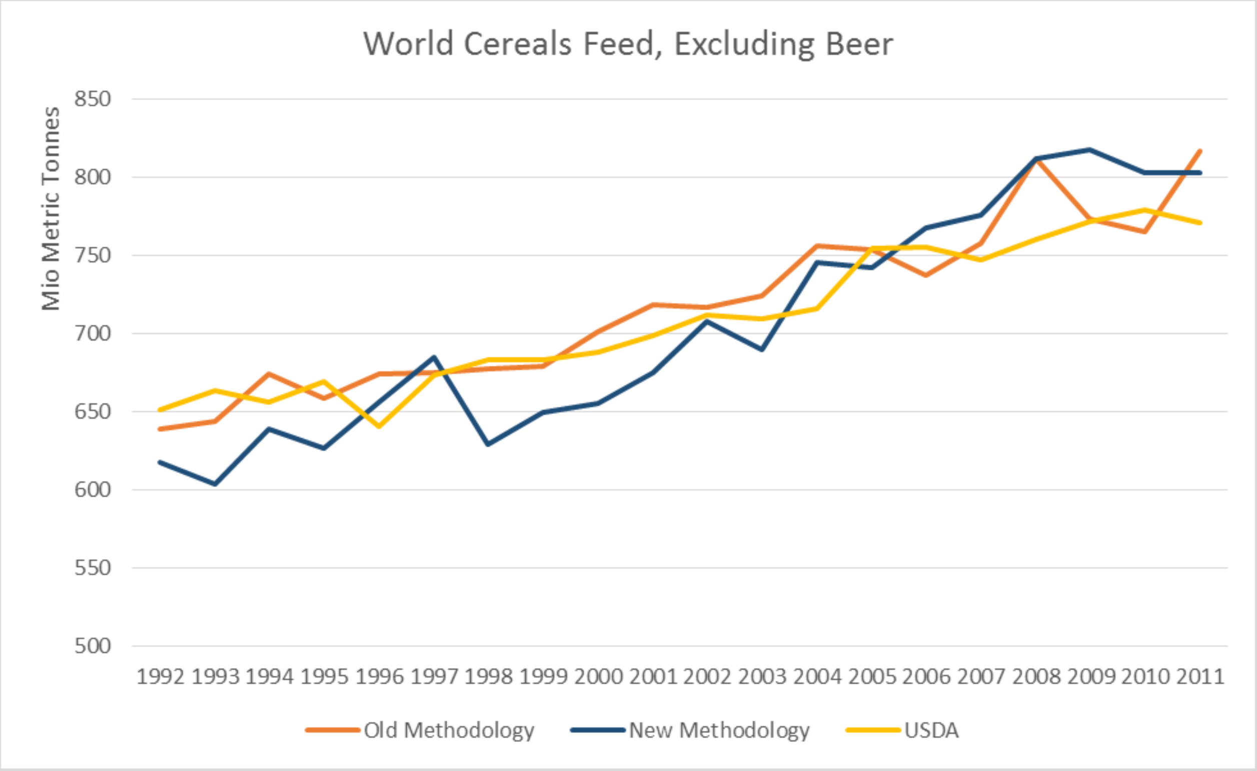


Figure 18: World cereal feeds, comparing results at the global level

A review of the empirical results

Overall, the new methodology suggests a slightly steeper increase of cereal feed than other data sources/estimates, which has several explanations. From the demand side, the sharp increase of herd sizes and animal production over the past decades certainly can lead to higher feed use than supply oriented allocations suggest. In addition, the growing intensification of animal husbandry has favoured the use of energy-rich grains, in particular maize. Combined with the shift towards pork and poultry, these factors support the faster increase of grains produced in the new methodology (Figure 18). What is more, feed use in aquaculture, which was not taken into account in the old methodology, in reality, has exhibited extraordinary growth and feed use rates, and also supports the faster increase produced by the new methodology as well. Perhaps most importantly, it shows that the results of the two approaches are much more different than Figure 18 suggests. Had aquaculture been taken into account in the old approach, the existing FAOSTAT estimates would have been higher by the equivalent of 47.3 million tonnes of aquafeed equivalents[23] in 2012, while in 1992 only 5.9 million extra tonnes of aqua-feed equivalents are not considered.

Secondly, the new methodology suggests more variation of feed use over time. It seems to be counterintuitive that cereal feed use is actually steady over time. Feed is at least the second biggest factor of disappearance for most cereals. Since food use tends to be stable if commodity availability varies, feed is expected to absorb parts of variation as determined by availability.

Thirdly, while the global use estimates may not be entirely dissimilar, even after accounting for aqua-feed and feedstuffs for camels, horses, etc. there are considerable differences across countries, which cancel out in the global aggregate. The examples at the end of this chapter illustrate cases for both over and underestimation of food use (Argentina and Myanmar).

Consistency and Quality checks of feed allocations

Overall, while the new method is more data and knowledge intensive, it nevertheless allows consistency checks of feed allocation vs the amount of nutrients that are required to produce the animal products reported by a country.

Negative protein balance

Arguably, one of the most reliable estimates of feed requirements in FBS comes from outside the FBS system; these are the estimates for the availability of oil meals and cakes. For one thing, oil meal and cakes are destined exclusively for feeding purposes [24] and for another, oil meals and cakes are by-products of oilseeds crushing. This means that they are industry products, and both production and imports of such variables are more reliable than estimates of primary products, not to mention use data. A similar rationale can be put forward for brans. What is more, both feedstuffs are protein-rich commodities and often the main ingredients to cover protein needs in a feed ration. The high confidence that can be put into these estimates is not only the basis for identifying them in the feed allocation process separately (1st step), but also offers possibilities to undertake consistency checks.

The main approach to consistency checks is to juxtapose the availability of these feedstuffs with the needs of protein to cover protein requirements, or rather total use of protein. The most obvious case of an inconsistency is when the available protein from oil cakes and meals, as well as brans exceeds the calculated needs/total use of protein (requirement times intensification rate). Figure 4 shows all 5256 data points, each representing the logarithm of the amount of protein required and the amount of protein available from oilmeals, oilcakes and brans in a given country

[23] Aquafeed, similar to compound feed for livestock, is composed by feedstuffs external to fish farms and may include a variety of ingredients, e.g. grains, meat or fish meals and oils. 47.3 million tonnes of aquafeed is equivalent to about 140 million tonnes of maize and 5.9 tonnes of aquafeed corresponds to about 17.3 million tonnes of maize

[24] Some oilcakes are being used as biofuels (e.g. cakes of olives), but they are generally negligible in terms of quantities and would not be considered feedstuffs under normal circumstances.

and year. In 526 cases the protein availability exceeds demand, which represents about 10% of all data. In Figure 19, this is apparent as the dots below the 45-degree line.

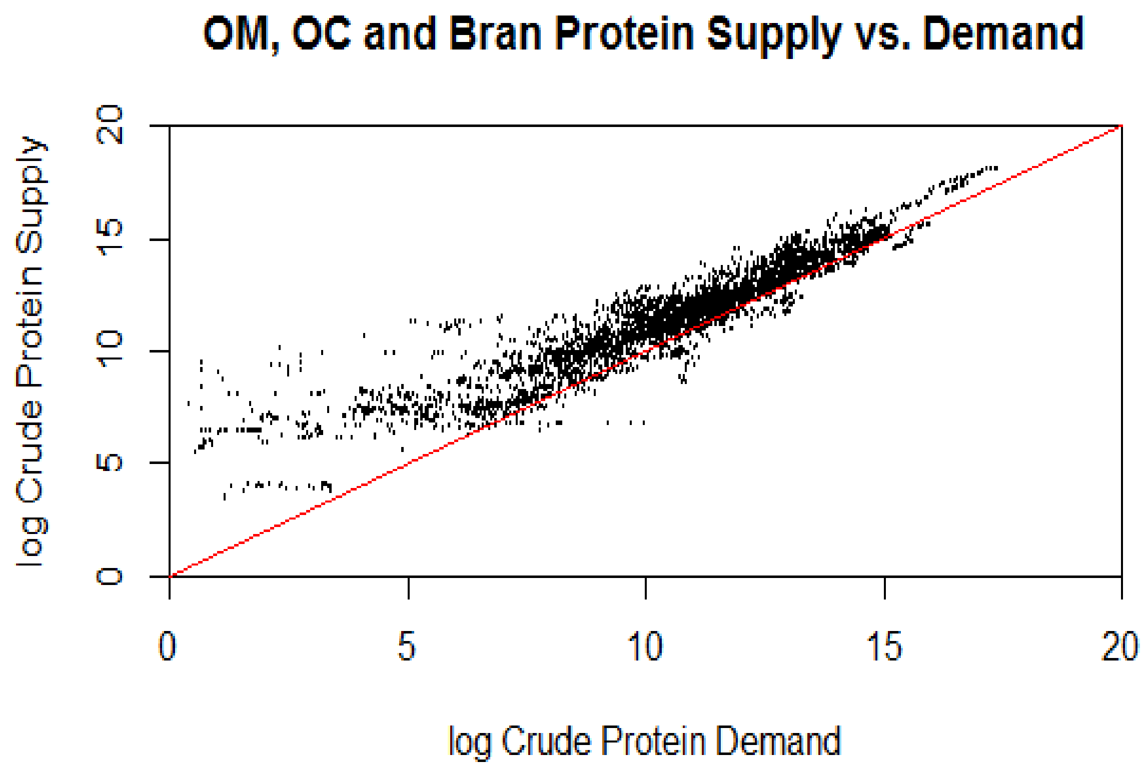


Figure 19: Oilcakes, oil meals and brans supply vs. demand

Such an imbalance can be due to two principal reasons. First, the intensification rate is too low and thus underestimates the actual needs. Second, either the number of animals or the feeding efficiency ratios are too low, which again would understate the actual needs. Such negative protein balances seem to be unlikely to occur, simply because they would suggest a cross underestimation of needs[25]. But in practice they have occurred in many instances after systematically triangulating protein needs with availability. The results have been reviewed on a case by case basis (country by country, year by year) and resulted in many adjustments of animal numbers and livestock productivity, in line with the newly obtained information from consistency checks. Where these are based on official estimates, the inconsistencies are being brought to the attention of the data suppliers. A separate document has been compiled providing a complete overview of the results.

[25] A negative balance would not only not leave room for any other CC feed, but suggest that availability from OM/OC already exceeds the requirements inherent in the needs calculations.

Insufficient availability

Just like there are negative balances, in some countries, at least in some years, vastly positive balances have occurred. While such large positive balances can occur without violating the feed approach in principle, they can absorb so many cereals and pulses, that they would render unreasonably low results for food and other uses.

Country cases for Argentina and Myanmar

In the following section, the feed estimation process for Argentina and Myanmar for the year of 2011 shall serve as a practical example. Both countries host the complete range of considered species and aquaculture and are therefore useful examples to examine the new approach. Both countries also represent the animal husbandry trends (from ruminants to monogastrics) as well as recent changes in feed utilization (mostly maize in Argentina and predominantly rice in Myanmar) which are characteristic of their respective regions and income class.

The starting point, as outlined earlier, is the animal numbers of each species. The biological requirements measure is expressed in terms of an animal unit index (AUI), one for energy needs and one for protein needs, and are measured using standard regression equations provided by state of the art animal science analysis. In particular, the National Research Council's publications on animal nutritive requirements of the National Academy of science of the United States of America have been used to estimate the requirements of each species in a given country and year.

The result of this process is set in relation to the standard requirements of the base unit, which is that of a mature 500kg cow that produces 3500kg milk and calves every 13 months. Such a cow is expected to metabolize 35600 MJ of energy and 319 kg of crude protein per year. Within the species, a distinction between dairy and beef cattle is made as well. Hence, applying the indices to the animal numbers will require the multiplication with the base unit requirements in order to obtain whole country requirements. Next, only the demand generated through intensive feeding systems, and not through roughage and waste feeds will be filtered out by applying the intensification rate (IR). Thus, in Table 6, the calculation goes as:

Table 6: Feed use requirements calculations for Argentina

Animal	Number	AUE	AUP	IR	Energy Demand (000 MJ)	Protein (metric tonnes)
Cattle	46000000	0.6414	0.3818	0.0757	79460350.42	423876
Sheep	14731000	0.0404	0.0352	0.0062	131010.3619	1021.628
Goats	4280000	0.0320	0.0278	0.0062	30138.76003	235.0242
Pigs	2350000	0.2012	0.1141	0.8145	13710611.58	69645.52
Chickens	1.00E+08	0.0210	0.0466	0.9386	70300170.37	1394261
Ducks	2550000	0.0230	0.0551	0.9386	1957614.199	42074.84
Geese	165000	0.0256	0.0636	0.9386	141020.3539	3141.334
Turkeys	3050000	0.0518	0.0480	0.9386	5278047.506	43872.91
Aqua					22155.82694	580.3475
		Demand			**171031119.4**	**1978709**
		OC, OM &Brans			**65268351**	**1699681**
		Cereal Feed Demand			**105762768.4**	**279028.1**

Table 7: Feed use requirements calculations for Myanmar

Animal	Number	AUE	AUP	IR	Energy Demand (000 MJ)	Protein (metric tonnes)
Cattle	14088043	0.4011	0.2609	0.0729	14668012.33	85479.11
Sheep	854383	0.0387	0.0338	0.1116	131647.2227	1026.595
Goats	3851919	0.0435	0.0379	0.1112	667026.0403	5201.518
Pigs	10497493	0.1811	0.1071	0.6822	46193124.93	244859.6
Chickens	176839000	0.0111	0.0224	0.8702	60900789.88	1100012
Ducks	15507000	0.0135	0.0270	0.8702	6468232.347	116336.4
Geese	1966000	0.0181	0.0403	0.8702	1106240.864	21984.46
Turkeys	3000	0.0348	0.0282	0.8702	3233.141735	23.53616
Aqua					10784477.33	220610
			Demand		**140922784.1**	**1795533**
			OC, OM &Brans		**44981407.66**	**661432.3**
			Cereal Feed Demand		**95941376.43**	**1134101**

Energy = Numbers * AUE * IR * 35600 and Protein = Numbers * AUP * IR * 0.319

Summing up all requirements of livestock as well as poultry species and adding those from aquaculture yields the aggregate CC Feed Demand. Subtracting the amounts of energy and protein available in OC, OM and Brans finally gives the remaining demand that has to be covered by cereals, pulses, tubers etc.

Having established demand, now the supply side is added to the picture. Availability shares are constructed focusing on the remaining feed items, mostly cereals, pulses, tubers, but also meat meals and milk. For each commodity and its corresponding nutrient supply, an availability share is constructed yielding two independent allocations. In the case of Argentina, the energy based feed allocation meets both demands, while in Myanmar it is the protein-based feed allocation that satisfies the dual demand structure.

Given now the decision on the allocation, the remaining requirements (Table 8 and Table 9) are allocated and finally converted from nutritive values into tonnes of feed through dividing by the respective content per unit of feed. In the case of energy, this is expressed as MJ per kg, while for protein it is in percent of feed.

Table 8: Feed Availability shares and nutrient conversion factors for Argentina

	Energy based	Protein based	Energy cont.[26]	Feed (tonnes)
Canary Seed	0.0007	0.0007	12.50	5804.3
Meat Meal	0.0002	0.0002	12.20	1669.0
Cassava	0.0070	0.0018	13.00	57121.1
Wheat	0.0160	0.0183	13.80	122742.0
Molasses	0.0284	0.0177	11.10	270989.5
Beet P. Dry	0.0000	0.0000	12.90	81.1
Wheat Gluten	0.0001	0.0003	15.50	697.4
Barley	0.0661	0.0698	14.00	498978.9
Carobs	0.0000	0.0000	12.20	0.0
Maize	0.4082	0.3755	14.30	3018928.8
Germ Maize	0.0013	0.0018	14.40	9792.2
Fr Pulp Feed	0.0307	0.0114	14.20	228598.6
Cocoa Husks	0.0000	0.0000	14.80	335.8
Rye	0.0030	0.0028	13.60	23196.1
Oats	0.0265	0.0261	14.00	200555.9
Millet	0.0000	0.0000	14.30	221.4
Sorghum	0.1627	0.1539	14.30	1203029.0
Skim Milk Cows	0.0237	0.0304	12.20	205424.9
Whey, Fresh	0.2253	0.2892	12.20	1953215.1

Table 9: Feed Availability shares and nutrient conversion factors for Myanmar

	Energy based	Protein based	Protein cont.[27]	Feed (tonnes)
Roots Tub Ns	0.0000	0.0000	0.054	11.3
Wheat	0.0039	0.0042	0.126	38184.4
Bagasse	0.0035	0.0005	0.018	32950.9
Beans, Dry	0.1330	0.2753	0.248	1259100.7
Chick-Peas	0.0150	0.0263	0.221	134780.7
Bran Pulses	0.0057	0.0100	0.215	52948.4
Rice, Paddy	0.4713	0.3461	0.083	4728434.3
Rice, Broken	0.2376	0.2138	0.104	2331410.7
Maize	0.0819	0.0707	0.105	763818.2
Oats	0.0000	0.0000	0.110	1.9
Millet	0.0106	0.0109	0.125	99089.1
Sorghum	0.0095	0.0085	0.108	88961.9
Dry Whey	0.0004	0.0004	0.125	4059.4
Whey, Fresh	0.0275	0.0332	0.125	301116.2

[26] In MJ per kg
[27] As a share of feed

Hence, the last columns provide an estimate of how much of each commodity is used as feed and can be implemented in the balance. Figure 20: Argentina use of CC feed, old v new methodology depicts aggregate cereals of this procedure for the whole time period from 1992 to 2012, comparing it to the previous estimate in place, and the amount of OC, OM and brans supply. In general, an increase of oilcake supply will, ceteris paribus, lead to a decrease of cereal feed, at an elasticity which is defined by respective conversion factors. Graphically, this would lead to an asymmetric reflexion of the cereal feed line and the OC, OM and bran line. Obviously, if levels of demand change, the lines will behave accordingly as to cover the additional or decreased feed required.

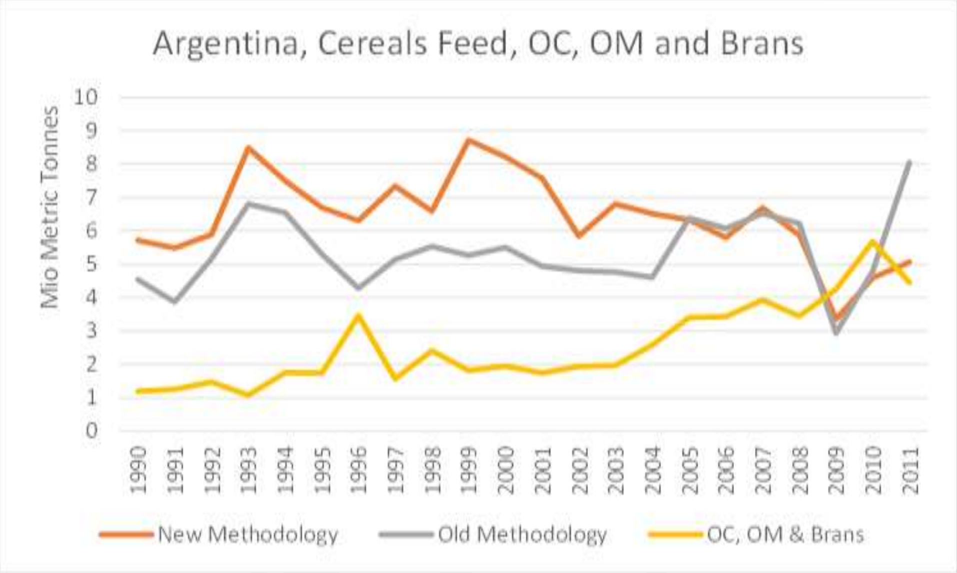


Figure 20: Argentina use of CC feed, old v new methodology

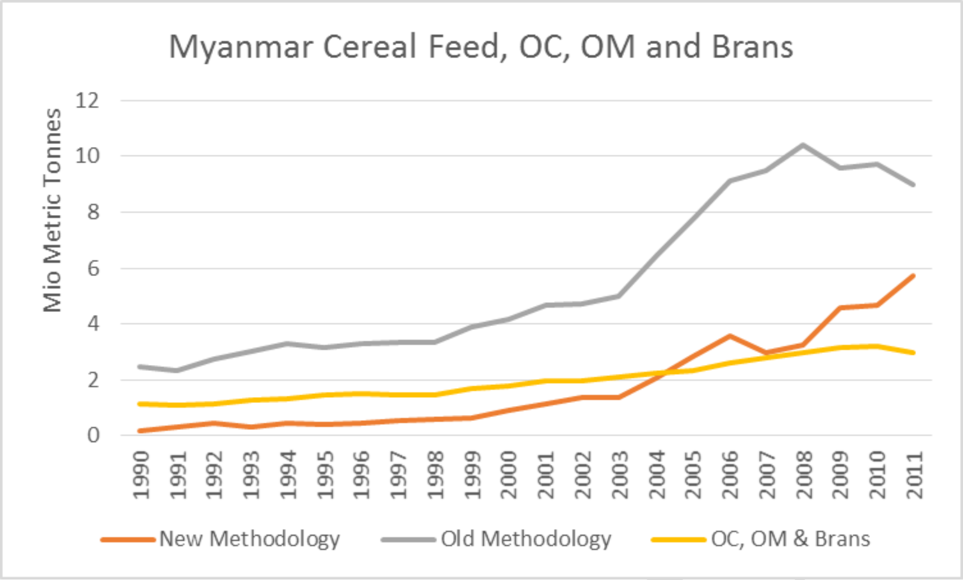


Figure 21: Myanmar use of CC feed, old v new methodology

In the case of Argentina (Figure 20), the new estimations lead to an overall higher level of cereal feed estimation, while the trends are somewhat similar, being very much in sync from 2005 on. Generally, this is linked to herd sizes and feeding intensity. Here, both suggest a higher use of cereal feed. The divergence from 1998 is due to an increase of cattle stock by more than 3%, combined with increased exports of oil cakes, and an overall steadily inclining productivity and numbers of animals in other species. The 2009 change is due to a decreased availability of cereals, which may have its source in the cereal price spikes from 2006 on, and is also in accordance with previously supply oriented estimates. Indeed, oilcake exports show a reduction by a considerable percentage. In particular soybean meal exports decreased by about 16% from 2007 to 2009. In addition, other feed items show higher levels of usage in order to compensate for the lower cereal use.

The estimates for Myanmar suggest a similar trend until 2006 like the old feed figures. However, productivity of ruminants started to decline sharply, for instance in cattle by 15%. After 2006, the number of pigs and poultry rose in tandem with productivity, while the number of ruminants was rather stagnant. This would support the idea of a shift from ruminants to monogastric species also taking place in Asia. Monogasters are reared in a more intensive manner than ruminants, which after all would not support the decreasing levels of feed use.

As both cases exhibit, as long as only the supply side is affected by any change, both estimation methods go hand in hand. However, as soon as there is some dynamism within the livestock and poultry sector, or even when human consumption preference shifts occur, the new methodology will capture these effects in contrast to the old, bringing new information on feed use to the table.

Food

Food use, in the definition of the FAO food balance sheets, covers food availability up to the retail level for a given reference period, i.e. typically one calendar year. This means that the FBS food variable is equivalent to the amount of food available for consumption at the retail level, rather than the amount consumed. It is typically considerably higher than actual consumption, in fact it exceeds consumption through food waste and losses at the retail and household level, i.e. food that ends up as table scraps, pet food, or is simply thrown away.

The FBS food consumption variable refers to the food available to the resident population of a country. The estimates for population come from the United Nations Population Division (UNPD) and include, to the extent known, also migrants, guest workers and refugees. However, they do not include tourists. This means that food consumed by tourists, or rather, food available to them, should not be included in food consumption. In the past, an ad hoc allowance of food consumed by tourists was included in other uses. The new approach makes this more transparent, as it estimates this directly and presents it as a separate variable. The methodology applied to estimate food available to tourists is presented separately in this document.

Estimates of food use

Harvesting and harnessing actual data

As is the case for all variables, estimates of food use should come from measured information. Such measured information is in short supply for many FBS variables. In the case of food, actual estimates are not only difficult to come by, but the empirically measured observations typically refer to a different definition. Most empirical estimates come from household surveys, and the majority of them come from household income and expenditure surveys (HIES). The information contained therein requires many steps of adjustment and even a comprehensive set of adjustments will not render a fully compatible definition. Adjustments that can be made include:

- Adjustment of the reporting period to a calendar year
- Adjustments of expenditures to quantities (applying appropriate prices)
- Intra-household distribution of consumption
- Inclusion, where necessary, of food consumed outside home, i.e. food consumed in restaurants, canteens and cafeterias, street food, etc.
- Adjustments for waste at the retail level

There are also adjustments that are more difficult to make, at least without the risk of introducing major inaccuracies. As the FBS cover all food disappearance in a country, they also include food consumption/disappearance in public households such as hospitals, prisons, the military, and so on. To make allowances for consumption in these entities is challenging, particularly when the adjustments go beyond the overall average levels, i.e. breaking consumption down for the 60 commodities contained in the FBS.

Creating expected values for time t

The basic idea of the previous FBS methodology was to derive food availability from the intersection of supply and utilization, i.e. food as a residual/balancing item. However, this principle could not, or at least not always, be implemented in practice. The reasons for this have been provided at the beginning of this document, the most important being the fact that food as a balancing item would have assumed the measurement errors inherent in all variables of the balance, and thus resulted in vast year-to-year fluctuations in food consumption. Clearly, such fluctuations would not only be incompatible with available evidence for measured food consumption data, but would also be at odds with the tenets of economic theory. Consumers would indeed make every effort to smooth and eventually stabilize consumption. Necessary adjustments would be assumed by other, more price/income-elastic forms of consumption, or in most cases, by changes in stocks.

This suggests that there are two main motivations for the new FBS food estimation procedure. First, food consumption should evolve gradually, ideally along economic variables such as changes in incomes and consumption patterns. Second, the FBS food classification seems to provide a good basis for such an economic approach; the example of distinguishing butter from solids-non-fat is a good point in case.

The approach used to calculate new expected values for food consumption is rather straightforward. It rolls out food consumption in year t as a function of income changes and trend factors, the latter reflecting things such as changing preferences or known, sudden supply disruptions.

$$Food_t = f(Food_{t-1}, \Delta Y, t) \qquad \text{(Equation 18)}$$

Whereby ΔY represents an average change in per capita income, and t the estimate for a simple time trend factor. Three different functional forms have been distinguished: a log-log specification, a semi-log, a log-inverse, and a log-log-inverse function. The choice of the functional form was taken in line with the functional form that was used for the estimation of the underlying income elasticities. All equations for all commodities and all countries have been parameterized with an income elasticity ε_j for every commodity j, country i, and a trend factor t_j.

log-log specification:

$$log\left(FoodPC_{t,ij}\right) = FoodPC_{t-1,ij} + \varepsilon_{ij}\ \ log\left(\Delta Y_{i,j}\right) + t_i * FoodPC_{t-1,ij}$$

(Equation 19)

semi-log specification:

$$\mathrm{FoodPC_{t,ij}} = \mathrm{FoodPC_{t-1,ij}} + \varepsilon_{\mathrm{i}}\ \ \log\left(\Delta \mathrm{Y_{i,j}}\right) + \mathrm{t_i} * \mathrm{FoodPC_{t-1,ij}}$$

(Equation 20)

Log-Inverse specification:

$$\log(\mathrm{FoodPC_{t,ij}}) = \mathrm{FoodPC_{t-1,ij}} + {}^{\varepsilon_{\mathrm{i,j}}}\!\Big/_{\Delta \mathrm{Y_{i,j}}} + \mathrm{t_i} * \mathrm{FoodPC_{t-1,ij}}$$

(Equation 21)

where $\Delta Y = \frac{GDP\_PC_t}{GDP\_PC_{t-1}}$

Initial values for food availability

The approach presented above leaves the important question open as to what the initial value of food use is and how it was derived. In other words, there is no "bootstrapping" that would allow rolling forward food availability by applying income growth rates. More importantly, if such an approach were valid and possible, it would replace the entire FBS procedure as next year's food consumption levels could simply be derived by predicting it from last year's level, an appropriate income elasticity and a change in income. In practice, such concerns are unwarranted, simply because these values are only initial values for the food consumption in time t. They will be over-written in the balancing mechanism, at least within the range of the assumed measurement error. In other words, the approach tries to capture the impact on food use stemming from population and income dynamics before the balancing algorithm consolidates that information with the overall commodity dynamism on country levels.

Finally and as assumed for all variables of the FBS, official estimates have precedence over imputed food consumption. Where official country estimates are available, they will be taken into the balancing process without a measurement error (see chapter on the balancing mechanism). The only exception to this rule is where countries provide estimates for processed (say flour) but not for primary food products (say wheat). In this case, it was decided to use the official estimates of processed food and work back to an equivalent level of primary food in the standardization process. This decision is based on a careful review of the extraction rates used in the standardization process, the results of which suggest that many of the implicit extraction rates are out of plausible ranges. The implicit extraction rate in this example is the ratio between primary wheat and wheat flour. This means that the food estimates are derived by dividing the amount of processed product reported by a country by a reasonable extraction rate (for details see section on standardization).

Tourist Consumption

Rationale

Creating sound Food Balance Sheets is a particularly challenging task for countries where the food available is not consumed in large measure by non-resident population. In the case of migrants and refugees, available food is assumed to be available to all population groups, and all population groups including refugees and migrants are assumed to be captured by UNPD estimates. This approach cannot be used for countries where non-resident populations are not included in the UNPD estimates. This mainly applies to tourists and consumption by tourists, and is a particularly important challenge where tourists account for a large share of the population and of food consumption. Moreover, consumption patterns of tourists are not necessarily congruent with those of the resident population.

The previous methodology

In the previous system, tourist consumption was accounted in the variable termed "Other uses"; but as it was not identified explicitly, both the amount and the composition of food consumed by tourists are difficult to gauge. As there is no methodology available, the amounts and the composition are difficult to reproduce.

Data availability

The basic data to estimate consumption by tourists is available from the United Nations World Tourism Organization (UNWTO), providing tourist flows between all pairs of countries. Unfortunately, no data is available regarding the consumption patterns of these individuals while they travel.

The new methodology

The "Other uses" category has been removed in the new methodology and replaced with two categories: tourist consumption and industrial uses[28]. This was done in part to move away from the temptation to use other uses as a sort of statistical discrepancy, but it was also done to more accurately account for and model global food consumption patterns. In particular, small countries/islands that are popular tourist destinations could be radically misrepresented, if tourist consumption is not accounted for. On the other hand, larger countries (i.e. countries with low tourism counts in comparison to their total population) will not be affected much by this element.

The methodology is rather straightforward: we have data on the number of day visitors, N_D, and overnight visitors, N_O, to and from each country, and we have information on the average number of nights stayed within each country, $\bar{D}$. The first step, then, is to compute the total number of "tourist days", N_T, from and to each country by adding the day visitor counts with the product of the overnight visitor counts and the average nights per visitor:

$$\boldsymbol{N_T} = \boldsymbol{N_D} + \boldsymbol{N_O}\boldsymbol{\bar{D}} \qquad \textbf{(}\text{Equation 22)}$$

[28] Plus residual other uses, where needed

In lieu of better information, we make the assumption that tourists follow the same consumption patterns abroad as they do at home, both in terms of quantity and preference. Thus, we multiply the total number of tourist days by the average daily consumption within the country of origin and allocate this amount to tourist consumption in the destination country. Moreover, we deduct this total from the food consumption in the country of origin, as the tourist will not be at home to consume these calories. Thus, we have the change in amount of food availability for commodity i in country j as

$$\boldsymbol{\Delta TC_{ij} = -\sum_{k=1,k\neq j}^{m} N_{jk} f_{ij} + \sum_{l=1,l\neq j}^{m} N_{lj} f_{il}} \qquad \text{(Equation 23)}$$

where N_{lj} represents the number of tourists travelling from country l to country j and f_{il} represents the historic amount of daily nutrients consumed within commodity i and in country l. This equation can be simplified a bit:

$$\mathbf{\Delta TC_{ij} = -N_T f_{ij} + \sum_{l=1,l\neq j}^{m} N_{lj} f_{il}} \qquad \text{(Equation 24)}$$

Average daily consumption is computed based on historic consumption patterns, and thus we can provide tourist consumption data at the full FBS level. The measurement error can likewise be derived based on the measurement error of food consumption of the resident population.

Industrial use

The 2000s have seen a number of important changes in the global commodity landscape. Many of them have originated outside of food and agriculture, but many have had a direct and sometimes massive impact on food and agricultural markets. Rising energy prices, in particular, have made a growing number of agricultural products competitive inputs in non-food markets, and have done so in growing volumes (Figure 22). Arguably, the most important example is the biofuels market, where rising energy prices in conjunction with policies to promote the use of agricultural feed stocks for energy uses have diverted a growing amount of (food) commodities into the fuel market. Maize for bioethanol in the US and rapeseed oil for biodiesel in the EU are the most popular and most important examples, but by far not the only ones. With rising energy prices, many countries have embarked on similar programmes, albeit to a smaller extent.

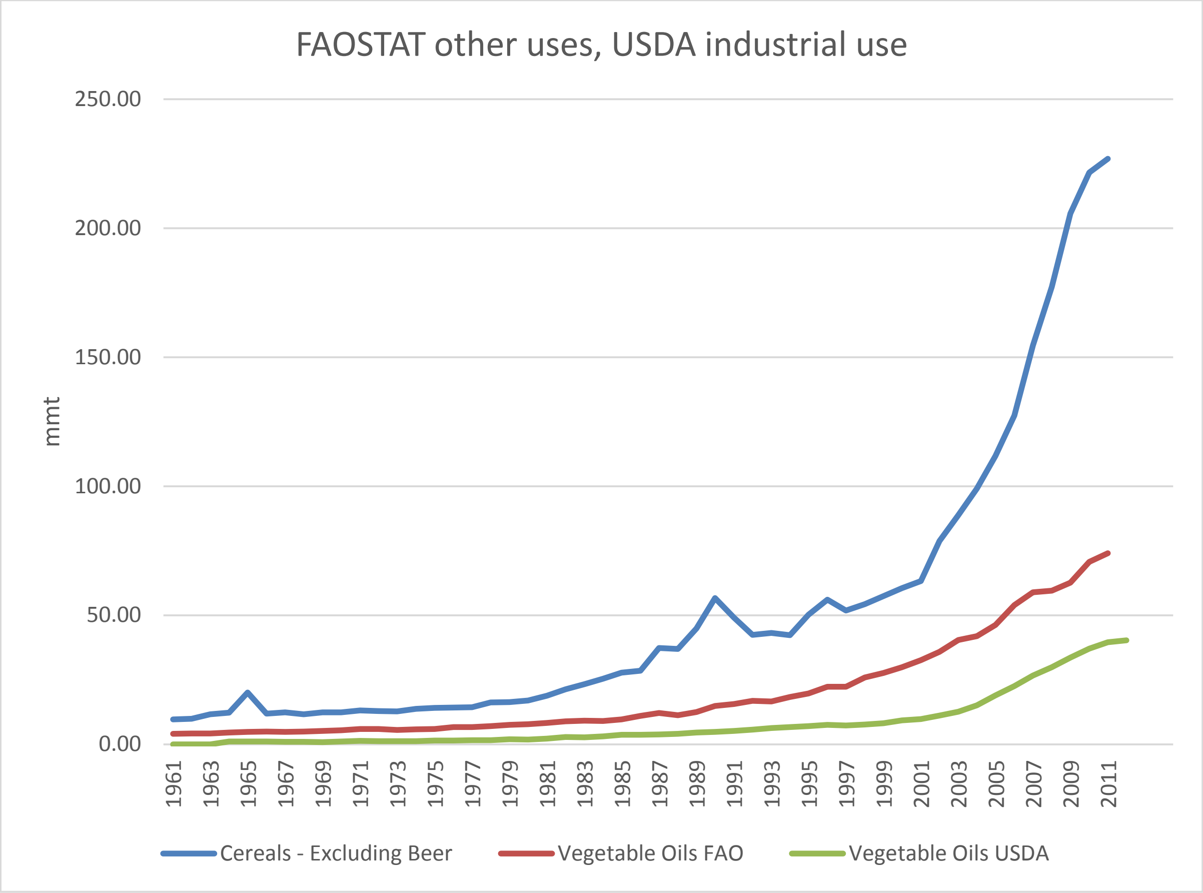


Figure 22: Vegetable oils: Other uses of FAOSTAT v industrial use of PSD

Much less prominent, but equally important, is a similar trend for industrial use outside the energy sector. Just like the biofuels market, this market affects both starch-rich and oil-rich food commodities. For example, a rapidly rising share of coconut oil and palm kernel oil are now diverted into the production of cosmetics, while many other vegetable oils have become key ingredients for the production of paints, soaps, and other detergents. Similarly, starch-based products have become increasingly important as construction materials and are being used for other, non-food purposes. The non-food use of vegetable oils is particularly important for the food balance sheets. These uses can easily be underestimated in quantity terms and given their high caloric content (9kcals/g), the impact on food availability and the DES is particularly high.

Unfortunately, the great variety of different uses (energy, paints, detergents, cosmetics, etc.) makes it difficult to gather actual information. What is more, there is no straightforward way of imputing information. Information on industrial uses is included in FAO questionnaires; however, very few countries provide information. The only area where information is readily available is the use of agricultural feedstuffs for biofuels production. This information is collected by the FAO for its medium term outlook work and is taken into account in the new version of the FBS. In the interim, information available from the USDA PS&D database provides the necessary information on industrial use of vegetable oils. The amount of calories implicit in these uses is considerably above the estimates of the FBS system, and will therefore affect food availability and the DES.

Past and new approach

The existing FBS system does not separately identify industrial use. It is instead part of "other uses", together with tourist consumption and other residual forms of use. The difference between industrial use and other uses can also be gleaned from figure xx, simply by comparing the industrial use of vegetable oils from USDA (PS&D)[29] and other uses of vegetable oils, as available from the FBS. The latter is nearly twice as high as the former and even when tourist consumption is deducted, other uses in the FBS system remain much above the sum of industrial use and tourist consumption. This implies that a considerable amount of "residual use" is included in the FBS variable "other uses"[30].

This also means that the previous approach included a number of different elements, not all of which have been identified clearly. The fact that other uses are higher in many countries than the sum of the two means that the variable other uses has also been used to absorb other, not always clearly identified, residual uses. This means that there are now three separate elements to be distinguished, namely industrial use, tourist consumption and residual other use.

Both tourist consumption and industrial use enter the new approach with their expected means and a measurement error. The approach for measuring tourist consumption is laid out in section xx of this document. Estimates for industrial use come from FAO questionnaires, the

[29] https://apps.fas.usda.gov/psdonline/

[30] For cereals such a comparison is not possible as the USDA PSD database does not provide separate information for industrial use.

OECD/FAO database used for their medium-term outlook, the USDA PSD database and other sources as available and suitable[31]. This leaves only the residual other uses to be defined.

Residual other uses (ROU)

The overall strategy was to identify, as far as possible, the contributing factors to the existing variable "other uses". Industrial use and tourist consumption have been singled out so far. Juxtaposing the existing aggregate "other uses" with the sum of industrial use and tourist consumption revealed that there is a remainder leftover in many countries, and consequently for the world as a whole. These residual other uses form the difference of other uses identified so far and those elements that have been identified separately, i.e. industrial use and tourist consumption.

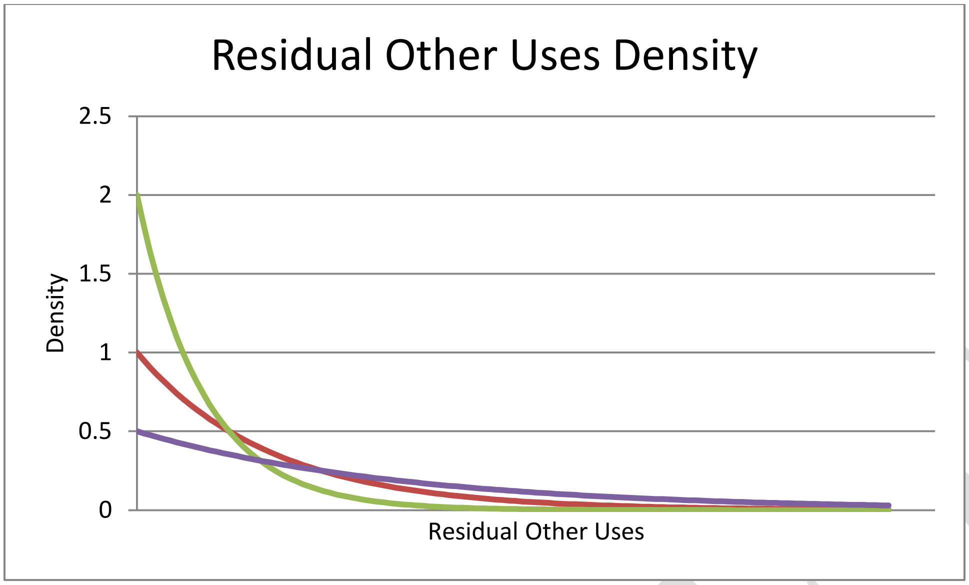
$$\boldsymbol{ROU_{i,j} = OU_{ij} - \left(IU_{i,j} + T_{i,j}\right)} \qquad \text{(Equation 25)}$$

For the balancing mechanism, we use this difference to be the expected value for $ROU_{i,j}$. We further assume that these residual other uses will follow an exponential distribution:

$$f_{ROU}(x) = \lambda e^{-\lambda x} \qquad \text{(Equation 26)}$$

A visual of this distribution is shown below for several different values of λ. This distribution has a mode of zero, suggesting that we expect Residual Other Uses to be zero. However, it has positive density to the right of zero, indicating that other positive values are allowed (and may be chosen if the balancing algorithm chooses).

[31] Overall, the rapid growth of industrial use suggests more efforts be placed on data collection at country level, both through statistical offices and/or direct industry contacts.


Residual Other Uses Density
2.5
2
1.5
1
0.5
0
Density
Residual Other Uses

Food losses and waste (FLW)

Definition of FLW in the FBS system

The FAO FBS system covers food availability up to the point of food purchases, typically the retail level. In line with the definition of food availability, food losses cover all losses up to the same level, i.e. up to the retail level (see Figure 23). Losses at the retail level or even the household level are not included in the FBS system. Postharvest losses (PHL) and losses that occur during storage and transportation as well as processing are included. Processing losses are estimated in the standardization process (see section on standardization), whereby a small allowance is made when extraction rates are applied to primary products. Figure 23 below provides an overview of losses across the entire value chain and delineates losses covered in the FBS system from losses outside the FAO food balances.

Data availability

The production questionnaire (see section on production) includes a section for countries to report estimates of waste and losses. While the vast majority of FAO member countries report production and/or productivity data, estimates for losses are notoriously scarce and, where they are made available by countries through the questionnaire, they are unlikely to be based on reliable measurement efforts. Indeed, the Global Strategy to Improve Rural and Agricultural Statistics found that there is not even a reliable method available to measure losses in an accurate manner. It therefore committed a study to devise an experimental design that could ultimately be used to measure food losses and waste in an accurate manner. This means that available estimates are not only limited, but the few available may not be fully reliable and comparable across countries.

These concerns become visible in the data on losses and waste reported by countries. They can differ, even as percentage loss rates, i.e. as shares of production and trade, considerably even across similar countries. A cross-country comparison has found such differences both for individual commodities, as well as for overall losses at country level estimate. No doubt, this is not an optimal basis to develop an imputation approach, but in the absence of better information, this is the only available basis to start to develop a new imputation method. The basic elements of the new method are laid out below. Before detailing the new method, a quick review of the previous approach is in order.

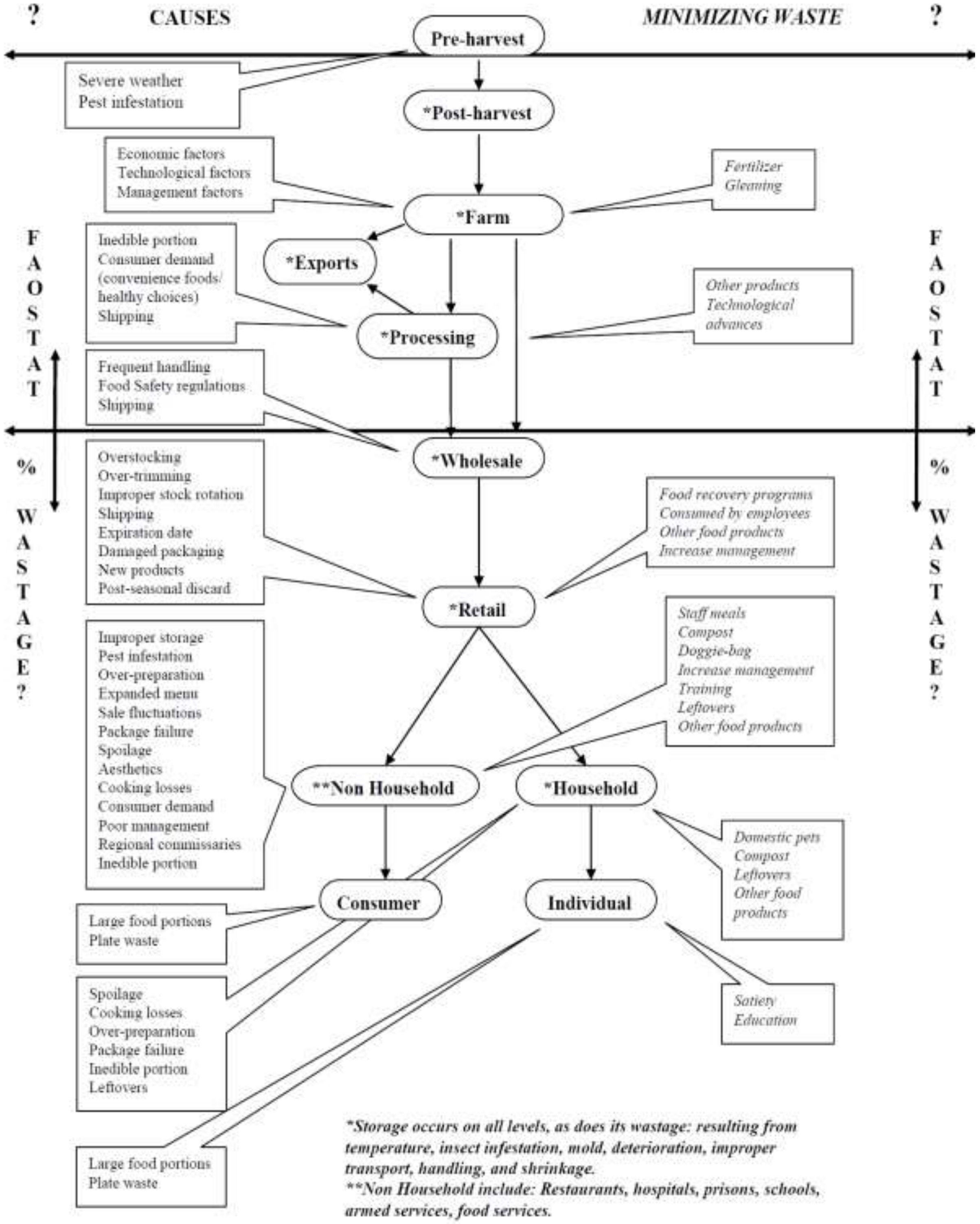


Figure 23: Flow chart of food losses and wastages

Source: Research Report on the Assessment of Intra-household Wastage of Food. FAO, ESS, 2004.

The previous imputation approach

Just like feed use, losses and waste have been estimated as a simple ratio (share) of availability; this means that the amount of food loss of commodity i in country j ($Lo_{i,j}$) is:

$$Lo_{i,j} = \left(P_{i,j} + I_{i,j} - X_{ij}\right) * \ r_{i,j} \qquad \text{(Equation 27)}$$

Again, similarly to the previous approach to estimate feed use, no differentiation is made between trade and domestic production, i.e. a uniform loss rate is applied to all elements of availability.

The rationale behind the choice of this approach is difficult to gather ex-post. On the face of it, it may have been motivated by the need to capture the dynamics of losses in a sub-optimal handling and storage system, whereby losses rise where and when supplies are ample, and are low, where and when supplies are limited. This could be a good approximation of a developing county's food economy, where losses are small in times of crop failures but rise often sharply, in years with bumper crops[32].

The new imputation approach

Given the simplicity of the old imputation method and notwithstanding the data limitations of official data on losses and waste, it was decided to develop a new method to estimate FLW from officially reported data. This meant in a first step to discard all previously imputed FLW data and base the new imputation methods solely on officially reported data.

The scarcity of official data made it necessary to pool available data into a panel. It was also felt necessary to distinguish losses that occur for domestic products from trade products as there are strong a priori reasons to assume that the latter are generally lower. The main reasons for lower trade losses are the fact that PHL do not apply to trade and that traded commodities are likely to be handled and processed in more advanced systems, which are typically less prone to losses.

The model/estimation procedure

Due to the reasons above, a hierarchical linear model (HLM) was developed with the loss quantity as response. The strength of the approach is its ability to pool together different levels of information for the best inference and optimal prediction.

Modelling is performed in a hierarchical fashion, with country/commodity specific estimates at the lowest hierarchical level followed by commodity specific estimates, food group estimates, and lastly food perishable group estimates. The coefficients in the hierarchy are estimated

[32] Such a rationale, however, would not be captured by a linear relationship as inherent in the current imputation method. Linearity would only be justified within the limits of storage and handling capacities and would rise faster than linearly once these capacities are exhausted.

simultaneously to ensure they are consistent between and within the hierarchy. The specific model employed is the following:

$$\log\left(\mathrm{Loss}_{\mathrm{ijklm}}\right) = \alpha_0 + \alpha_1 \mathrm{t} + \alpha_{2\mathrm{ijkl}} \log\left(\mathrm{Production}_{\mathrm{ijklm}} + 1\right) + \mathrm{A}_{\mathrm{ijklm}}$$

$$\begin{gathered}
\alpha_{2\mathrm{ijkl}} = \beta_{20} + \beta_{2\mathrm{ijk}}(\mathrm{Country{:}\ Commodity})_{\mathrm{ijkl}} + \mathrm{B}_{\mathrm{ijkl}} \\
\beta_{2\mathrm{ijk}} = \gamma_{20} + \gamma_{2\mathrm{ij}}(\mathrm{Commodity})_{\mathrm{ijk}} + \mathrm{C}_{\mathrm{ijk}} \\
\gamma_{2\mathrm{ij}} = \delta_{20} + \delta_{2\mathrm{i}}(\mathrm{Food\ Group})_{\mathrm{ij}} + \mathrm{D}_{\mathrm{ij}} \\
\delta_{2\mathrm{i}} = \zeta_{20} + \zeta_{21}(\mathrm{Food\ Perishable\ Group})_{\mathrm{i}} + \mathrm{E}_{\mathrm{i}}
\end{gathered}$$

(Equation 28a-e)

where the i_{th} index represents the effect of food perishable group, the j_{th} index the effect of food group, the k_{th} index the effect of commodity, the l_{th} index the effect of country/commodity, and the m_{th} index the individual observations. The capital letters represent random errors, and these are all assumed to follow normal distributions with some unknown variance. The Greek letters indicate coefficients to be estimated from the data.

The utility of this model is that if a coefficient is missing due to non-availability of official data for estimation, the algorithm will impute losses based on the next level of hierarchy. This means that if loss estimates are missing for a specific country and commodity, the global loss rate for that commodity is applied. If that is not available, the global loss rate for the food group of that commodity is used, and if that is even still not available, the loss rates for perishable food is used. However, if large quantities of data for a particular commodity within a particular country are available, these data are used for imputation without moving up the hierarchy.

The particular advantage of the model lies in the fact that a single model will suffice, where the common approach usually demands several models to impute when a single model fails. In addition, when separate models were estimated individually, they do not incorporate the hierarchical information present in the data.

Seed use

A brief review of the existing approach

The current definition of seed use includes "all amounts of the commodity in question used during the reference period for reproductive purposes, such as seed, sugar cane planted, eggs for hatching and fish for bait, whether domestically produced or imported". Account is taken of double or successive sowing or planting whenever it occurs. Seed use also includes, at least when and where available, the quantities necessary for sowing or planting of crops for fresh use of fodder or food (e.g. green peas, green beans, maize for forage). On average, the amount of seed needed per hectare planted in any given country does not greatly vary from year to year.

Seed data are collected through a special section in the FAO production questionnaire; where no information is provided by countries directly, seed use information is collected through the websites of the relevant national authorities, i.e. the National Statistical Offices or the Ministries of Agriculture. Where neither of these sources provide official data, seed estimates are calculated or estimated either as a percentage of production (e.g. eggs for hatching) or by multiplying the area sown/harvested with a seed rate. Ideally, seed estimates are based on the area sown rather than the area harvested; in practice, however, sparse data available for area sown make it inevitable that seed use estimates are based on the area harvested.

This can be formulated in the simple identity:

$$Se_{i,j} = (AS_{i,j}) * r_{i,j} \quad \text{(Equation 29)}$$

where Se is the seed usage, AS is area sown and r is the seed rate for commodity i in country j. The seed rates vary by both country i and commodity j to account for different agricultural demands in different climates/regions of the world.

Seed rates

The seed rates currently used in the FBS/SUA system are based on an ad-hoc publication entitled "Technical conversion factors for agricultural commodities". It provides, inter alia, seed rates for every country and primary commodity/item of FCL classification. It brings together information on seed rates by country and commodity, reflecting current production practices under different growing conditions. The compilation of seed rates benefited from information provided through the questionnaire replies, as well as from FAO expert advice.

The publication also provides information about the share of eggs that is typically used for hatching[33]. Additional information is provided to better assess the reliability of hatching rates. All hatching and seed rates have been reviewed for the new FBS system and have been changed where necessary.

[33] Only 11 countries have provided official data in recent years.

The new imputation approach

Data for the area sown are collected through the FAO production questionnaire. However, while overall response rates to the questionnaire have been rising, data on seed use remains sparse overall. Where neither seed rates nor official seed use information is available, seed use is imputed. In practice the steps are:

1. Impute area sown, when missing.
2. Apply seed rates as per previous methodology, when questionnaire results are unavailable.
3. Impute seed, when neither area sown nor seed rates are available.

Imputation of Area Sown

To impute the actual area sown, the following approach is taken:

- If previous values of the area sown and the area harvested are available, then an average ratio of the area sown (in year t) to the area harvested (in year t+1) is computed. Then, if the area sown is unavailable in one year, it is imputed by multiplying the area harvested in the following year by the average ratio.
- If no prior information on the area sown is available, than a ratio of 1 is used.

Imputation of Seed

The imputation of seed is performed via a hierarchical linear model (mentioned previously in the loss imputation). The rationale for this model is that it is capable of capturing and modelling complicated trends when data is available. Moreover, the hierarchy of the model allows accurate imputation on countries with very sparse data by pooling together global data. The mathematical model can be written as follows:

$$\log\left(\mathrm{Se}_{i,j,k}\right) = \beta_0 + \beta_1 \mathrm{Temp}_i + \beta_2 \mathrm{Time} + \beta_{3,j,k} \log\left(\mathrm{Area\ Sown}_{i,j,k}\right) + \varepsilon_{i,j,k}$$

$$\beta_{3,j,k} = \gamma_{30} + \gamma_{31,j,k}(\mathrm{CPC\ Code{:}\ Country})_{j,k} + \delta_{j,k}$$

$$\gamma_{31,j} = \kappa_{310} + \kappa_{311}(\mathrm{CPC\ Group})_j + \zeta_j$$

(Equation 30 a-c)

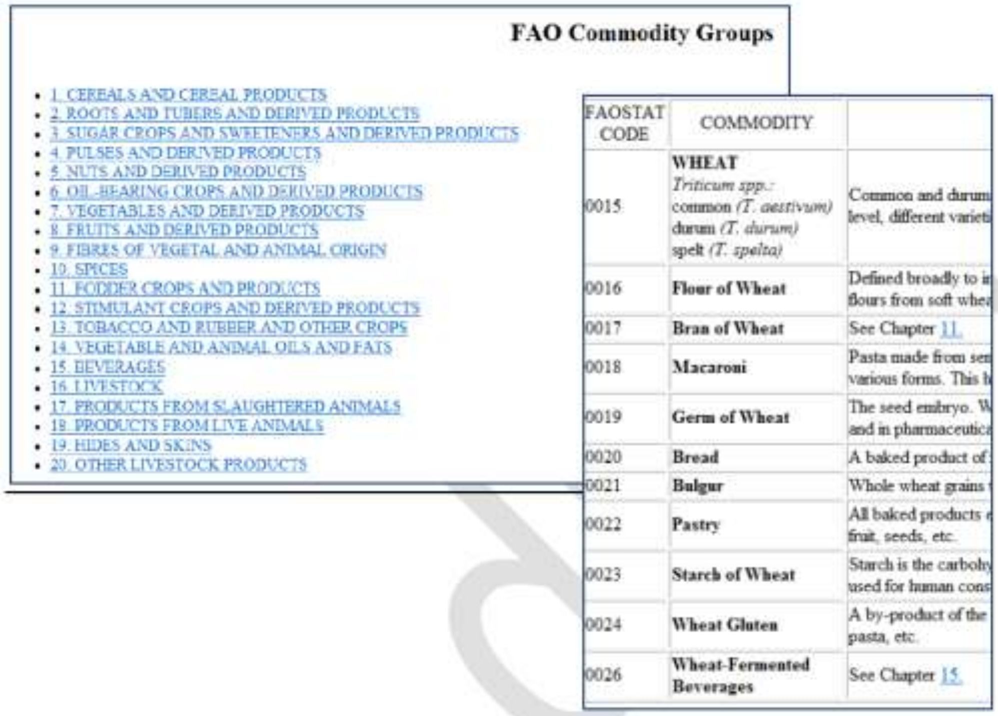
where β, γ, κ are coefficients to estimate from the data, $\varepsilon, \delta, \zeta$ are error estimates, $Temp_i$ is the average annual temperature of country i (provided by the World Bank), and $Time$ is measured in years and is included to capture linear trends over time. The i indices run over all countries, the j indices over all CPC groups, and the k indices over all unique country/CPC code combinations.

Thus, the model estimates seed use proportional to the area sown. The model also accounts for changes over time and differences across countries; the latter are captured by the annual temperature variable, assuming essentially that seed rates need to be higher where production conditions are difficult, with a potential of late and frequent frosts (Russia), and can be lower where production conditions are more favourable (UK).

If data for a particular country and commodity are sparse, then κ_{310} and κ_{311} will likely be estimated to be close to 0. Thus, $\gamma_{31,j,k}$ will be close to its mean value, and the model will only account for availability within commodity groups. However, if data are available for a country or commodity, the estimates of κ_{310} and/or κ_{311} will be far from 0, and thus the model can adapt to the individual characteristics of a particular scenario.

Classifications

The FAOSTAT commodity list (FCL)[34] is the classification of commodities, which has been used in FAOSTAT since the 1960's. Originally it was based on the UN Standard International Trade Classification (SITC)[35]. It includes **683 commodities**, grouped in **20 chapters** (or groups, Figure **24**) and covers crops, livestock and their derived products (Figure 24). It excludes agricultural inputs (such as fertilizers, pesticides and machinery), fishery and forest products, for which different classifications and lists are used in FAOSTAT.



FAO Commodity Groups

- 1. CEREALS AND CEREAL PRODUCTS
- 2. ROOTS AND TUBERS AND DERIVED PRODUCTS
- 3. SUGAR CROPS AND SWEETENERS AND DERIVED PRODUCTS
- 4. PULSES AND DERIVED PRODUCTS
- 5. NUTS AND DERIVED PRODUCTS
- 6. OIL-BEARING CROPS AND DERIVED PRODUCTS
- 7. VEGETABLES AND DERIVED PRODUCTS
- 8. FRUITS AND DERIVED PRODUCTS
- 9. FIBRES OF VEGETAL AND ANIMAL ORIGIN
- 10. SPICES
- 11. FODDER CROPS AND PRODUCTS
- 12. STIMULANT CROPS AND DERIVED PRODUCTS
- 13. TOBACCO AND RUBBER AND OTHER CROPS
- 14. VEGETABLE AND ANIMAL OILS AND FATS
- 15. BEVERAGES
- 16. LIVESTOCK
- 17. PRODUCTS FROM SLAUGHTERED ANIMALS
- 18. PRODUCTS FROM LIVE ANIMALS
- 19. HIDES AND SKINS
- 20. OTHER LIVESTOCK PRODUCTS

FAOSTAT CODE	COMMODITY	
0015	**WHEAT** *Triticum spp.:* common *(T. aestivum)* durum *(T. durum)* spelt *(T. spelta)*	Common and durum level, different varieti
0016	**Flour of Wheat**	Defined broadly to i flours from soft whea
0017	**Bran of Wheat**	See Chapter 11.
0018	**Macaroni**	Pasta made from sem various forms. This h
0019	**Germ of Wheat**	The seed embryo. W and in pharmaceutica
0020	**Bread**	A baked product of
0021	**Bulgur**	Whole wheat grains
0022	**Pastry**	All baked products fruit, seeds, etc.
0023	**Starch of Wheat**	Starch is the carboh used for human cons
0024	**Wheat Gluten**	A by-product of the pasta, etc.
0026	**Wheat-Fermented Beverages**	See Chapter 15.

Figure 24: Printed version of the FAOSTAT commodity list (extract)

The purpose of the FCL is to provide a framework for collecting and analysing data on production and trade and, ultimately, to compile the Supply Utilization Accounts and Food Balance Sheets (SUA/FBS). SUA and FBS "*provide a picture of the pattern of a country's food supply during a specified reference period*"[36] and are at the basis of food security and undernourishment statistics in FAO.

[34] The FCL structure and definitions are available on FAO Statistics Division at: www.fao.org/waicent/faoinfo/economic/faodef/faodefe.htm

[35] http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=28; http://unstats.un.org/unsd/publication/SeriesM/SeriesM_34rev4e.pdf

[36] http://www.fao.org/docrep/003/X9892E/X9892E00.htm#TopOfPage

ESS collects **production data** through a production questionnaire (PQ) that is sent to National Statistical Offices and Ministries of Agriculture around the world on annual basis. Product lists included in PQs are country-specific, which means that the number and type of commodities can vary from country to country. The generic template of the PQ includes **209 primary commodities** (167 crops and 42 livestock) and only **47 processed products** (34 vegetable oils and cakes, 4 dried fruit, 6 alcoholic beverages, 3 sugar products). The classification used is the **FCL** (links to the **CPC** have been added for the last two data collection rounds as further explained in the next paragraphs).

Trade data in ESS are not collected through a questionnaire: countries send ESS their full trade files in **Harmonized System** format; data on food and agriculture are then extracted from trade files. Trade data are eventually converted from HS to FCL format. Once trade data are converted to FCL they are combined with production data to compile SUA/FBS. Trade data are also published on FAOSTAT in FCL format.

The FCL is a tried and tested tool that has been used by ESS for a long time and that has provided a major reference for statistical definitions of agricultural products. However:

- it has never been updated (7 versions of HS since the ‘80s, the same FCL)
- there is no governance mechanism to manage and review it
- no consultation to meet the users’ needs and the developments in the agro-food system has ever been carried out
- codes are not always consistent
- commodity definitions are structured in a way to combine data from the production and the trade domain, which may introduce ambiguities and errors
- it is “isolated”: only used by FAO and not by countries or other organizations; it does not allow integration of agricultural statistics with statistics from other domains

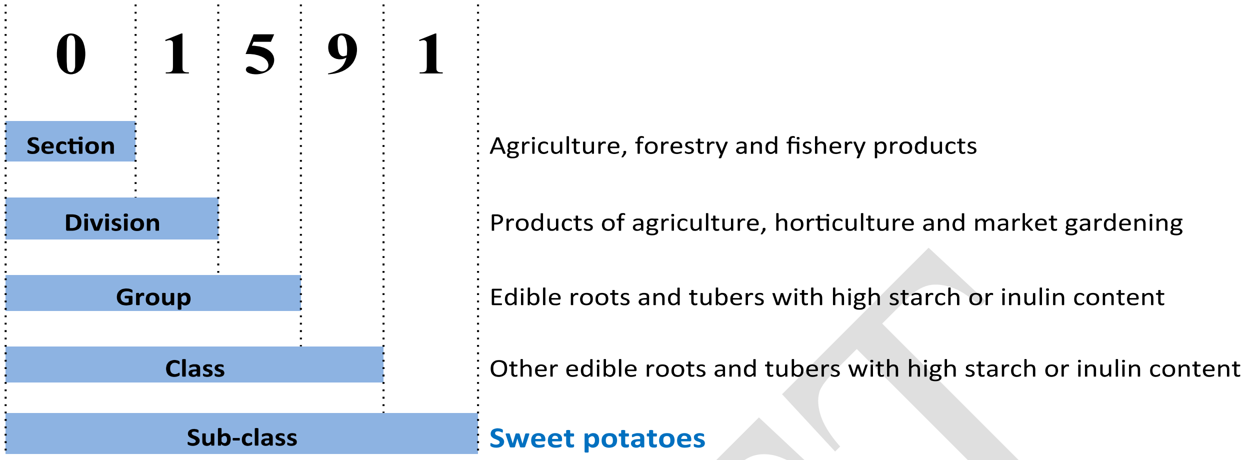

0 1 5 9 1
Section
Division
Group
Class
Sub-class
Agriculture, forestry and fishery products
Products of agriculture, horticulture and market gardening
Edible roots and tubers with high starch or inulin content
Other edible roots and tubers with high starch or inulin content
Sweet potatoes

Figure 25: Codes for “Sweet potatoes” in CPC Ver.2 (0122 in FCL)

For these reasons, in 2011 the decision was taken to replace the FCL with the Central Product Classification of the UN in the new Statistical Working System (SWS). The CPC is a comprehensive classification of products[37], in a system of categories that are both exhaustive and mutually exclusive and based on a set of internationally agreed concepts, definitions, principles and classification rules. The CPC has a five-level hierarchical structure where each digit provides information on product grouping (Figure 25).

The latest CPC version (2.1) will be complemented with an official annex developed by FAO to meet the needs of agricultural statistics; this structure is called “*CPC expanded for agricultural*

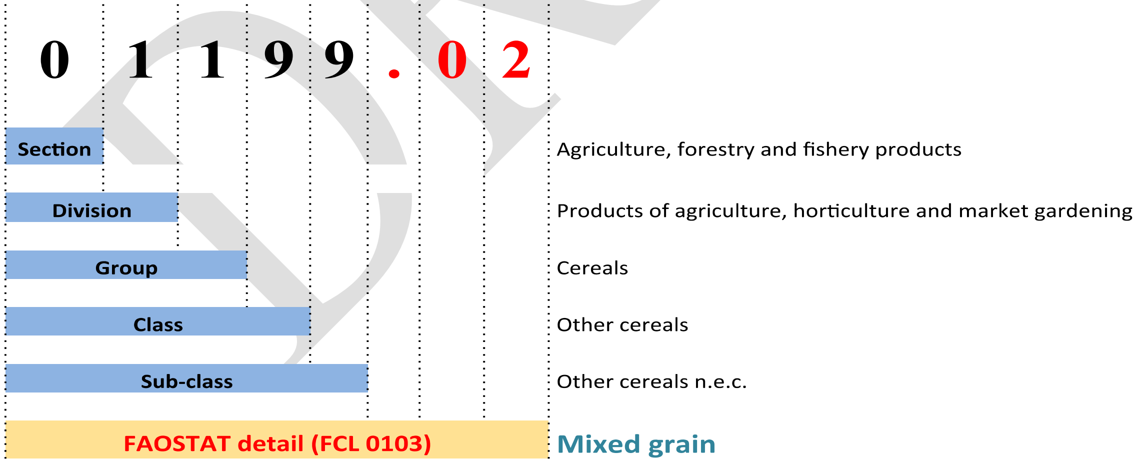

0 1 1 9 9 . 0 2
Section
Division
Group
Class
Sub-class
FAOSTAT detail (FCL 0103)
Agriculture, forestry and fishery products
Products of agriculture, horticulture and market gardening
Cereals
Other cereals
Other cereals n.e.c.
Mixed grain

Figure 26: CPC expanded code for “Mixed grain” (0103 in FCL)

[37] Products follow the SNA definition i.e. *all output of economic activities* that can be the object of domestic or international transactions or that can be entered into stocks including transportable goods, non-transportable goods, services and other products.

statistics". The CPC expanded provides additional details on agricultural commodities (primary products) and is obtained by adding one level (two digits) to the lowest level of the standard CPC[38] (Figure 26).

The CPC classifies products based on the *physical properties and the intrinsic nature* of the products, as well as on the principle of *industrial origin* (harmonized with ISIC - although in some cases products can be the output of several ISIC industries). **HS** subheadings are used as building blocks for the goods part of the CPC: high harmonization with the HS is therefore ensured.

The CPC is a general-scope classification i.e. covers products of all economic activities (not sector specific), but it can be customized for sectoral applications. It is also a general-purpose classification, so that potential applications range from production, to trade, prices and consumption.

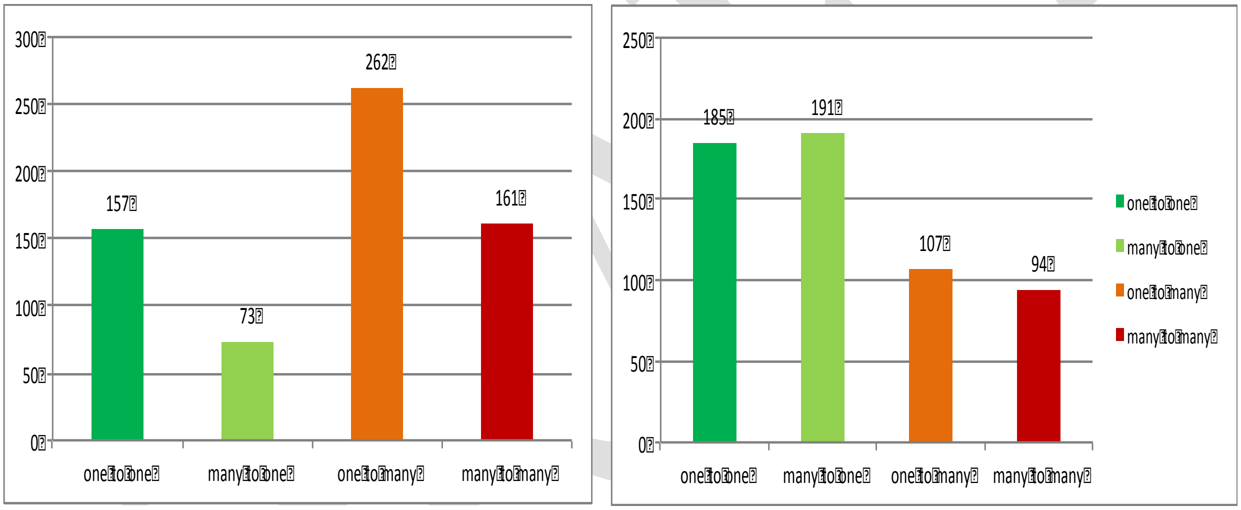


Figure 27: HS-FCL correspondences

- an international classification, constantly updated and reviewed by the Expert Group on International Classifications (chaired by UNSD, participants are countries and international organizations)
- used by other organizations and statistical domains: allows data comparability across statistical domains
- used by countries: reduces reporting burden; in addition the CPC expanded is designed not only for FAO but also for countries engaged in the collection and dissemination of data on agriculture and food products: it provides a flexible tool that allows increased

[38] Valentina Ramaschiello, "*CPC Ver.2 Review and Harmonization with Food and Agriculture Statistics in FAO*". Food and Agriculture Organization of the United Nations. presented at the Expert Group Meeting on International Classifications, UNSD, New York, May 2011, http://unstats.un.org/UNSD/class/intercop/expertgroup/2011/AC234-15.PDF

granularity at the lower level, including local species and varieties, while maintaining comparability across countries at the higher level

- alignment with ISIC and HS: when these classifications are updated, the CPC is also updated; as it is highly aligned with HS, data conversion for SUA/FBS is improved compared to the FCL (65% of the HS07-CPCver.2 correspondences are one-to-one or many-to-one vs. 35% in the case of HS-FCL) (Figure 27)
-

CPC implementation in the new SWS (production domain)

CPC implementation in the new SWS is a long-term activity, which started in 2011. A prerequisite to the implementation of the CPC in the new SWS was to increase the detail on agriculture, forest and fishery products in the CPC. To this end, FAO contributed significantly to CPC ver.2 and ver.2.1. together with UNSD and the Expert Group on International Classifications. In addition, a CPC expanded for agricultural statistics was developed and added as an official annex to CPC version 2.1. When detail in CPC 2.1 expanded is not sufficient, the classification is expanded further in FAOSTAT according to FAO needs and data available. In this way high harmonization between the FCL and the CPC is ensured.

The CPC is planned to be used for future data collection and to be applied to old time-series, in order to allow data comparability over time and avoid breaks in the series.

As mentioned above, countries provide production data to FAO on annual basis, by means of a questionnaire sent by ESS to NSOs and MoAs. In the PQ agricultural products are identified and coded according to the FCL: in view of the change in the classification correlations to CPC codes have been introduced for the last two rounds of data collection.

Although the basic condition for data back cast is to have double coded data for at least one year, it seemed difficult for FAO to increase its data requests to countries: additional burden on national offices might have lowered the response rate and hampered the data collection process. Therefore ESS identified alternative solutions to allow for progress in the change of the classification and data back cast, while reducing the cost of this operation. The solution adopted depended on the type of link encountered and allowed **full alignment between FC and CPC**:

- **One-to-one** cases are resolved quite easily as old data are transferred to the new classification assigning codes and definitions according to the new classification while data remain the same ("key method" [39]).

[39] A classification at the lowest aggregation level is directly recoded to the revised classification. For example, the old code 12345 is recoded to 56789 and the historical data for 12345 are assigned to 56789. This method, also called "*key method*", assures a straightforward relationship between the old and the new results, as the old data are simply transferred to the new classification. The process and outcomes should, however, be documented and communicated to the users. The "key method" is described in Gert Buiten, Jarl Kampen and Sidney Vergouw, 2009, "*Producing*

- Also for **many-to-one** cases data conversion is straightforward as data in the FCL are aggregated into the target classification (CPC). Such an aggregation entails a loss of information, as the CPC is less detailed than the FCL. In order to avoid i losing information in FAOSTAT, many-to-one cases have been turned into one-to-one correlations: first the target classification is expanded further according to the detail available in the FCL and then the "key method" is applied. When detail in the CPC 2.1 expanded is not sufficient, the classification is expanded further for FAOSTAT purpose.

More difficulties are faced for one-to-many and many-to-many types of links. In these cases data are converted based on statisticians' best judgment according to the *predominant* correspondence. Coefficients of conversion have not been calculated, given the lack of information in both formats for at least one year, therefore risk threatening data quality in the conversion. Conversion keys used are 1 and 0 exclusively:

- **One-to-many** relations between the FCL and the CPC are managed identifying the dominant correlation based on statistician's best judgement and assigning the conversion key "1" accordingly.
- In **many-to-many** cases, which represent a minority in the FCL-CPC correlations, the target classification is modified and aligned to the source one.

Details and examples are provided in Appendix 1.

historical time series for STS-statistics in NACE Rev.2", Discussion paper (09001), Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/A8A9AB3B-37F6-480A-BA76-253979DED22D/0/200901x10pub.pdf

CPC implementation in the new SWS (SUA/FBS)

The compilation of SUA and FBS is based on commodity trees. A "commodity tree" in FAO terms should not be confused with a classification tree or "hierarchy".

A **commodity tree** (CT) is a "*symbolic representation of the flow from a primary commodity to various processed products derived from it, together with the conversion factors from one commodity to another*"[40].

A **statistical classification** is "*a set of categories which may be assigned to one or more variables"* where "the *categories are defined in terms of one or more characteristics of a particular population of units of observation. A statistical classification may have a flat, linear structure or may be hierarchically structured, such that all categories at lower levels are sub-categories of a category at the next level up.*"[41]

The FAOSTAT commodity list is a flat classification (or "a list") where commodities are listed following an ascendant order (in most cases). The FCL itself does not set the relations amongst commodities as all categories are on the same level: to distinguish primary from processed products the printed version of the FCL uses capital letters, which is not a classification feature (Box 1). It is the commodity tree that sets the links amongst commodities listed in the FCL through the application of extraction rates. Extraction rates "*indicate, in percent terms, the amount of the processed product concerned obtained from the processing of the parent/originating product, in most cases a primary products*"[42] (Figure 28)[43].

Group 1: Cereals

0070 MILLET

0080 Flour of millet

0081 Bran of millet

0082 Beer of millet

Box 1 Classification of millet and its derived products in the FCL

[40] FAO. 2011. "*Food Balance Sheet: A Handbook*" http://www.fao.org/docrep/003/X9892E/X9892E00.HTM

[41] Andrew Hancock, "*Best Practice Guidelines for Developing International Statistical Classifications*", presented at the Meeting of the Expert Group on International Classifications, UNSD, New York, May 2013 http://unstats.un.org/unsd/class/intercop/expertgroup/2013/AC267-5.PDF

[42] FAO. 2011. "*Food Balance Sheet: A Handbook*" http://www.fao.org/docrep/003/X9892E/X9892E00.HTM

[43] Millet is a simple example as it only includes 1st level processed products

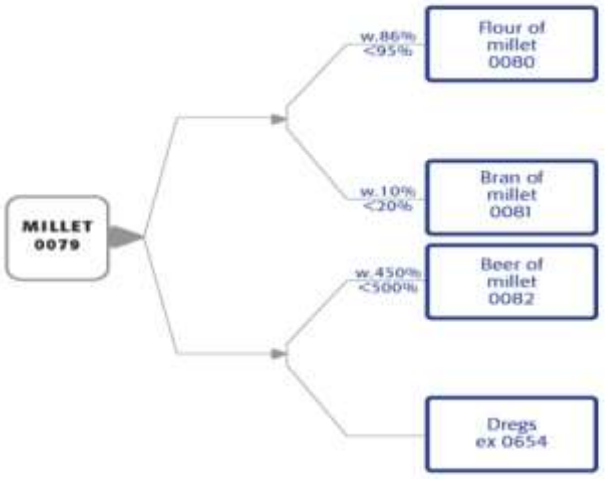


Figure 28: Commodity tree of millet

As far as single commodities are identified in the reference product classification (one-to-one and one-to-many correspondences), commodity trees can be developed. Commodity trees are "independent" from the statistical classification used, or better: their structure does not depend on the reference classification hierarchy. Indeed, relationships set in the trees should not be confused with the classification hierarchy. In a hierarchical classification items at the lower level can be grouped/aggregated into the one at the higher level. For example: millet, wheat, barley and maize can be grouped into the group "cereals" or seeds and grains of milled can be grouped into the class "millet". This is not true for commodity trees where flour, bran and beer cannot be grouped into millet unless quantities are first expressed in terms of primary equivalents, by applying extraction rates. Therefore in CT the key is the *relation* set amongst commodities while commodities are listed in the reference products classification.

The new SWS will be able to manage different trees i.e. CPC classification tree (Box 2) and commodity trees. Through the FCL-CPC conversion table it is possible to translate CT from FCL to CPC (Figure 29).

Box 2: Classification of millet and its derived products in CPC[44];

Section 0: Agriculture, forestry and fishery products
Division 01: Products of agriculture, horticulture and market gardening
Group 011: Cerals
Class **0118: MILLET**

Section 2: Food products, beverages and tobacco [...]
Division 23: Grain mill products, starches and starch products; other food products
Group 231: Grain mill products
Class 2312: Other cereals flour
Subclass 23120: Other cereal flours
FAO Expansion **23120.05 Flour of millet**
Division 24: Beverages
Group 243: Malt liquors and malt
Class 24310: Beer made from malt
FAO Expansion **24310.03 Beer of millet**

Section 3: Other transportable goods
Division 39: Wastes or scraps
Group 391: Wastes from food and tobacco industry
Class 3912: Bran and other residues from the working of cereals or legumes
Subclass 39120: Bran and other residues from the working of cereals or legumes
FAO Expansion **39120.07: Bran of millet**

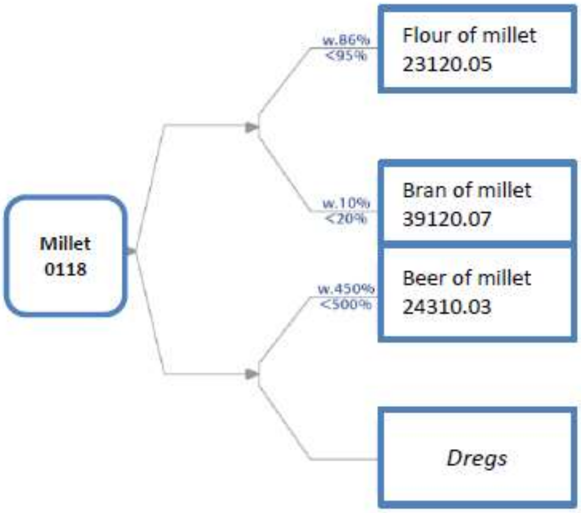


Figure 29: Commodity tree of millet in CPC

[44] CPC hierarchy reflects the economic activity of origin

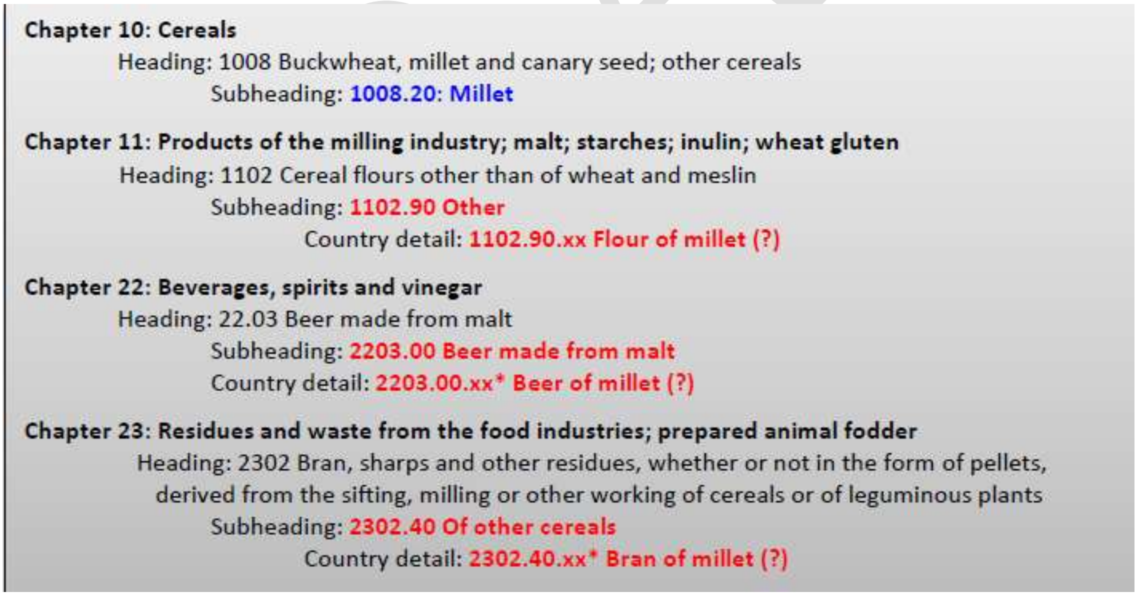
Open questions for the new SWS

Challenges in the compilation of SUA/FBS are related to the data conversion process from HS to FCL. In particular, except for the primary commodity on the left of the tree (209), all related commodities in the trees are processed (474). Out of the 474 processed products listed in the FCL, production data are available for only **47** through the production questionnaire (secondary sources are also used for some other processed products). This means that **for 90% of the processed commodities used in commodity trees (60% of total commodities) official data come from trade only, while production and other variables in the SUA are estimated/calculated/imputed. In terms of classifications, commodity trees are HS-driven.**

Some challenges are related to the use of HS for SUA/FBS, given the current methodology:

○ *Less detail in the HS than the FCL for some commodities.*

For example, details on the cereal of origin for the majority of flours in the FCL is not available in the HS: in the FCL there are 12 different types of cereal flour (wheat, rice, barley, maize, rye, millet sorghum, buckwheat, fonio, triticale, mixed grain, cereals n.e.s.) while in the HS at 6 digits there are only two (wheat and maize flour). Some countries may provide data in national HS format (beyond 6 digits) but how is information obtained when the data are reported at 6 digits only?



Chapter 10: Cereals
Heading: 1008 Buckwheat, millet and canary seed; other cereals
Subheading: **1008.20: Millet**

Chapter 11: Products of the milling industry; malt; starches; inulin; wheat gluten
Heading: 1102 Cereal flours other than of wheat and meslin
Subheading: **1102.90 Other**
Country detail: **1102.90.xx Flour of millet (?)**

Chapter 22: Beverages, spirits and vinegar
Heading: 22.03 Beer made from malt
Subheading: **2203.00 Beer made from malt**
Country detail: **2203.00.xx* Beer of millet (?)**

Chapter 23: Residues and waste from the food industries; prepared animal fodder
Heading: 2302 Bran, sharps and other residues, whether or not in the form of pellets, derived from the sifting, milling or other working of cereals or of leguminous plants
Subheading: **2302.40 Of other cereals**
Country detail: **2302.40.xx* Bran of millet (?)**

Box 3: Classification of millet and its derived products in HS

○ *Heterogeneity in HS beyond 6 digits.*

While the international standard of the HS has 6 digits, trade files received from countries may go beyond 6 digits to meet national needs (8, 10, 12 digits), however this does not apply to all countries and not to all commodities; even when countries use the HS beyond 6 digits, it should be noted that there is no international standard to expand the HS and so each country can do it in a different way.

Recommended actions

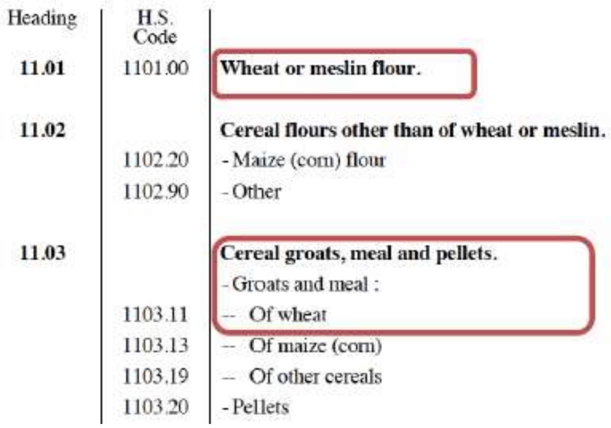
- The HS – FCL map applied so far in the SWS should be documented and checked to make sure it reflects the most recent update in the HS.
- The list and definitions of commodities in the FCL (and therefore those added to the expansion of the CPC ver.2.1) should be verified to make sure they are all relevant and up-to-date.

  For example: the definition of flour of cereals in the FCL not only includes the "fine" flour but also groats, meals and pellets that are classified in different subheading of the HS; all these categories are aggregated in the FCL as flour. However, the extraction rate of flour, groats, meals and pellets may be different, and so their use[45]. The change in the classification should constitute an opportunity to review and either modify or confirm definitions in use so far in the system. (Appendix 2)

[45] **0016** Wheat flour is defined as: "Defined broadly to include meal, groats and pellets. Strong flours from hard wheat are used for bread, while durum wheat flour is used primarily for pasta. Weaker flours from soft wheat are mainly used in cakes, pastries, biscuits and certain noodles."

HS explanatory notes for **1103.11** read: "Durum wheat meal, or semolina, is the principal raw material in the manufacture of macaroni, spaghetti or the like. Semolina is also used directly as a foodstuff (e.g., in making semolina puddings)." Despite the HS definition, the extraction rate of macaroni from wheat flour is 100%.

HS explanatory notes for **1102** "Cereal flours other than of wheat and meslin" read "The heading also covers " swelling " (pregelatinised) flours which have been heat treated to pregelatinise the starch. They are used for making preparations of heading 19.01, bakery improvers or animal feeds or in certain industries such as the textile or paper industries or in metallurgy (for the preparation of foundry core binders)": not all products under this code are edible." It should be verified if the inedible component is taken and considered as edible in CT.



Heading	H.S. Code	
11.01	1101.00	**Wheat or meslin flour.**
11.02		**Cereal flours other than of wheat or meslin.**
	1102.20	- Maize (com) flour
	1102.90	- Other
11.03		**Cereal groats, meal and pellets.**
		- Groats and meal :
	1103.11	-- Of wheat
	1103.13	-- Of maize (com)
	1103.19	-- Of other cereals
	1103.20	- Pellets

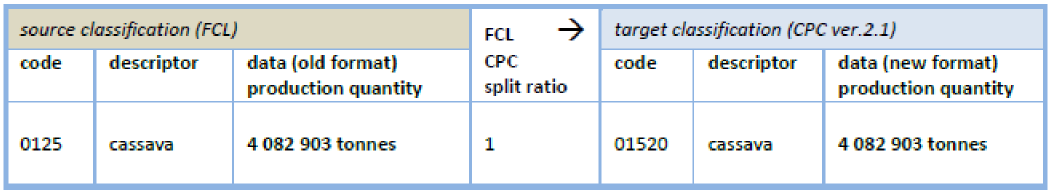
Box 4: classification of flour, groats, meal and pellets in HS 2012 (0016 “Flour of Wheat” in FCL)

Appendix 1

Examples of solutions adopted to convert FAOSTAT data on agricultural commodities from FCL to CPC format

One-to-one cases are resolved quite easily as old data are transferred to the new classification i.e. codes and definitions are re-assigned according to the new classification while data remain the same (Example 1).

Example 1: Data conversion from FCL to CPC in case of one-to-one type of link

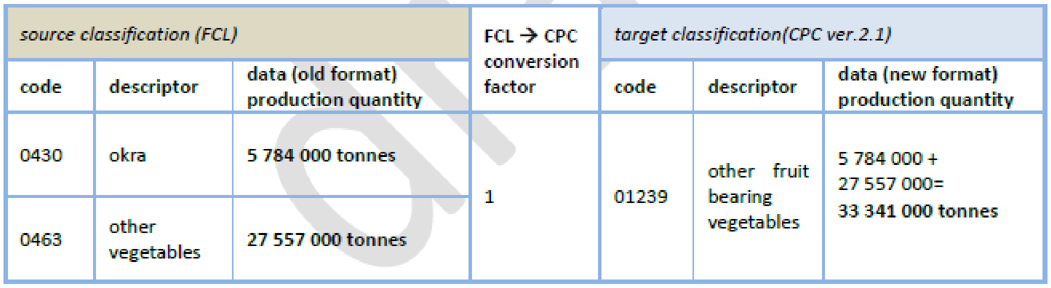


source classification (FCL)			FCL → CPC split ratio	*target classification (CPC ver.2.1)*		
code	descriptor	data (old format) production quantity		code	descriptor	data (new format) production quantity
0125	cassava	4 082 903 tonnes	1	01520	cassava	4 082 903 tonnes

Data are taken as example and refer to the production of cassava in Cameroon, 2011 (source: FAOSTAT)

Also for **many-to-one** cases data conversion is straightforward as data in the source classification (FCL) are aggregated into the target classification (CPC). Such an aggregation entails a loss of information, as the target classification is less detailed than the source one (Example 2).

Example 2: Data conversion from FCL to CPC in case of many-to-one type links

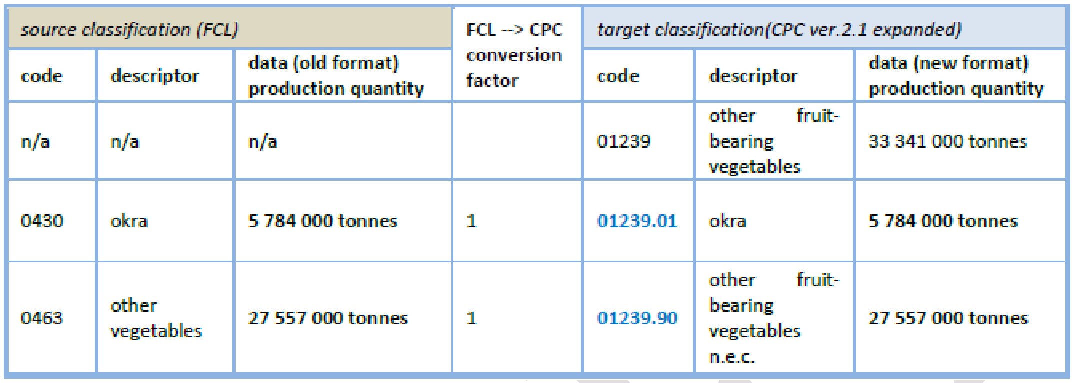


source classification (FCL)			FCL → CPC conversion factor	*target classification(CPC ver.2.1)*		
code	descriptor	data (old format) production quantity		code	descriptor	data (new format) production quantity
0430	okra	5 784 000 tonnes	1	01239	other fruit bearing vegetables	5 784 000 + 27 557 000= 33 341 000 tonnes
0463	other vegetables	27 557 000 tonnes				

Data are taken as example and refer to the production of okra and other fresh vegetables in India, 2011 (source: FAOSTAT)

In order to not lose information in FAOSTAT, many-to-one cases have been turned into one-to-one correlations: first the target classification has been expanded further according to the detail available in the FCL (new CPC expanded codes 01239.01 and 01239.90 in Example 3) and then the "key method" is applied as in Example 1. When the level of detail in the CPC 2.1 expanded is not sufficient, the classification is expanded beyond for FAOSTAT purpose.

Example 3: Data conversion from FCL to CPC in the case of a many-to-one type of link turned into a one-to-one relations (the codes in **bold blue** text are the CPC expanded codes developed by the FAO for FAOSTAT purpose)



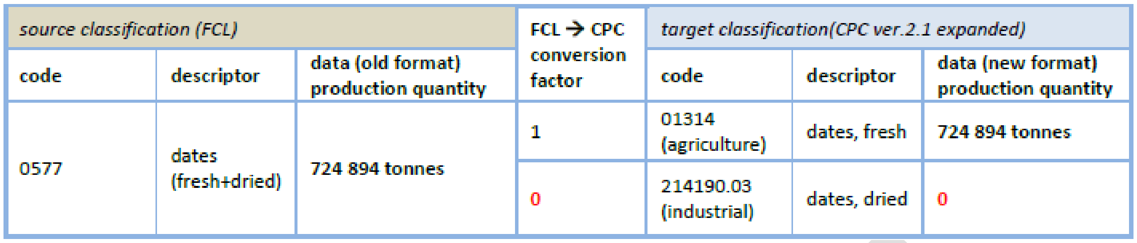
source classification (FCL)			FCL --> CPC conversion factor	*target classification(CPC ver.2.1 expanded)*		
code	descriptor	data (old format) production quantity		code	descriptor	data (new format) production quantity
n/a	n/a	n/a		01239	other fruit-bearing vegetables	33 341 000 tonnes
0430	okra	5 784 000 tonnes	1	**01239.01**	okra	5 784 000 tonnes
0463	other vegetables	27 557 000 tonnes	1	**01239.90**	other fruit-bearing vegetables n.e.c.	27 557 000 tonnes

Data are taken as example and refer to the production of okra and other fresh vegetables in India, 2011 (source: FAOSTAT)

More difficulties are faced for one-to-many and many-to-many types of links. In these cases data have been converted based on statisticians' best judgment according to the *predominant* correspondence. Coefficients of conversion have not been calculated, given the lack of information in both formats for at least one year and, therefore, the risk to threaten data quality in the conversion. Conversion keys used are 1 and 0 exclusively.

One-to-many relations between the FCL and the CPC mainly concern agricultural (primary) vs. industrial (processed) products. For example, fresh and dried fruit in the FCL are sometimes classified together while they are separated in the CPC. This is due to the fact that the CPC is closely linked to the International Standard Industrial Classification of All Economic Activities (ISIC) and dried fruit is considered as an output of the manufacturing industry and not of agriculture. The solution adopted for data conversion in FAOSTAT when dried fruit is not dedicated a specific class (as in the case of dates) is to associate FCL data only to the items in the agricultural section of the CPC, leaving blanks in correspondence with the industrial goods section. In Example 4 below, the one-to-many correlation is converted into one-to-one, assigning the conversion factor "1" to the class that, based on statistician's best judgment, is the best one covering the FCL boundaries (predominant correspondence). In the metadata it will be noted that 01314 may, in some years for some countries, include information on dates dried on farm.

Example 4: Data conversion from FCL to CPC in case of one-to-many type of link



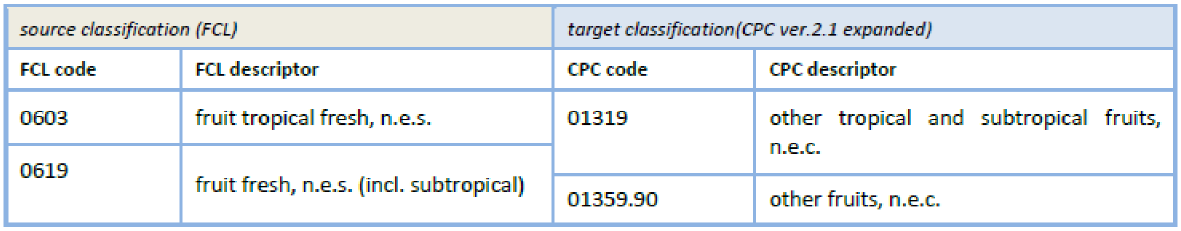
source classification (FCL)			FCL → CPC conversion factor	*target classification(CPC ver.2.1 expanded)*		
code	**descriptor**	**data (old format) production quantity**		**code**	**descriptor**	**data (new format) production quantity**
0577	dates (fresh+dried)	724 894 tonnes	1	01314 (agriculture)	dates, fresh	724 894 tonnes
			0	214190.03 (industrial)	dates, dried	0

Data are taken as example and refer to the production of dates in Algeria, 2011 (source: FAOSTAT)

In **many-to-many** cases, which represent a minority of cases in the FCL-CPC correlations, the CPC is modified and aligned to the FCL.

In Example 5, the FCL includes "subtropical fruit" under "fruit fresh n.e.s." (0619) while in the CPC subtropical fruit is classified with "other tropical and subtropical fruits, n.e.s." (01319). This generates a mismatch between the two classifications. Given the impossibility of estimating split ratios, and not introducing breaks in the series, the CPC is adapted and aligned to the FCL (Example 6): the component "subtropical fruit" in the CPC is moved under "other fruits n.e.c." in FAOSTAT (01359.90). Definitions in the metadata are adjusted accordingly.

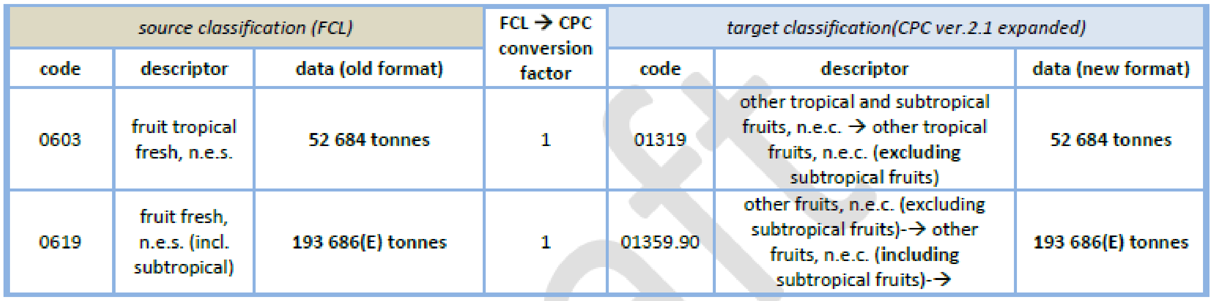
Example 5: Many-to-many relations between the FCL and the CPC concerning tropical, subtropical and other fruit n.e.c.



source classification (FCL)		*target classification(CPC ver.2.1 expanded)*	
FCL code	**FCL descriptor**	**CPC code**	**CPC descriptor**
0603	fruit tropical fresh, n.e.s.	01319	other tropical and subtropical fruits, n.e.c.
0619	fruit fresh, n.e.s. (incl. subtropical)	01359.90	other fruits, n.e.c.

Example 6:

Data conversion from FCL to CPC in case of many-to-many type of link



source classification (FCL)			FCL → CPC conversion factor	*target classification(CPC ver.2.1 expanded)*		
code	**descriptor**	**data (old format)**		**code**	**descriptor**	**data (new format)**
0603	fruit tropical fresh, n.e.s.	52 684 tonnes	1	01319	other tropical and subtropical fruits, n.e.c. → other tropical fruits, n.e.c. (excluding subtropical fruits)	52 684 tonnes
0619	fruit fresh, n.e.s. (incl. subtropical)	193 686(E) tonnes	1	01359.90	other fruits, n.e.c. (excluding subtropical fruits)-→ other fruits, n.e.c. (including subtropical fruits)-→	193 686(E) tonnes

Data are taken as example and refer to the production of tropical fruit n.e.s. and fruit n.e.s. in Ecuador, 2011 (source: FAOSTAT); (E) = FAO estimates

Appendix 2

Proposal for a working group to review definitions in the FCL

As all commodities in the FCL have been duplicated in the CPC ver.2.1 expanded, forming a working group to review the commodities in the FCL and their definitions to ensure that all items need to be kept is recommended. The aim should be to improve definitions and delete redundancies, inconsistencies or obsolete items or components.

The group should be allocated adequate resources and activities should be included in PEMS to ensure participation: work is extremely time consuming and requires dedicated staff. Unsuccessful attempts of working groups have already been made in the past: due to low participation, activities were interrupted.

Out of 683, not all commodities in the FCL need to be verified: to start with the work a subset of most difficult cases has been identified:

- one-to-many and many-to-many cases in the FCL-CPC correspondences
- one-to-many and many-to-many cases in the HS-FCL correspondences