Production

### Abstract

This paper describes the data collection process for production data at FAO. It discusses the types of data collected, the difficulties with this process, and the data availability.

In addition, this paper discusses the production imputation methodology. This methodology is based on ensemble learning and hence is very flexible and able to capture the trends in the data. Moreover, it is able to generate reasonable estimates in all cases (i.e. where data availability is very low or very high).

### Introduction

The FAO Statistics Division collects data on agricultural production via an annual questionnaire. The questionnaire is dispatched to more than 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 gaps of missing data, implying that they have not collected the data in most cases[[1]](#footnote-2).

The lack of data provided in the questionnaires reflects the lack of data available at country level and the inability of countries to compile their own FBSs; countries will face the same problems as FAO in completing the SUA and FBS sheets. It cannot be stressed often enough that the recommended approach to fill the data gaps is to undertake surveys, collect conversion factors, extraction rates and other coefficients. FBS compilers should make every effort to obtain empirically measured information. No imputation methods, no matter how sophisticated, can replace measurement. Needless to say, this also holds for all imputation methods presented in this document. These methods cannot be surrogates for surveys, censuses, and measurement in general. Only in the absence of empirically collected/measured information should missing data/conversion factors/etc. be imputed. But while imputation cannot replace measurement, there are wide differences in the suitability of imputation methods. Good imputation methods are not only necessary to generate missing information, they can also be used to triangulate measured information and potentially discover inconsistencies in measured information.

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 databases (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 **1** 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 1: 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 indicates the number of responding countries while RR refers to the 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, 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

quality checks. These checks include a considerable range of operations, e.g. adjusting units of measurement, weeding out outliers, identifying transcription errors, filling in obvious gaps, and so forth. A general concept towards improved data quality is laid out in the Statistical Quality Assurance Framework (SQAF)[[2]](#footnote-3) of FAO, which is also available to all FBS compilers at country level.

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[[3]](#footnote-4). FAO Statistics Division will then calculate the missing variable (i.e. yield) by applying the simple identity that productivity=production/activity.

### Data imputation: a new approach

Once the QA/QC has been undertaken, outliers weeded out, and data gaps filled with information available from alternative sources[[4]](#footnote-5), the remaining data gaps can be completed through imputation procedures that harness all information available up to this point. To this end, a new imputation method has been developed, which combines a robust 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. Such imputation methods have been examined and, based on this review, the basic insight is that no single approach/model taken by its own is able to render satisfactory results across all countries and commodities. Based on this insight, an “ensemble learning model” was developed, tested and adjusted to the needs of the FAO production domain. The tests have also shown that the particular strength of the ensemble model is in successfully handling the great heterogeneity across different countries; however, FBS compilers at country level may find the same or similar approaches particularly useful for their individual country imputation challenges.

A detailed description of the methodology is available in a separate document. Here presentation is limited to a few main features. First, the imputation of yield in any given country incorporates available information from notionally all countries for that commodity. 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.

The performance of the new imputation method has been tested and reviewed; the 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 four main steps to the imputation procedure:

1. Productivity Imputation
2. Imputation by balancing via Productivity = Production / Activity.
3. Production Imputation
4. Imputation by balancing via Productivity = Production / Activity.

Step 1 will be skipped if productivity is currently available. If balancing is then possible, imputation via balancing will be done. Then, production imputation will be performed if production is still not available. If one variable is still missing, it will be imputed in the final step.

#### 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. It should be noted that an appropriate length of time should be chosen for this imputation: a time series that is too short will not have enough data to learn from trends while a time series that is too long may include unreliable data from many years ago.

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:

(Equation 5)

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

(Equation 6)

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

(Equation 7)

If the non-linear regression model fails to converge, then 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 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.

1. 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.
2. 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.
3. 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.
4. Splines: A cubic spline is fit to the observed values and used for interpolation.
5. MARS: A Multivariate Adaptive Regression Spline, which is a mathematical model appropriate for piecewise linear time series, is fit to the observed values.
6. 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 then 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, , 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:

(Equation 8)

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. Note that in this case, all weights must be adjusted to ensure they still sum to one.

#### 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. Likewise, if productivity and production are available, we impute activity by the identity activity = production/productivity.

#### 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.

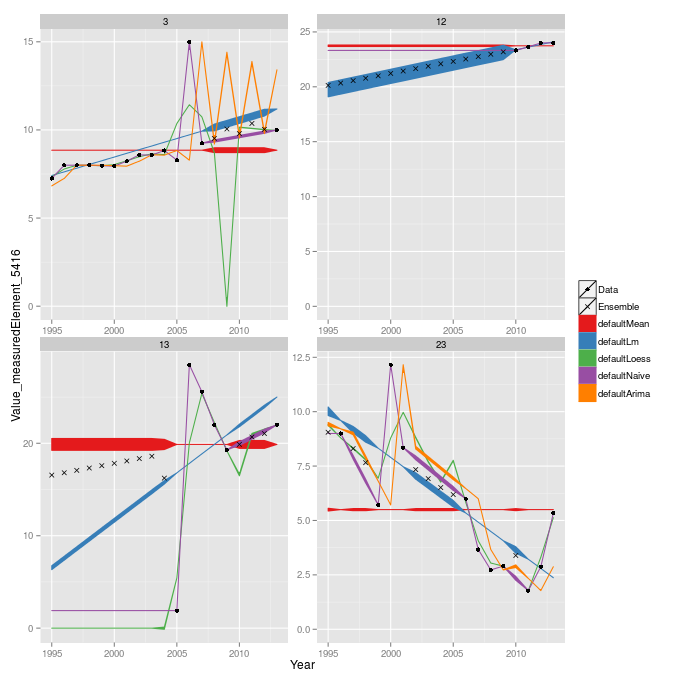
****

Figure 1: A plot of the ensemble imputation method on four different datasets. The circles indicate the known data points, the lines provide the fitted models, and the “x”’s are the imputed values. The thickness of the line represents the weight it was given, and it is evident that weights change in different scenarios.

1. 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. [↑](#footnote-ref-2)
2. <http://www.fao.org/docrep/019/i3664e/i3664e.pdf> [↑](#footnote-ref-3)
3. 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). [↑](#footnote-ref-4)
4. 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. [↑](#footnote-ref-5)