

Imputation Methodology for FAOSTAT Production Domain

Michael. C. J. Kao

Food and Agriculture Organization
of the United Nation

Outline

- 1 Current Methodology
- 2 Case Study and Exploratory Data Analysis
- 3 Proposed Methodology
 - Imputation for Yield
 - Imputation for Area Harvested
- 4 Results
 - Results for Wheat
 - Results for Cassava
- 5 Further Improvements
- 6 Discussion

Why do we need imputation?

The agricultural production domain is integral to the compilation of Food Balance Sheets. In particular to estimate consistent food supplies, imputation is required to ensure that data are non-sparse. Owing to the potential impact of imputation when often data are missing, accuracy and reliability of food estimates cannot be compromised.

The relationship of production and its components can be expressed as:

$$P_t = A_t \times Y_t \quad (1)$$

Where P_t , A_t and Y_t denotes production, area harvested and yield, respectively, at time t .

Outline for section 1

- 1 Current Methodology
- 2 Case Study and Exploratory Data Analysis
- 3 Proposed Methodology
 - Imputation for Yield
 - Imputation for Area Harvested
- 4 Results
 - Results for Wheat
 - Results for Cassava
- 5 Further Improvements
- 6 Discussion

The presently applied methodology aims to capture the variation of relevant commodity and/or geographic characteristics through the application of aggregated growth rates. a five-hierarchy was designated represented by:

- ① Same country/commodity aggregate
- ② Sub-region aggregate/same commodity
- ③ Sub-region aggregate/commodity aggregate
- ④ Regional aggregate/same commodity
- ⑤ Regional aggregate/commodity aggregate

In short, the aggregation imputation method computes the commodity/regional aggregated growth of both area and production, the growth rate is then applied to the last observed value. The formulae of the aggregated growth can be expressed as:

$$r_{s,t} = \sum_{c \in \mathbb{S}} X_{c,t} / \sum_{c \in \mathbb{S}} X_{c,t-1} \quad (2)$$

The imputation can then be computed as:

$$\hat{X}_{c,t} = X_{c,t-1} \times r_{s,t} \quad (3)$$

There are several shortcomings of the current methodology,

- Divergence of area and production, there are mainly two reasons for this.
 - ① Due to missing values, the aggregated growth can be heavily biased.
 - ② The basket used to compute the aggregated growth rate is not the same over time and between area and production.
- Assumes perfect correlation between group and country series.
- Cannot support and incorporate additional information.

Outline for section 2

- 1 Current Methodology
- 2 Case Study and Exploratory Data Analysis
- 3 Proposed Methodology
 - Imputation for Yield
 - Imputation for Area Harvested
- 4 Results
 - Results for Wheat
 - Results for Cassava
- 5 Further Improvements
- 6 Discussion

To illustrate the proposed imputation, we have chosen Wheat from the crop group and Cassava from the root group as case studies. Both are important agricultural products but Wheat is much more commercialised and traded than is Cassava.

We have log-transformed the data to make the relationship an additive one, so that the production can be decomposed.

$$\log(P_t) = \log(A_t) + \log(Y_t) \quad (4)$$

Area dictates the level and changes in the production

Same for Cassava

Let us dig deeper

The co-movement appears weaker

What is the data telling us?

What the data have shown is that the level, trend of the production is mainly determined by a smooth monotonic area occasionally affected by shocks, while the yield generates the variation from year-to-year reflecting climate or economic conditions.

This leads to the proposed methodology to estimate the year-to-year variation of yield while a stable method for area.

Outline for section 3

- 1 Current Methodology
- 2 Case Study and Exploratory Data Analysis
- 3 Proposed Methodology**
 - Imputation for Yield
 - Imputation for Area Harvested
- 4 Results
 - Results for Wheat
 - Results for Cassava
- 5 Further Improvements
- 6 Discussion

First of all, we propose to impute the yield and area, and along with the restriction of the new model this almost guarantees that area and production will not diverge.

Second, instead of applying the changes directly, the model estimates the relationship between the country and the aggregated series and applies the factors accordingly.

Finally the proposed model allows incorporation of additional information such as prices, vegetation indices and other data that may improve the accuracy of the imputation.

Linear Mixed Model

To capture the co-movement of yield and model sub-regional differences, we have proposed to model the yield with a Linear Mixed Model (LME), which can be expressed as follows in matrix notation:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i \\ \mathbf{b}_i &\sim \mathbf{N}_q(\mathbf{0}, \boldsymbol{\Psi}) \\ \epsilon_i &\sim \mathbf{N}_{ni}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i) \end{aligned} \tag{5}$$

More specifically, the equation for the imputation is

$$Y_{i,t} = \underbrace{\beta_{0j} + \beta_{1j}t}_{\text{Fixed effect}} + \underbrace{b_{0,i} + b_{1,i}t + b_{2,i}\bar{Y}_{j,t}}_{\text{Random effect}} + \epsilon_{i,t} \quad (6)$$

The average yield can be calculated as:

$$\bar{Y}_{j,t} = \sum_{i \in j} \omega_i Y_{i,t} \quad (7)$$

Which acts as a proxy to reflect the change in climatic conditions and other factors which can simultaneously affect multiple countries.

Currently we have adopted **linear interpolation** and **last observation carry forward** to impute area harvested.

First the area harvested displays close to monotonic behaviour and much little year-to-year fluctuation and thus linear interpolation is suitable. Furthermore, a previous simulation study has shown linear interpolation gives best result.

While "last observation carry forward" is useful when the last observed value is a true zero, we will not impute a positive value.

We are, however, investigating improvements in the imputation of area harvested using information such as area sown.

Outline for section 4

- 1 Current Methodology
- 2 Case Study and Exploratory Data Analysis
- 3 Proposed Methodology
 - Imputation for Yield
 - Imputation for Area Harvested
- 4 Results**
 - Results for Wheat
 - Results for Cassava
- 5 Further Improvements
- 6 Discussion

Outline for section 5

- 1 Current Methodology
- 2 Case Study and Exploratory Data Analysis
- 3 Proposed Methodology
 - Imputation for Yield
 - Imputation for Area Harvested
- 4 Results
 - Results for Wheat
 - Results for Cassava
- 5 Further Improvements**
- 6 Discussion

- Test multilevel structured additive regression to account for non-linearity in yield.
- Test EM-algorithm to estimate rather than compute the sub-regional average yield.
- Understand and improve the methodology for area harvested.

Outline for section 6

- 1 Current Methodology
- 2 Case Study and Exploratory Data Analysis
- 3 Proposed Methodology
 - Imputation for Yield
 - Imputation for Area Harvested
- 4 Results
 - Results for Wheat
 - Results for Cassava
- 5 Further Improvements
- 6 Discussion

The newly proposed methodology demonstrates the ability to resolve issues in the current methodology and extended to incorporate additional information.

We welcome any information which can enhance the performance of the imputation.

Thank you for your time and attention.