

Imputation Methodology for FAOSTAT Production Domain

Michael. C. J. Kao

Food and Agriculture Organization
of the United Nation

Outline

- 1 Current Methodology
- 2 Proposed Methodology
 - Imputation for Area Harvested
 - Imputation for Yield
- 3 Results
 - Individual imputation
 - Simulation Results
- 4 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 .

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.

In the new methodology, 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.

Currently we have adopted **linear interpolation** and **last observation carry forward** to impute area harvested. This is what we called the naive imputation.

First the area harvested and particularly carcass weight per animal 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.

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:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i \\ \mathbf{b}_i &\sim \mathbf{N}_q(\mathbf{0}, \boldsymbol{\Psi}) \\ \epsilon_i &\sim \mathbf{N}_{ni}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i) \end{aligned} \tag{4}$$

More specifically, the equation for the imputation is

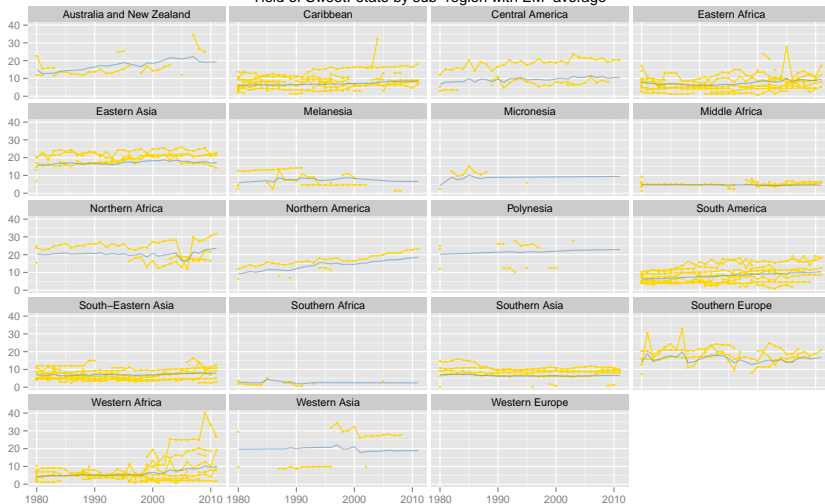
$$Y_{i,t} = \underbrace{\beta_{0j} + \beta_{1j}t}_{\text{Fixed effect}} + \underbrace{b_{0,i} + b_{1,i}t + b_{2,i,t}\bar{Y}_{j,t}}_{\text{Random effect}} + \epsilon_{i,t} \quad (5)$$

The average yield can be calculated as follow,

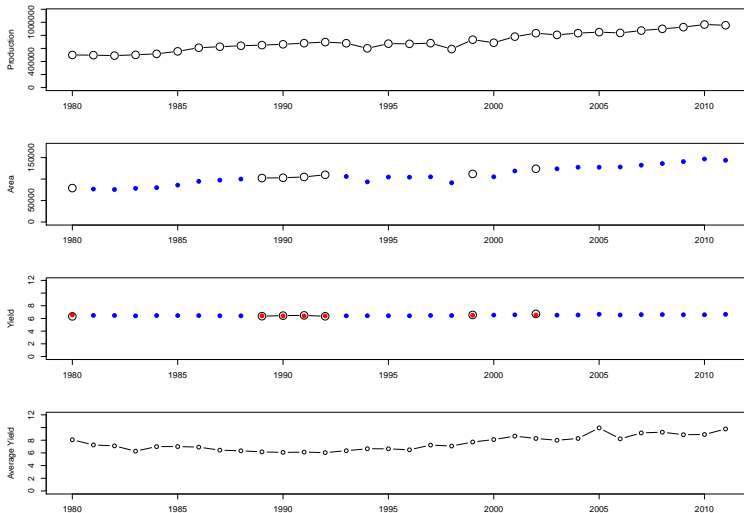
$$\bar{Y}_{j,t} = \sum_{i \in j} \omega_i Y_{i,t} \quad (6)$$

Which acts as a proxy to reflect the change in climatic conditions and other factors which can simultaneously affect multiple countries. However, due to missing values, this quantity is not computed directly from the raw data. The EM-algorithm is implemented for the estimation for the unbiased average.

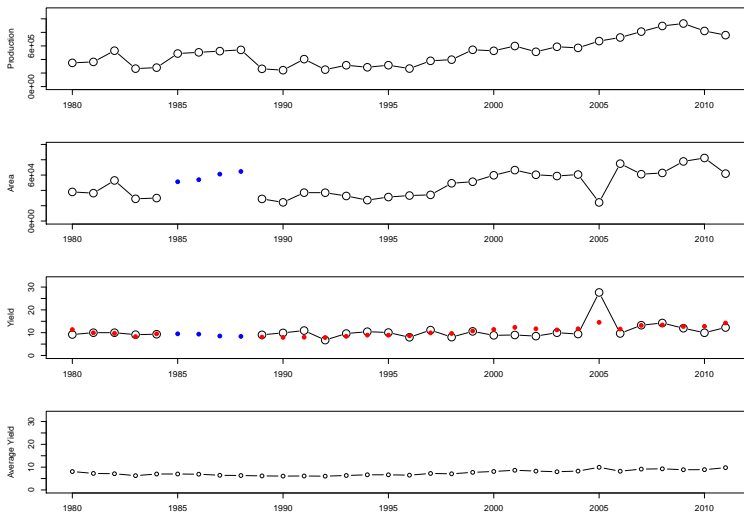
Yield of SweetPotato by sub-region with EM-average

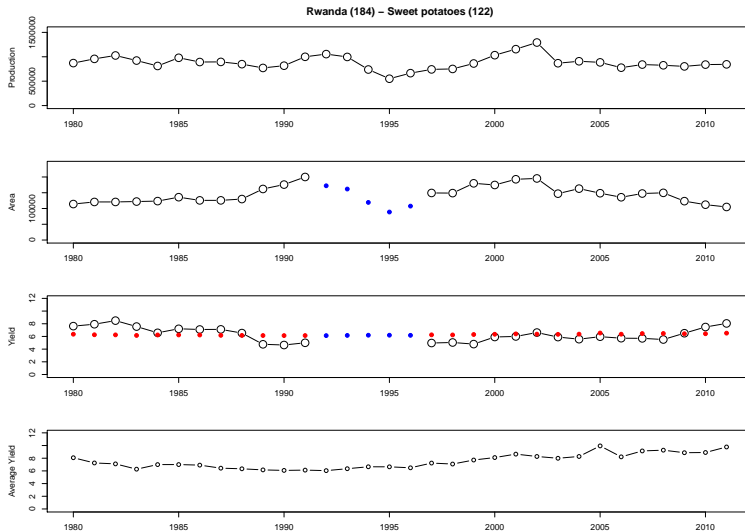


Burundi (29) – Sweet potatoes (122)

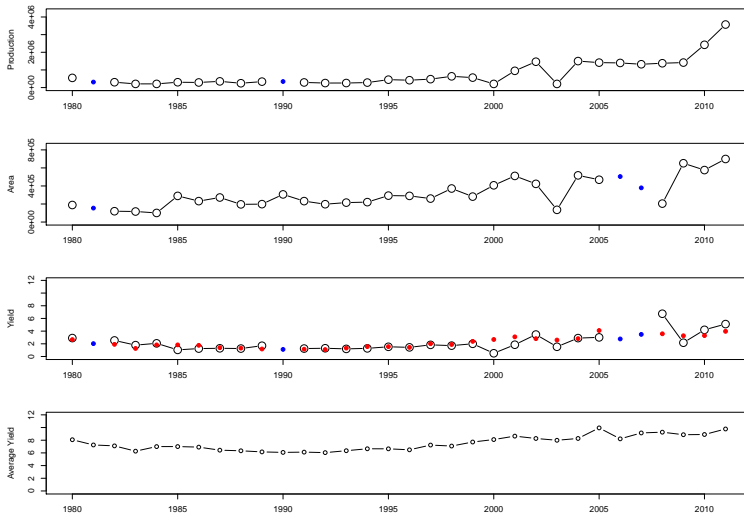


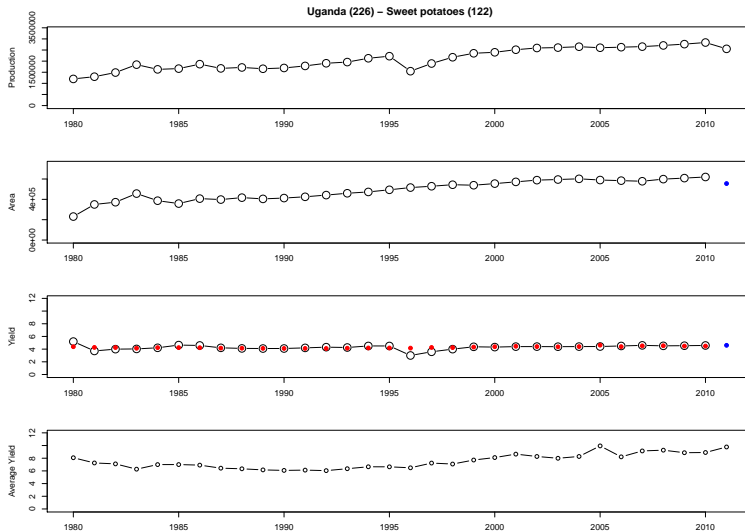
Kenya (114) – Sweet potatoes (122)

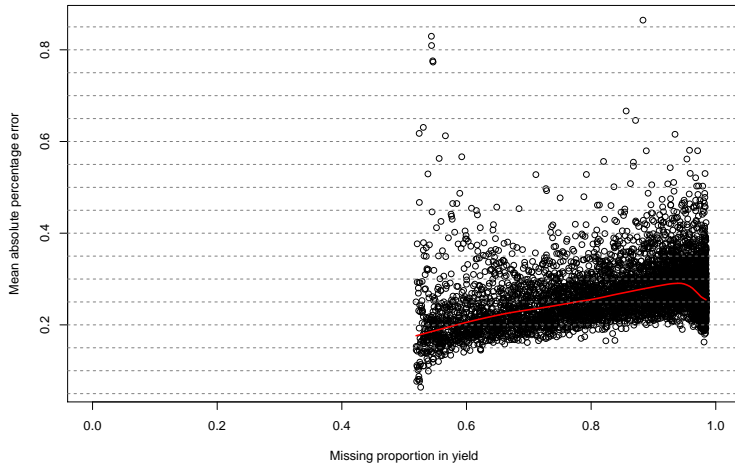




United Republic of Tanzania (215) – Sweet potatoes (122)







- Incorporation additional information for imputation.
- Develop a better grouping classification.

Thank you for your time and attention.