Statistical Working Paper on Imputation and Validation Methodology for the FAOSTAT Production Domain

Michael C. J. Kao
Food and Agriculture Organization
of the United Nations

Abstract

This document is a presentation of the continue on-going improvements and refinement of the imputation methodology of the FAOSTAT production domain. The paper proposes a new imputation method for the FAOSTAT production domain. The proposal provides resolve to many of the shortcomings of the current approach, and offers a flexible framework to incorporate further information to improve performance.

We first provide the necessary background of the prior methodology and address issues faced during the last phase of implementation which motivated the proposal.

Visualization of the data are presented in order to identify the potential driving factors of the production of various commodity and to provide insight to potential relationships which underlies the new model.

The description of the new methodology is provided, with a visual decomposition of the model and accompanying explanation. Finally, conclusion and future work are presented for discussion.

Keywords: Imputation, Mixed model, Production.

1. Introduction

Missing values are commonplace in the agriculturl production domain, stemming from non-responding in surveys or a lack of reliable measurement. Yet a consistent and non-sparse production domain is of critical importance to Food Balance Sheets, thus accurate and reliable imputation is essential and a necessary requisite for continuing work. This paper addresses several shortcomings of the current work and a new methodology is proposed in order to resolve these issues and to increase the accuracy of imputation.

The relationship between the variables in the production domain can be expressed as:

$$\mathrm{P}_t = \mathrm{A}_t \times \mathrm{Y}_t \tag{1}$$

Where P, A and Y represents production, area harvested and yield respectively indexed by time t. The Yield is however, unobserved and can only be calculated when both production and area are available. For certain commodity harvested area may not exist or sometimes they may be represented under a different context.

The main aim for the need of imputation is to incorporate all available and reliable information in order to provide estimates of food supply in Food Balance Sheets (FBS).

2. Background and Review of the Current Methodology

There have been two classes of methodology proposed in the past in order to account for missing values in the production domain. The first type utilises historical information and implements methods such as linear interpolation and trend regression; while the second class of methodology aims to capture the variation of relevant commodity and/or geographic characteristics through the application of aggregated growth rates. The imputation is carried out independently on both area and production, with the yield calculated implicitly as an identity.

Nevertheless, both approaches only utilise one dimension of information and improvements can be obtained if the information usage can be married. Furthermore, these methods lack the ability to incorporate external information such as vegetaion indices, precipitation or temperature that may provide valuable information and enhance the accuracy of mputation.

The simulation results of the prior attempts indicate that linear interpolation is a stable and accurate method but lack the capability to utilize cross sectional informations. Furthermore, it does not provide a solution for extrapolation where connection points are not available. As a result, the aggregation method was then implemented as it provide a high coverage rate for the imputation with satisfactory performance.

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 of the respective series. The formulae of the aggregated growth can be expressed as:

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

Where $\mathbb{S}$ denote the relevant set of products and countries within the relevant commodity group and regional classification after omitting the item to be imputed. For example, to compute the *country cereal aggregated growth* in order to impute wheat production, we sum up all the production of commodities listed in the cereal group in the same country excluding wheat. On the other hand, to impute by *regional item aggregated growth*, wheat production data within the regional profile except the country of interest are aggregated.

Imputation can then be computed as:

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

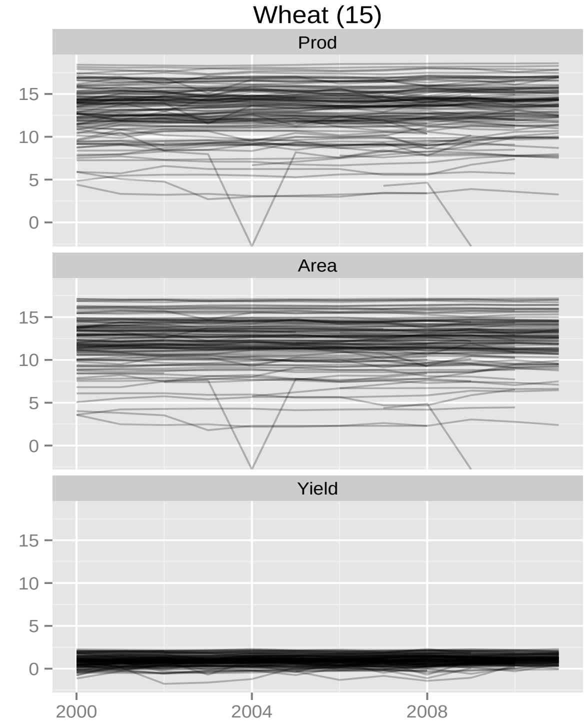
There are however, several shortcoming of this methodology. The achillies heel lies in the fact that area and production are imputed independently, cases of diverging area havrvested and production have been observed that result in inconsistency between trends as well as exploding yield. The source of this undesirable characteristic is nested in the computation of the aggregated growth rate. Due to missing values, the basket computed may not be comparable over time and result in spurious growth or contraction. Furthermore, the basket to compute the changes in production and area may be considerably different.

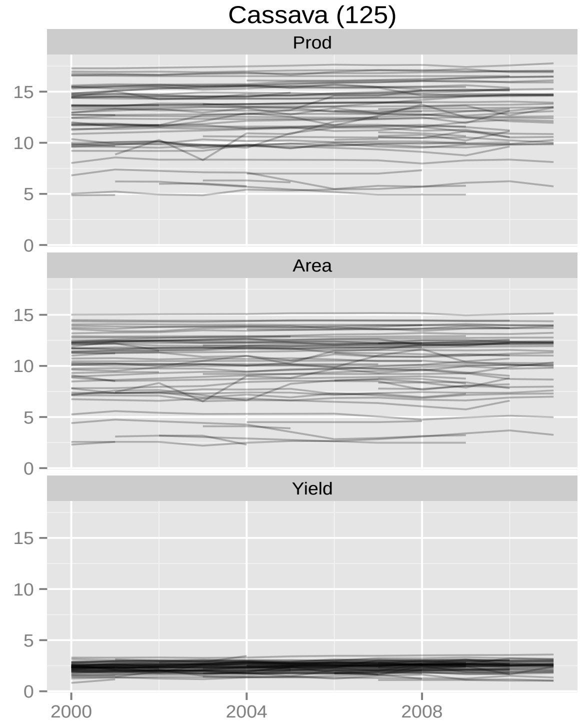
Finally, the methodology does not does not provide insight into the underlying driving forces of the production for understanding and intepretation.

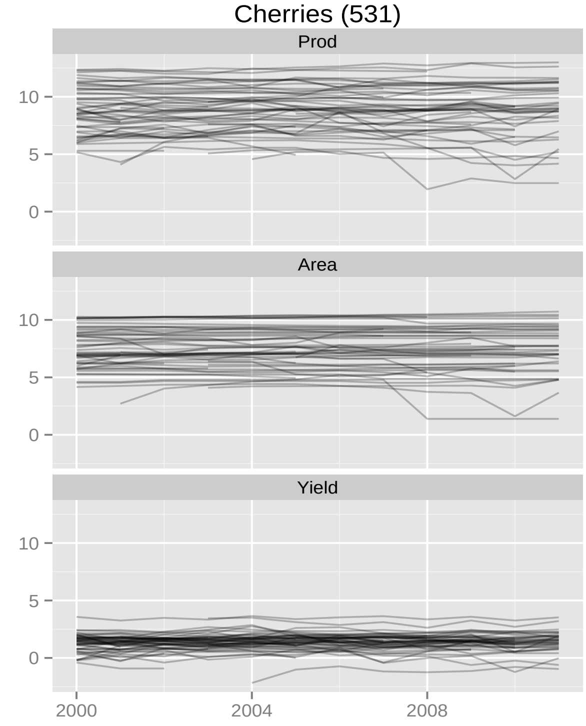
3. Exploratory Data Analysis

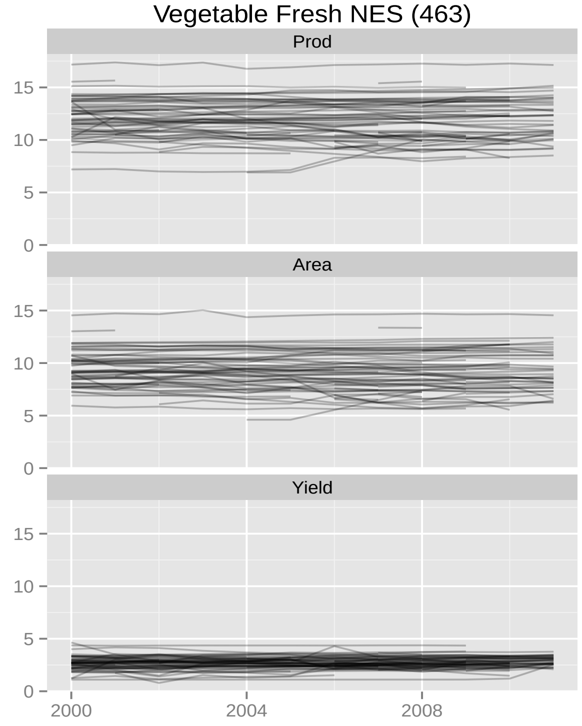
Before any modelling or statistical analysis, a grasp of the data is essential. This section is devoted to some basic exploratory analysis of the data in order to understand the nature of the series and their drivers. First, let us explore the relationship between the identity in equation 1. To make the relationshp clearer we have log-transformed the data so the relationship becomes an additive one rather than multiplicative.

$$\log(P_t) = \log(A_t) + \log(Y_t) \tag{4}$$


Wheat (15)

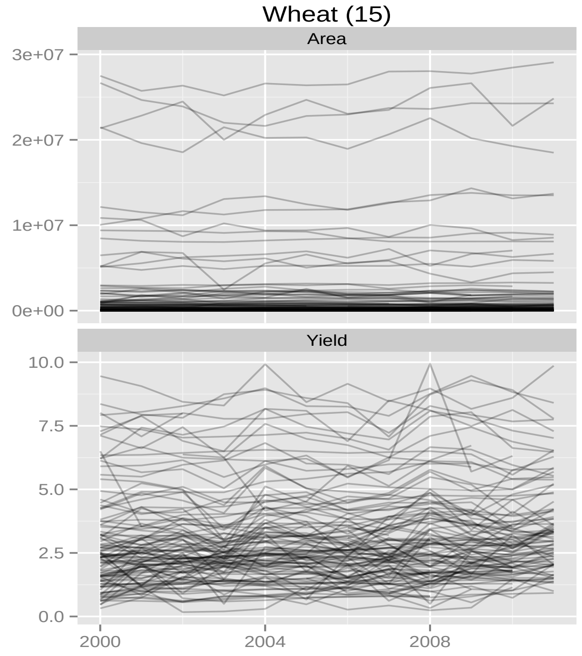

Cassava (125)

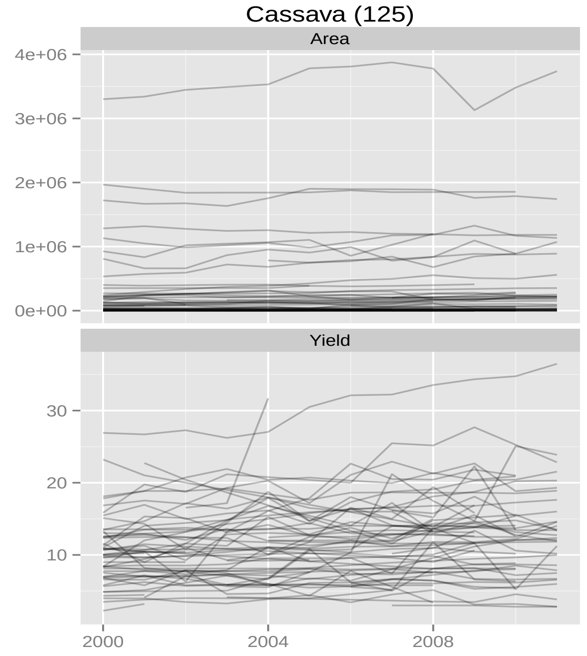

Cherries (531)

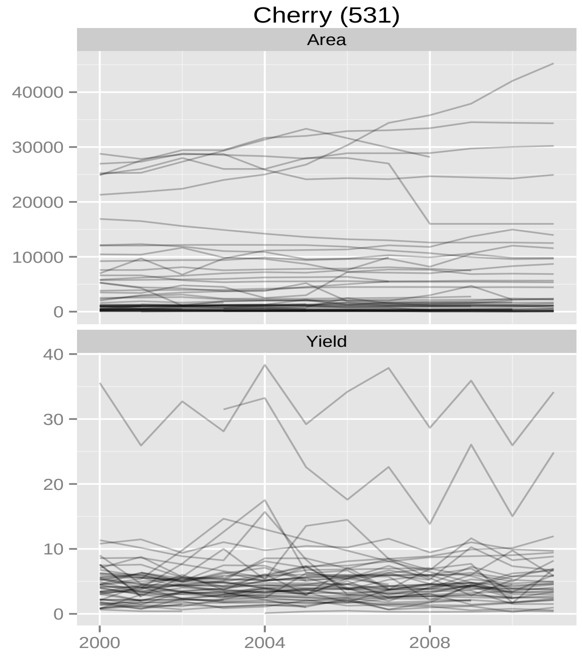

Vegetable Fresh NES (463)

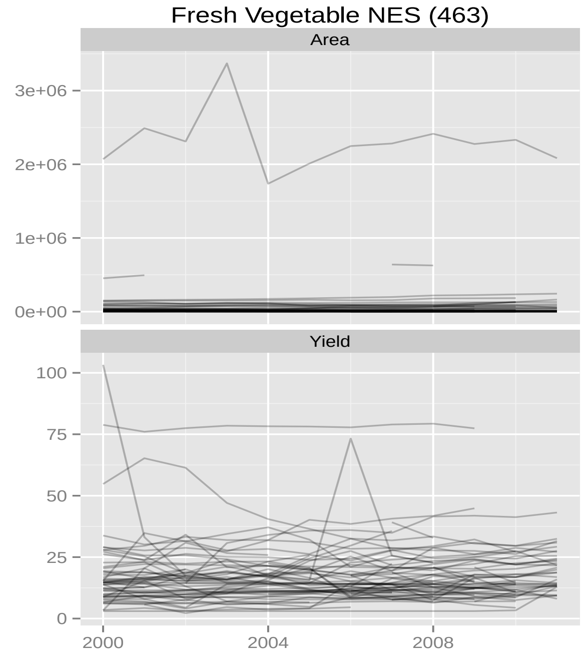
In the above plots, the log of production, area and yield of a specific commodity is depicted within each panel for comparison. Each line represents a country and the production is the sum of area and yield. The first notable thing we can observe from the relationship between the series is that the level of production is mainly determined by the level of harvested area, furthermore shocks are typically refleceted by a significant change in the area rather than the yield. It is commonly believed that harvested area are stable and predictable over time while vulnerable to drastic changes such as wild fire or tsunamies. Secondly, the range of the variability for yield is very small in comparison to area, this matches with our intuition that there are physical constraints on the potential yield of a crop within a given size of area. The results are very similar even between different commodities.

After exploring the relationship between the identity, let us delve deeper into the constitutents of area and yield. Depicted below are area and yields for the same set of commodities but on an original scale.









We can first observe that the area are in general much more stable and smoother when compared to the yield. The yield fluctutate from year to year with correlation to a certain extent and more preminently observed in wheat. This implies there maybe underlying factors such as climatic conditions, which may impact the yield in different countries simultaneously. However, this characteristic is not observed in the NES category which suggests the impact of the factor are stronger within the same commodity but weak in general.

The figures strongly suggests that both the trend and level of the production is largely determined by the area harvested, but the year-to-year fluctuation is driven by the yield, which may be associated with the climate conditions. The exploratory data analysis reveals valuable insight into the nature of the time series, and underlies the proposed model decomposition of variability and in attributing the fluctuation to area and yield.

4. Proposed Methodology

In order to avoid identification problems and to capture the correlation of yield between countries, we propose to impute the yield and area in contrast to production and area. The additional advantage of this approach, with well designed validation, almost guarantees that the series will not diverge as is the case with the current approach.

4.1. Imputation for Yield

The proposed model for imputing the yield is a linear mixed model, the usage of this model enables all the information available both historical and cross-sectional to be incorporated. In addition, proposed indicator such as the vegetaion index, CO_2 concentration and other drivers can be tested and incorporated if proven to improve predictive power.

The general form of the model can be specified as:

$$\begin{aligned} \mathbf{y_i} &= \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \epsilon_i \\ \mathbf{b_i} &\sim \mathbf{N_q}(\mathbf{0}, \boldsymbol{\Psi}) \\ \epsilon_i &\sim \mathbf{N_{ni}}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i) \end{aligned} \tag{5}$$

Where the fixed component $\mathbf{X_i}\boldsymbol{\beta}$ models the regional level and trend, while the random component of $\mathbf{Z_i}\mathbf{b_i}$ captures the country specific variation around the regional level. More specifically the proposing model for FAOSTAT production has the following expression:

$$Y_{i,t} = \overbrace{\beta_{0j} + \beta_{1j}t}^{\text{Fixed effect}} + \overbrace{b_{0,i} + b_{1,i}t + b_{2,i}\bar{Y}_{j,t}}^{\text{Random effect}} + \epsilon_{i,t} \tag{6}$$

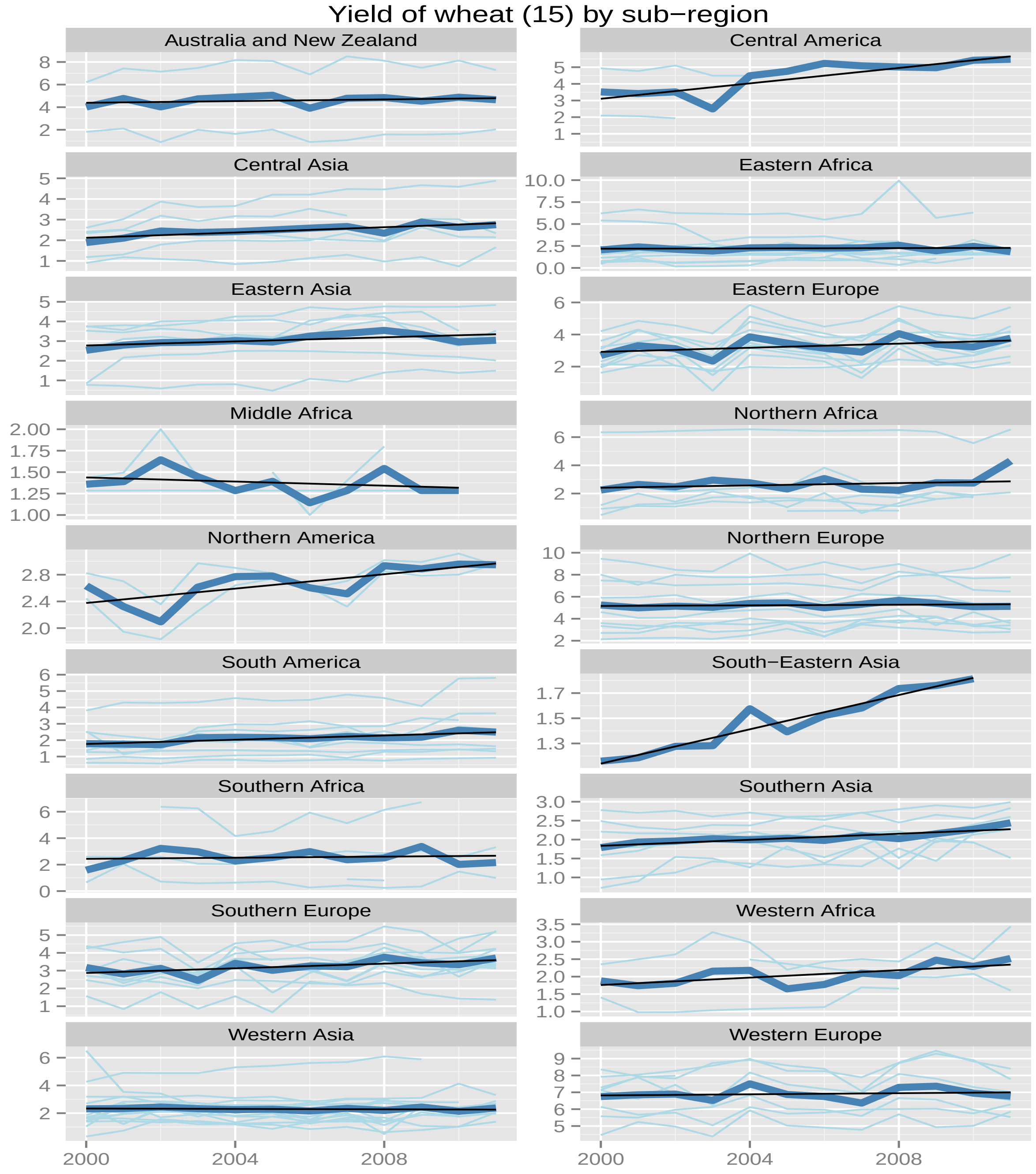
Where Y denotes yield with $\bar{Y}$ being the grouped averaged yield, i for country, j represents the designated regional grouping and t denotes time. Where the grouped averaged yield is computed as:

$$\bar{Y}_{j,t} = \frac{1}{N_i}\sum_{i \in j} \hat{Y}_{i,t} \tag{7}$$

However, since the grouped average yield is only partially observed given the missing values, the average yield is estimated through the EM-algorithm.

In essence, the imputation of the yield is based on the country specific level and historical regional trend while accounting for correlation between country and regional fluctuation. In contrast to the previous methodology, where the full effect of the change is applied, the proposed methodology measures the size of relationship between the individual time series and the regional variability to estimate the random effect for the country. Since both historical and cross-sectional information are utilised, imputed values display stable characteristics while reflecting changes in climatic conditions.

To better understand the methodology, shown below is the sub-regional decomposition with both the regional level (black line) and the grouped yield (dark blue line) superimposed. The model assigns a sub-regional level and trend depicted as the black line to each series and models the correlation with the dark blue grouped yield series.



4.2. Imputation for Harvested area

After imputing the yield and compute areas and production where available, we then impute the area with linear interpolation and carry forward the last observation when both production and area are not available.

Following prior research and current investigation, we believe linear interpolation is suitable because much of the harvested area data exhibits extremely stable trend and linear interpolation yields a satisfactory result. Despite the stability, shocks are sometimes observed in the area series. However without further understanding of the nature and the source of the shock, blindly applying the model will introduce vulnerability rather than an anticipated improvement of imputation. At the current stage, we have chosen to carry forward and backward the latest available figure where linear interpolation is not applicable. The major advantage of this approach is that if the production cease to exist and both production and area are zero, we will not impute a positive value. Nevertheless, we are continuing to explore the data and investigate methods which may be applied to the imputation of area.

$$\hat{A}_t = A_{t_a} + (t - a) \times \frac{A_{t_b} - A_{t_a}}{t_b - t_a} \tag{8}$$

Then for values which we can not impute with linear interpolation, we impute with the latest value.

$$\hat{A}_t = A_{t_{nn}} \tag{9}$$

5. Conclusion and Future improvements

The aim of the paper is to further refine and incorporate feedback from the last round of the imputation methodology.

The proposed model demonstrated the ability to solve issues such as diverging area and production series and biased growth as a result of the missing value. Furthermore, the proposal takes into account of the feed back for incorporating relevant information and establish a flexible framework to accomodate additional information.

This is however an draft interim report, and the teams are continuing to collaborate in order to seek deeper understanding of the data and improve the model. Application of the state-space model are in place to be tested for imputation of the production, area and yield simultaneous.

Annex 1: Geographic and classification

The geographic classification follows the UNSD M49 classification at `http://unstats.un.org/unsd/methods/m49/m49regin.htm`. The definition is also available in the `FAOregionProfile` of the R package **FAOSTAT**.

Annex 2: Code implementation

The codes and data can be find at the github repository at `https://github.com/mkao006/Imputation`.

Algorithm 1: EM-Algorithm for Imputation

```
Initialization;
    $\hat{Y}_{i,t} \leftarrow f(Y_{i,t})$;
    $\mathcal{L}_{\text{old}} = -\infty$;
    $\mathcal{E}$ = 1e-6;
    n.iter = 1000;
begin
    for i=1 to n.iter do
        (1) Compute the expected group average yield;
            $\bar{Y}_{j,t} \leftarrow 1/N \sum_{i \in j} \hat{Y}_i$;
        (2) Fit the Linear Mix Model in 6;
        if $\mathcal{L}_{new} - \mathcal{L}_{old} \geq \mathcal{E}$ then
            $\hat{Y}_{i,t} \leftarrow \hat{\beta}_{0j} + \hat{\beta}_{1j}t + \hat{b}_{0i} + \hat{b}_{1i}t + \hat{b}_{2j}\bar{Y}_{j,t}$;
            $\mathcal{L}_{\text{old}} \leftarrow \mathcal{L}_{\text{new}}$;
        end
        else
            break
        end
    end
end
```

```
Algorithm 2: Imputation Process
Data: Production (element code = 51) and Harvested area (element code = 31) data
Result: Imputation
Missing values are denoted $\emptyset$;
Initialization;
begin
    if $A_t = 0 \wedge P_t \neq 0$ then
        $A_t \leftarrow \emptyset$;
    end
    if $P_t = 0 \wedge A_t \neq 0$ then
        $P_t \leftarrow \emptyset$;
    end
end
Start imputation;
begin
    forall the commodities do
        (1) Compute the implied yield;
                $Y_{i,t} \leftarrow P_{i,t}/A_{i,t}$;
        (2) Impute the missing yield with the imputation algorithm 1;
        forall the imputed yield $\hat{Y}_{i,t}$ do
            if $A_t = \emptyset \wedge P_t \neq \emptyset$ then
                $\hat{A}_{i,t} \leftarrow P_{i,t}/\hat{Y}_{i,t}$;
            end
            if $P_t = \emptyset \wedge A_t \neq \emptyset$ then
                $\hat{P}_{i,t} \leftarrow A_{i,t} \times \hat{Y}_{i,t}$;
            end
        end
        (4) Impute area ($A_{i,t}$) with equation 8 then 9;
        forall the imputed area $\hat{A}_{i,t}$ do
            if $\hat{Y}_{i,t} \neq \emptyset$ then
                $\hat{P}_{i,t} \leftarrow \hat{A}_{i,t} \times \hat{Y}_{i,t}$;
            end
        end
    end
end
```

Affiliation:

Michael. C. J. Kao
Economics and Social Statistics Division
Economic and Social Development Department
United Nations Food and Agriculture Organization
Viale delle Terme di Caracalla 00153 Rome, Italy
E-mail: michael.kao@fao.org
URL: https://github.com/mkao006/Imputation