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 paper proposes a new imputation method for the FAOSTAT production domain based on linear mixed model and the EM-algorithm. The proposal provides resolve to many of the shortcomings of the current approach, and offers a flexible and robust framework to incorporate further information to improve performance.

We first examine the factors that drive changes in production by commodity, after which a brief acount of the current approach and its shortcomings. A description of the new methodology is provided, with a visual decomposition of the model and accompanying explanation.

Finally, a case study on wheat is given with the fit, diagnostic and simulation results presented and closed with discussion.

Keywords: Imputation, Linear Mixed Model, Agricultural Production, EM.

Disclaimer

This Working Paper should not be reported as representing the views of the FAO. The views expressed in this Working Paper are those of the author and do not necessarily represent those of the FAO or FAO policy. Working Papers describe research in progress by the author and are published to elicit comments and to further discussion.

1. Introduction

Missing values are commonplace in the agricultural production domain, stemming from non-response in surveys or a lack of capacity by the reporting entity to provide measurement. Yet a consistent and non-sparse production domain is of critical importance to Food Balance Sheets (FBS), 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 represent production, area harvested and yield of crops, respectively, indexed by time t. In the case of livestock, A represents number of slaughtered animal while Y represents the carcass weight per animal. The yield is, however, unobserved and can only be calculated when both production and area are available. For certain commodities, harvested area may not exist or sometimes it may be represented under a different context.

The primary objective of imputation is to incorporate all available and reliable information in order to provide best estimates of food supply in 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 utilizes historical information and implements methods such as linear interpolation and trend regression; while the second class aims to capture the variation of relevant commodity and/or spatial 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 utilize one dimension of information and improvements can be obtained if information usage can be married. Furthermore, these methods lack the ability to incorporate external information such as vegetation indices, precipitation or temperature that may provide valuable information and enhance the accuracy of imputation.

Simulation results of the prior attempts indicate that linear interpolation over small period is a stable and accurate method but it lacks the capability to utilize cross-sectional information. 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 was found to provide a high coverage rate for imputation with seemingly 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 formula 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 $\mathbb{S}$ denotes 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* with the aim 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 shortcomings of this methodology. The Achilles heel lies in the fact that area and production are imputed independently, cases of diverging area harvested and production have been observed that result in inconsistency between trends as well as exploding yields. The source of this undesirable characteristic is nested in the computation of the aggregated growth rate. Owing to missing values, the basket computed may not be comparable over time and consequently results in spurious growth or contraction. Furthermore, the basket to compute the changes in production and area may be considerably different.

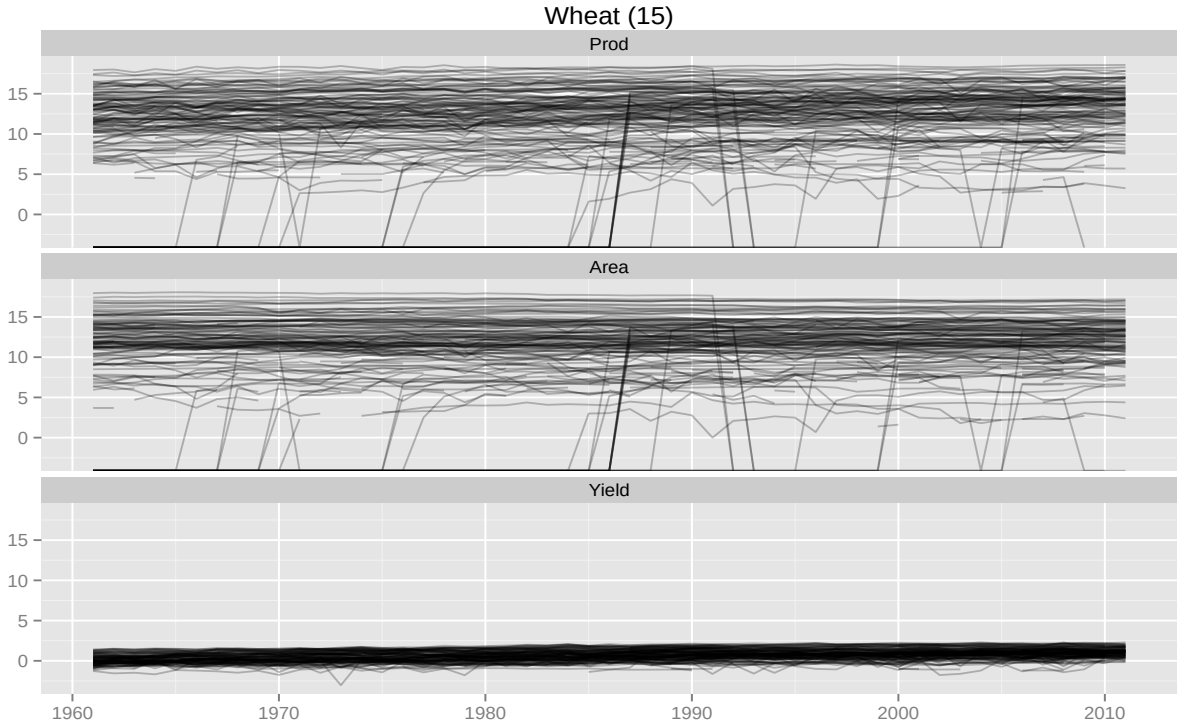
Finally, the methodology does not provide insight into the underlying driving factors of production that are required to better understand the phenomenon and hence for interpretation.

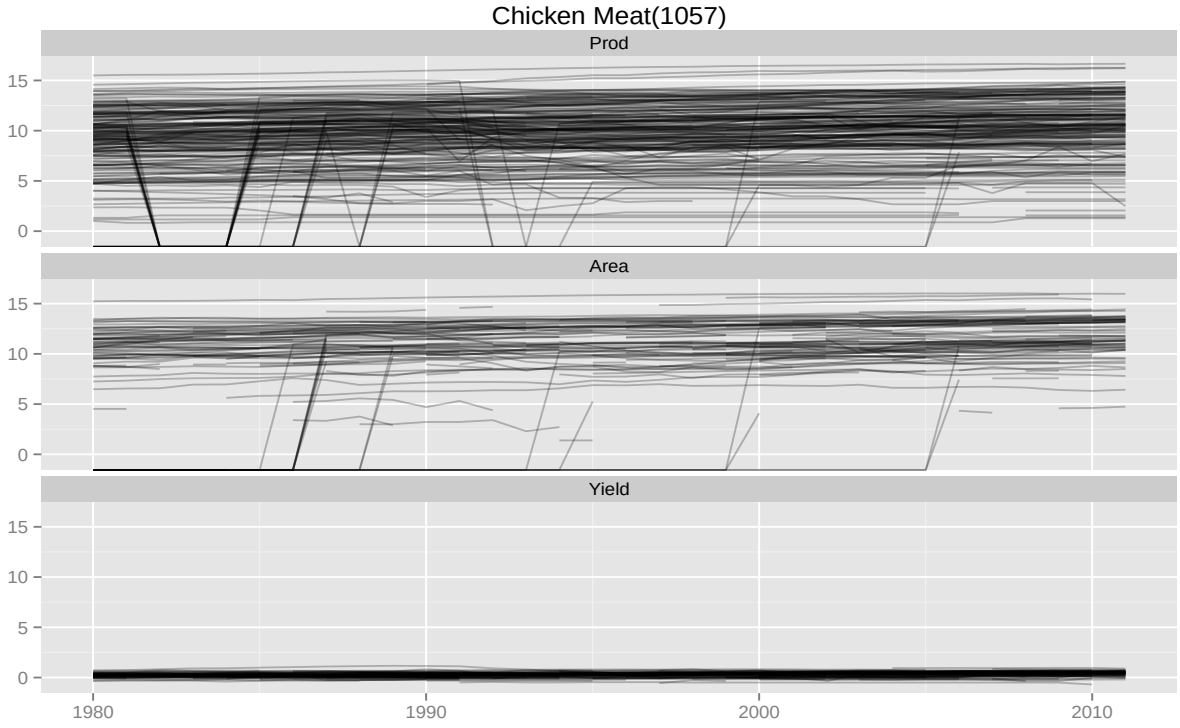
3. Exploratory Data Analysis

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 relationship clearer we have log-transformed the data so the relationship becomes an additive one rather than multiplicative.

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

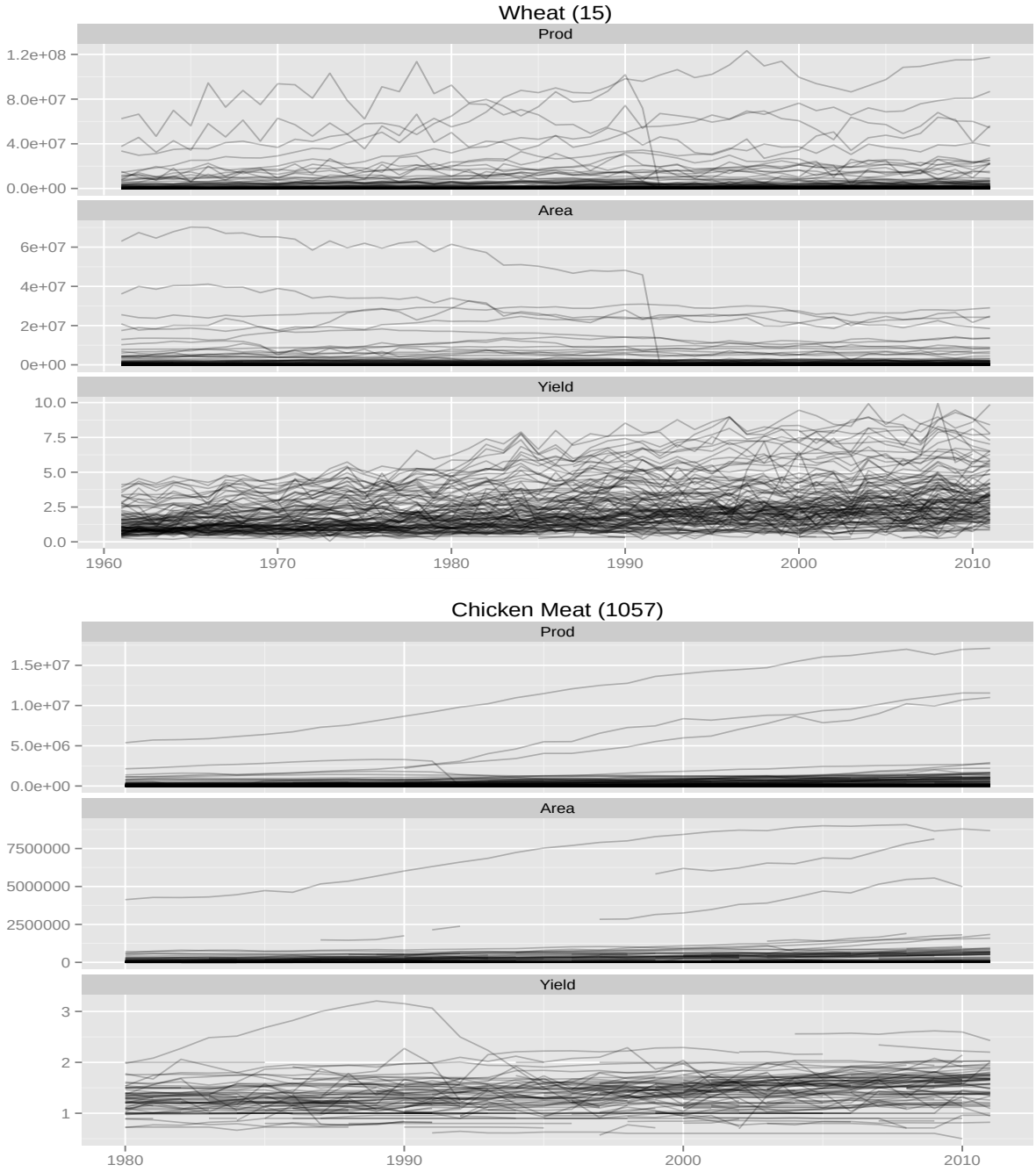




In the above plots, the log of production, area and yield of a specific commodity is depicted within each panel for comparison. Each line represents a country and the production is the

sum of area and yield. The first noteworthy feature we 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 reflected by a significant change in the area rather than the yield. It is commonly believed that harvested area is stable and predictable over time while vulnerable to shocks stemming from the natural world. Secondly, the range of the variability for yield is very small in comparison to area, consistent 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 constituents of production: 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 area is in general much more stable and smoother when compared to yield. The yield fluctuates from year-to-year with correlation to a certain extent, which

is more prominently observed in wheat. This implies that there maybe underlying factors such as climatic shocks, which may impact the yield in different countries simultaneously. However, this characteristic is not observed in the livestock and there are no reason to suspect that the carcass weight per animal vary significantly from year-to-year and correlated amongst countries.

The figures strongly suggests that both the trend and level of production is largely determined by area harvested, but the year-to-year fluctuation is driven by the yield, which may be associated with changing climate conditions. The exploratory data analysis reveals valuable insights into the nature of the time series, and underpins 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 added 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 Harvested area

From the exploratory analysis we can observe that the series is extremely stable, and it is impossible to predict the shocks given the current information set. The methodology proposed here is what we called naive imputation with linear interpolation and last observation carry forward or backward. This method has proven to work extremely well especially when support points are added after imputing the yield.

After imputing the yield and computing area 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 exhibit extremely stable trends while linear interpolation yields a satisfactory result. Despite the stability, shocks are sometimes observed in the area series. However, without a further understanding of the nature and the source of the shocks, blindly applying the model will introduce vulnerability rather than an anticipated improvement of imputation performance. At the current stage, we have chosen to carry forward and backward the latest available data where linear interpolation is not applicable. The major advantage of this approach is that if production ceases 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 superior methods which may be applied to the imputation of area.

$$\hat{A}_t = A_{t_a} + (t-a) \times \frac{A_{t_b} - A_{t_a}}{t_b - t_a} \tag{5}$$

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

$$\hat{A}_t = A_{t_{nn}} \tag{6}$$

4.2. Imputation for Yield

The proposed model for imputing the yield is a linear mixed model, the usage of this model enables all information available both historical and cross-sectional to be incorporated. In ad-

dition, proposed indicators such as the vegetation 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 expressed as:

$$\begin{aligned} \mathbf{y_i} &= \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \epsilon_i \\ \mathbf{b_i} &\sim \mathbf{N_q}(\mathbf{0}, \boldsymbol{\Psi}) \\ \epsilon_i &\sim \mathbf{N_{ni}}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda_i}) \end{aligned} \tag{7}$$

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

$$Y_{i,t} = \overbrace{\mathbf{X_i}\boldsymbol{\beta}}^{\text{Fixed effect}} + \overbrace{b_{0,i} + b_{1,i}t + b_{2,i}\bar{Y}_{j,t}}^{\text{Random effect}} + \epsilon_{i,t} \tag{8}$$

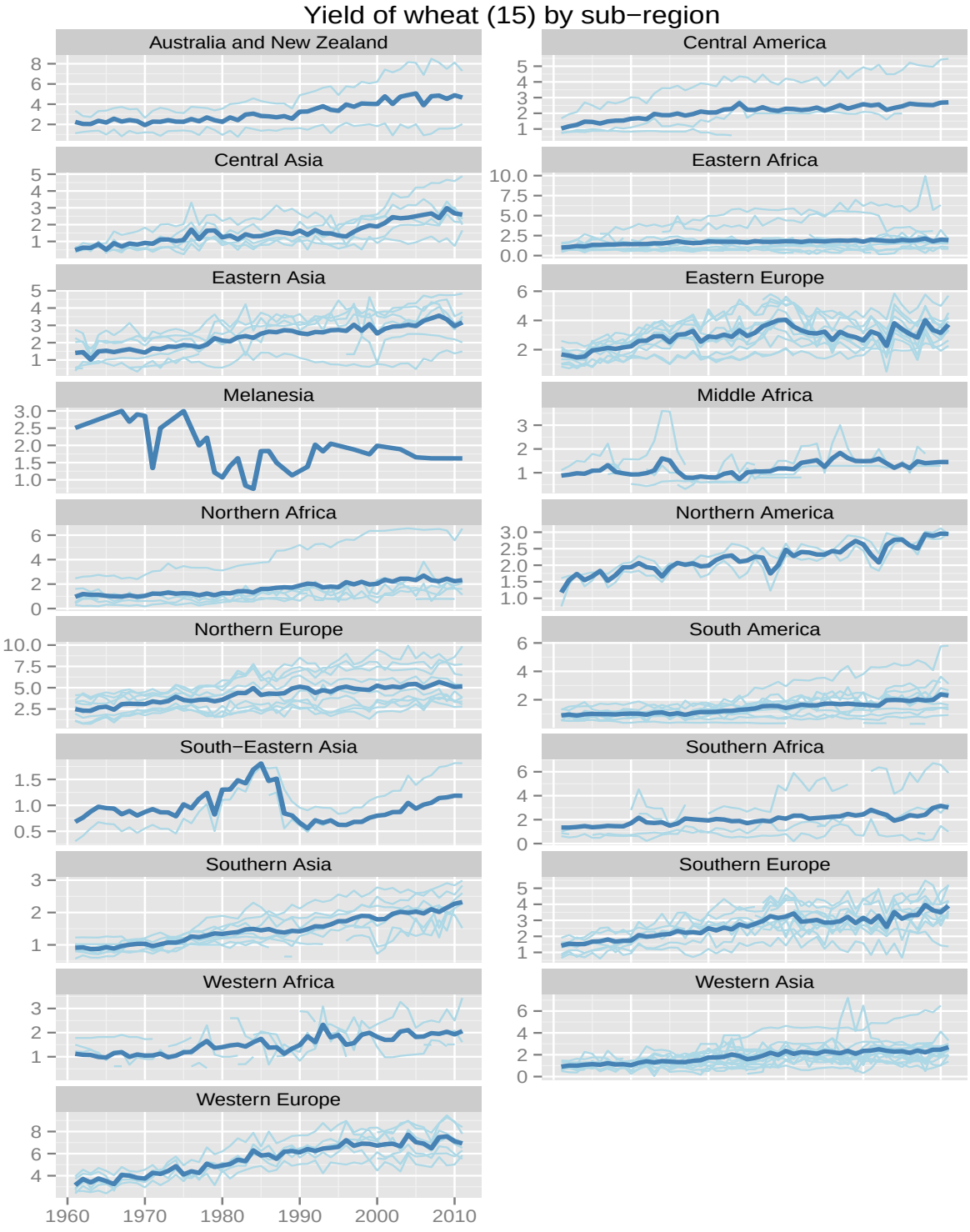
Where Y denotes yield with $\bar{Y}$ being the grouped averaged yield, i for country, j represents the designated regional grouping and t denotes time. The fixed effect is left for external drivers such as precipitation, the grouped averaged yield is computed as:

$$\bar{Y}_{j,t} = \frac{1}{N_i}\sum_{i \in j} \mathbb{1}_{Y^O_{i,t}} Y_{i,t} + \mathbb{1}_{Y^M_{i,t}} \hat{Y}_{i,t} \tag{9}$$

However, as 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 fluctuations. 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 utilized, imputed values display stable characteristics while reflecting changes in climatic conditions.

To better understand the methodology, shown below is the sub-regional decomposition with the grouped average yield (dark blue line) superimposed. The model estimates the regional average while accounting for country specific deviation and effects.



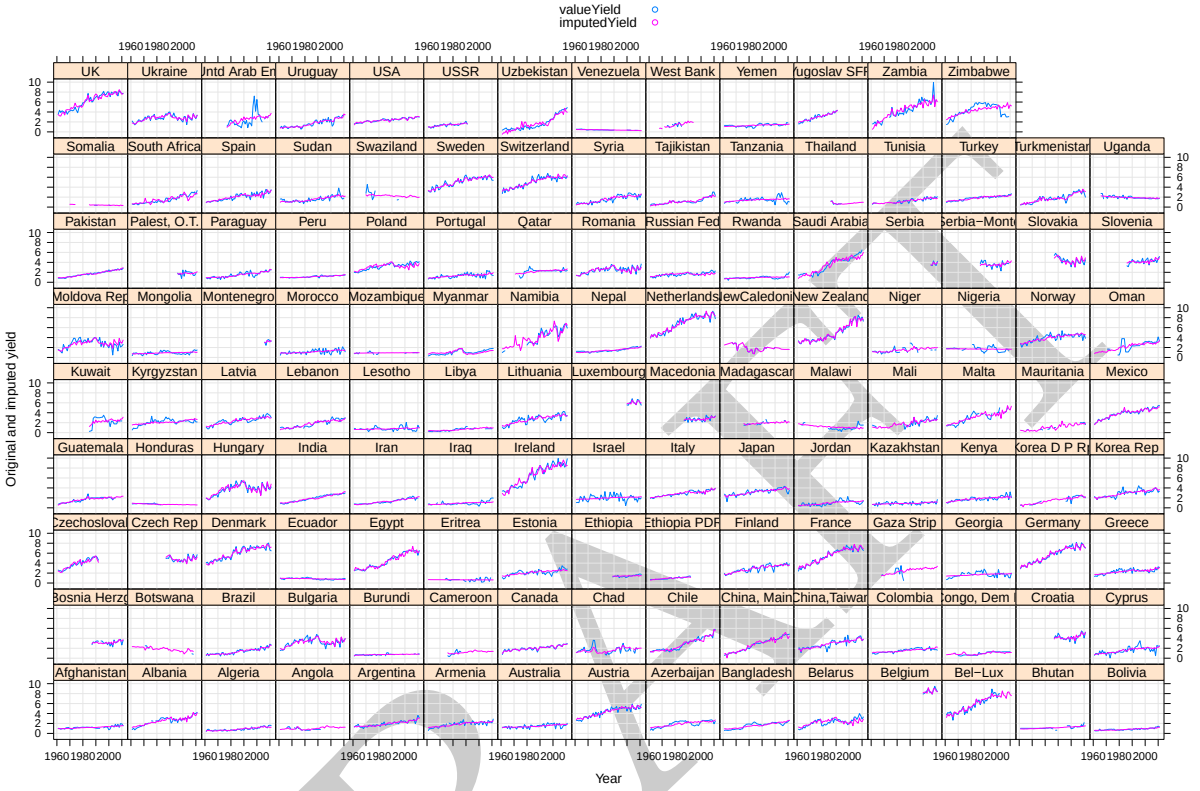
5. Wheat Case Study

In this section we present the imputation result and diagnostic of the proposed methodology.

5.1. Model fit

Shown below is the model fitted to the wheat data set, the pink are the fitted value of the model and act as the imputation value where the original data is missing. Overall, we can

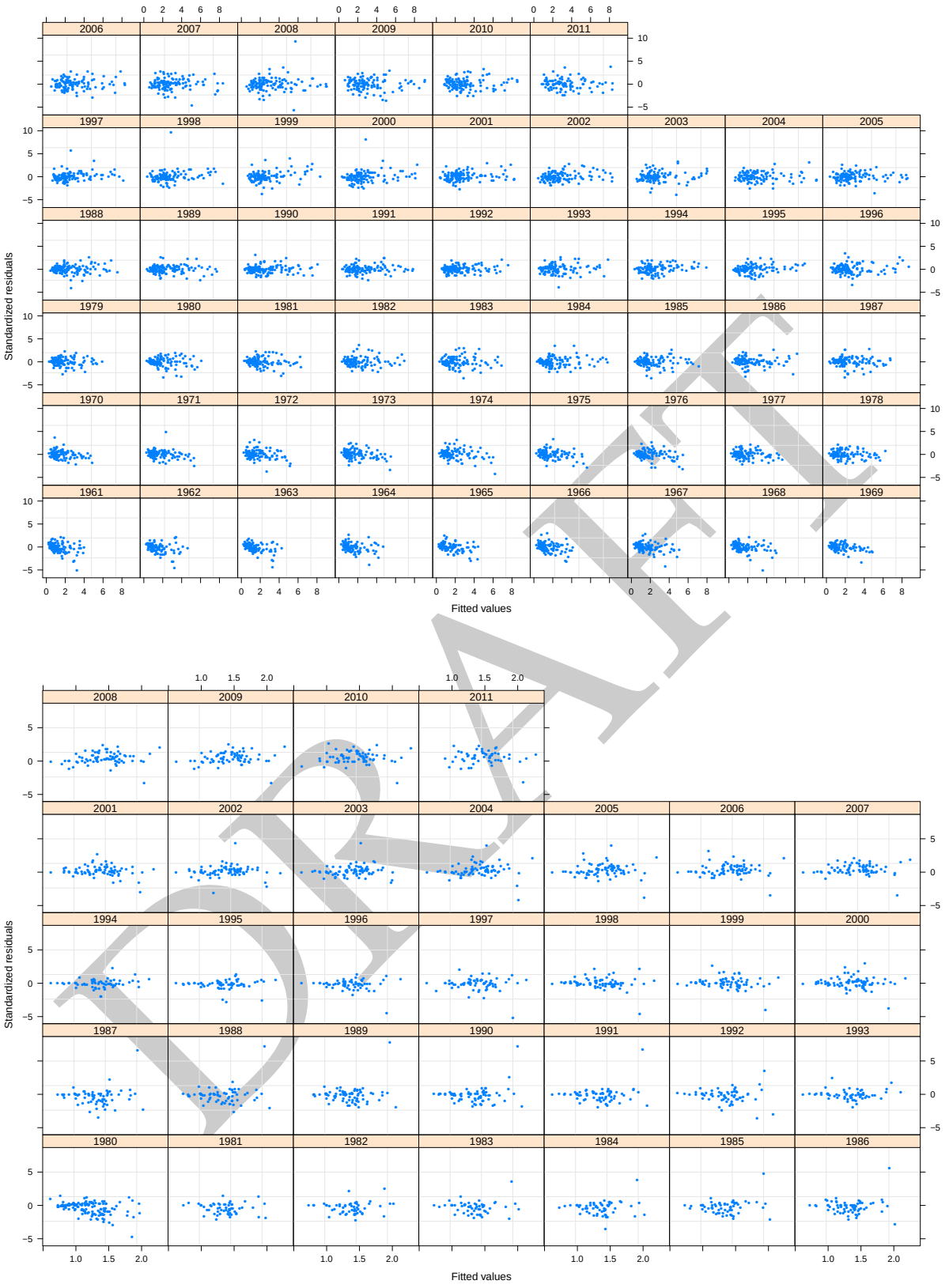
see the fit is fine but there are several fit which shows sign of unsatisfactory. In particularly, Zimbabwe, United Arab Emirates, Oman and Kuwait. They all share the same characteristic of drastic change over a point in time or a certain period in which in-depth analysis is required for understanding.



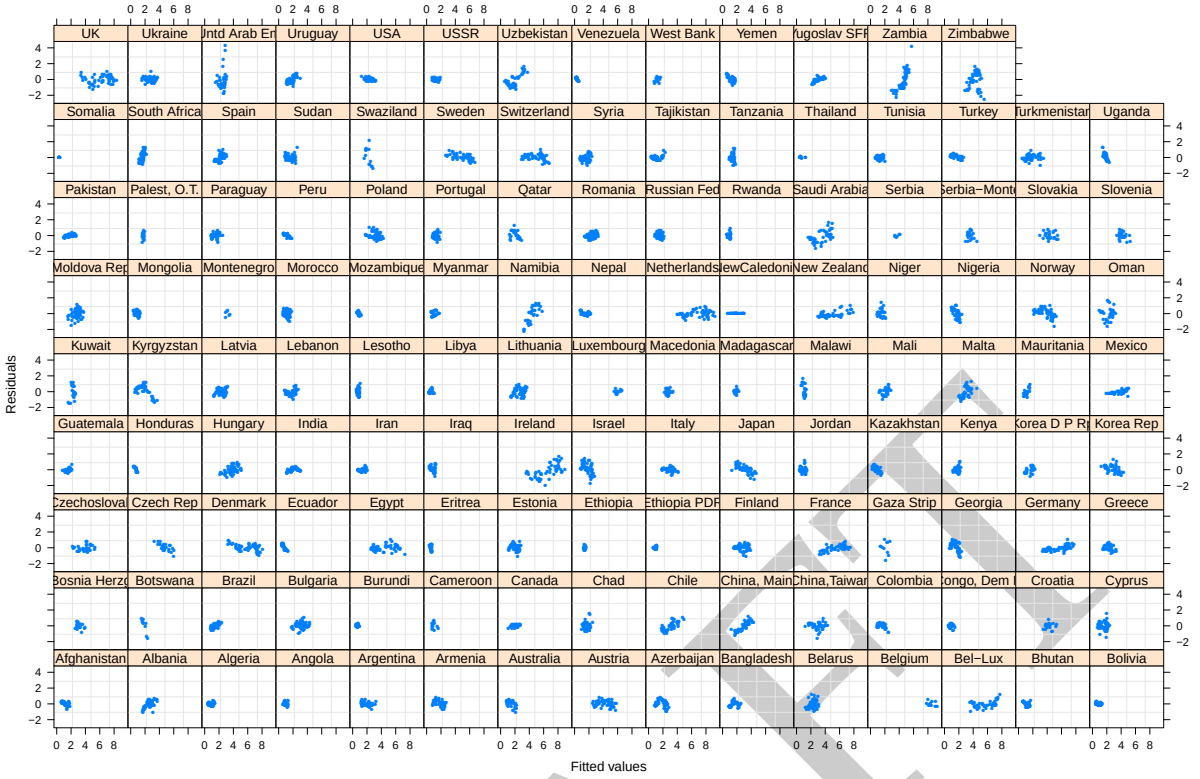
5.2. Diagnostic

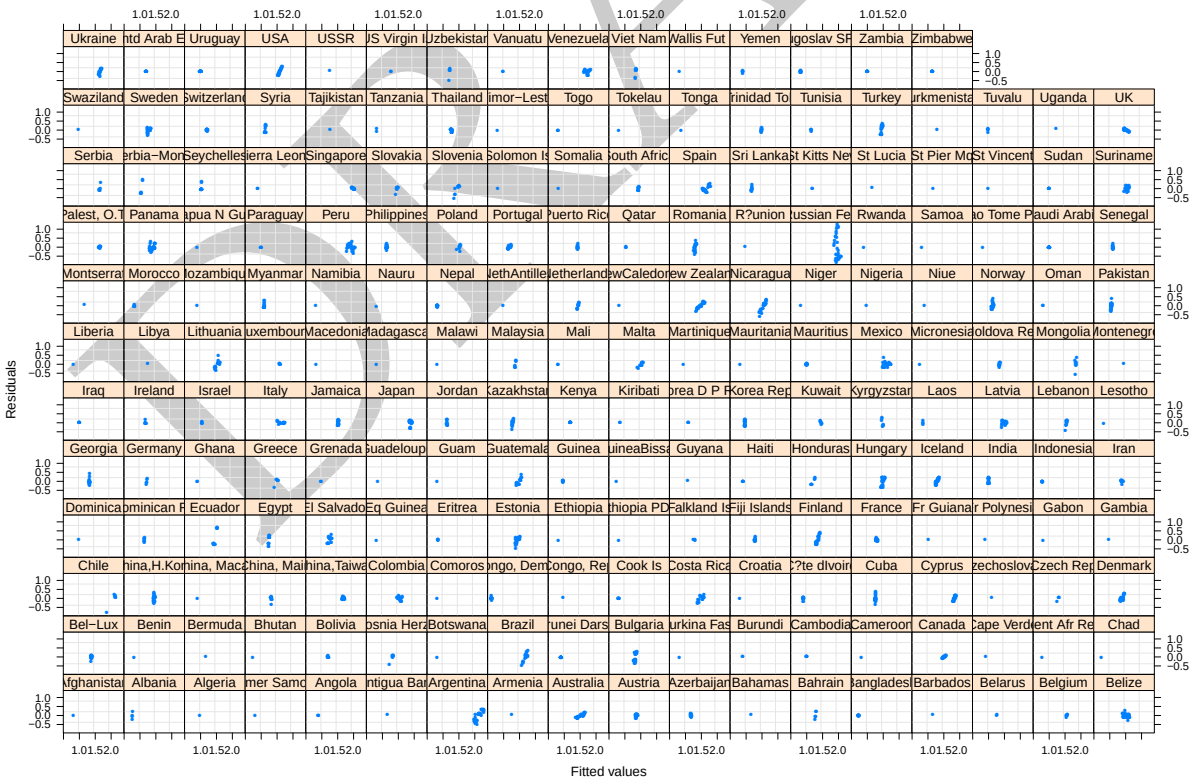
In this section we provide diagnostic plots which will assist us in assess the validity and feasibility of the model for imputation.

We first examine the linearity assumption with respect to time. The following two plots ilustrates the standardized residual against the fitted value, we can observe that both plot does display any peculiar pattern or information which we are able to utilize to enhance the model.

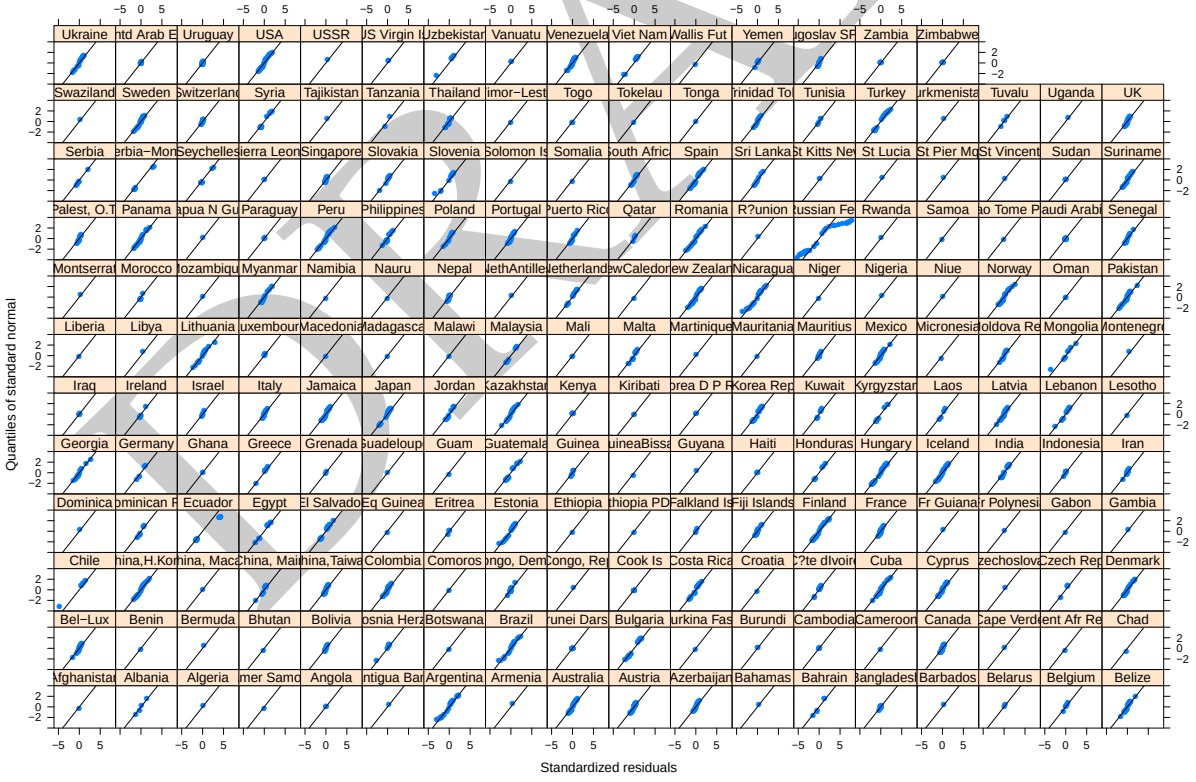


Secondly, the same plot is generated for at the country level. In some countries like Zambia, it seems that there is a positive associate with the residual and the fitted value. After examination, we found that this is caused by an influencial point where the linear fit was pulled towards. However, accounting for such influencial points on a case-to-case bassis maybe difficult and inpractical.





Finally, the qqnorm plot is depicted and remarkbly that most of the residuals does not depart from the normality assumption.



5.3. Simulation Results

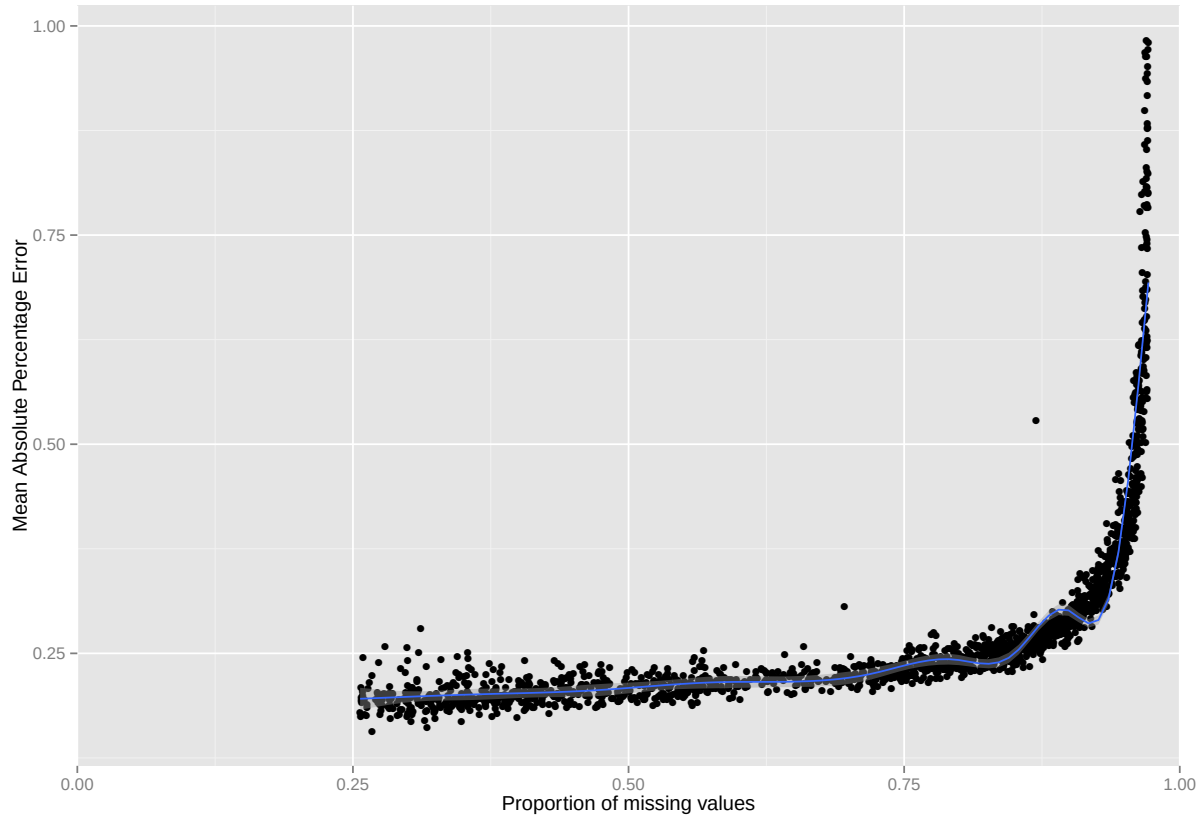
In order to assess the model, we have also carried out a simulation by which observation were withdrawn, imputed then benchmark with the original value.

The wheat data set has been taken for the simulation study, the error rate will and very likely

to differ between commodities. Two thousand simulation were performed, the result is given in the following figure.

Given that approximately 25% of the original production were missing we were only able to carry out the simulation for missing proportion greater than 25%.

The evaluation criteria used was the mean absolute percentage error (MAPE), the figure shows that even with 50% data we can still achieve an out of sample error rate below 25%. In practical modelling this is more than aceptable in particularly that the quality of the data contributes significantly to these error.



6. Conclusion and Future improvements

The aim of the paper has been to overhaul the current imputation methodology, with a more consistent, yet better performing approach.

The proposed model demonstrates the ability to resolve issues such as diverging area and production series and biased growth as a result of missing values. Furthermore, the proposal takes into account the attribute of incorporating relevant information as well as establishing a flexible framework to accommodate additional information.

This is, however, work in progress, as the technical teams are continuing to collaborate in order to seek a deeper understanding of the data in order to improve on the model. Application of the state-space model might be a candidate to succeed in this regard, as it allows for production, area and yield to be imputed simultaneously. In addition, additive mixed model are under-investigation for country non-linearity departures.

7. Acknowledgement

This work is supervised by Adam Prakash with assistance from Nicolas Sakoff, Onno Hoffmeis-

ter and Hansdeep Khaira whom were crucial in the development of the methodology. The author would also like to thank the team members which participated in the first round of the discussion providing valuable feedbacks. Finally, credits to Cecile Fanton and Frank Cachia whom devoted their time to translate the paper into French.

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: Pseudo Codes

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;
(1) Estimate model without grouped average effect;
$\hat{Y}_{i,t} \leftarrow \hat{\beta}_{0,j} + \hat{\beta}_{1,j}t + \hat{b}_{0i} + \hat{b}_{1i}t$;
(2) Estimate model with grouped average effect;
begin
  for *i=1* **to** *n.iter* **do**
    E-step: Compute the expected group average yield;
    $\bar{Y}_{j,t} \leftarrow 1/N \sum_{i \in j} \hat{Y}_i$;
    M-step: Estimate the Linear Mix Model in 8;
    if $\mathcal{L}_{new} - \mathcal{L}_{old} \geq \mathcal{E}$ **then**
      $\hat{Y}_{i,t} \leftarrow$ fitted value of the model;
      $\mathcal{L}_{\text{old}} \leftarrow \mathcal{L}_{\text{new}}$;
    end
    else
      break
    end
  end
end

Algorithm 2: Imputation Procedure

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 5 then 6;
    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

Annex 3: Supplementary Resources

The data, code implementation and documentation can all be found and downloaded from https://github.com/mkao006/sws_imputation. This paper is generated on October 15, 2013and is subject to changes and updates.

References

[1] Douglas M. Bates *lme4: Mixed-effects modelling with R* 2010

[2] Data Collection, Workflows and Methodology (DCWM) team, *Imputation and Validation*

Methodologies for the FAOSTAT Production Domain. Economics and Social Statistics Division, 2011.

[3] Nan M. Laird, James H. Ware, *Random-Effects Models for Longitudinal Data*. Biometrics Volume 38, 963-974, 1982.

[4] R Core Team, *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2013.

[5] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar and the R Development Core Team, *nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-108. 2013

[6] Douglas Bates, Martin Maechler, Ben Bolker and Steven Walker (2013). *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.0-4. http://CRAN.R-project.org/package=lme4

[7] Donald B. Rubin, *Inference and Missing Data*, Biometrika, Volume 63, Issue 3, 581-592, 1976

[8] Valentin Todorov, Matthias Templ *R in the Statistical Office: Part II* 2012

[9] Nam M. Laird, James H. Ware *Random-Effects Models for Longitudinal Data* Biometrics, Volume 38, Number 4, pp.963-974 1982

[10] A. P. Dempster, Nam M. Laird, D. B. Rubin *Maximum Likelihood from Incomplete Data via the EM Algorithm* Journal of Royal Statistical Society. Series B (Methodological), Volume 39, Number 1, pp1-38 1977

[11] Randy C. S. Lai, Hsin-Cheng Huang, Thomase C. M. Lee *Fixed and random effects selection in nonparametric additive mixed models* Electronic Journal of Statistics, Volume 6, pp810-842 2012

Affiliation:

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