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.

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.

It is in the view of the author that imputation should be implemented as a last resort, rather as a replacement for data collection. Imputation itself does not create information it merely create observations based on assumption.

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:

$$\mathrm{P}_t = \mathrm{A}_t \times \mathrm{Y}_t \qquad P_t \geq 0,\, A_t \geq 0,\, Y_t > 0 \tag{1}$$

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.

Presented in table 1 is a description of the existing flags in the current Statistical Working System (SWS). In this exercise, non-official/semi-official data which are marked as either **F**, **E** and **T** are the target values to be imputed.

Table 1: Description of the flags in the Statistical Working System

Flags	Description
	Official data reported on FAO Questionnaires from countries
/	Official data reported on FAO Questionnaires from countries
*	Commodity International Organizations
X	Commodity International Organizations
P	Estimated data using trading partners database
F	FAO estimate
C	Calculated data
B	Data obtained as balance
T	Extrapolated/interpolated
M	Not reported by country
E	Expert sources from FAO (including other divisions)

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

3.1. Relationship Between Production, Area, and Yield

In the this subsection we have log transformed the variables in 1 to make the visualization comparable and to allow the decomposition simpler with an additive relationship.

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

Shown below is the scatter plot matrix of the three variables of interest, production, area harvested and implied yield of a certain year (2011) of all countries producing wheat.

There is a strong relationship between area and production, but much weaker with yield. High production are associated with high value of harvested area and vice versa, which is intuitive as to produce a large quantity you need to harvest a large area. While one would expect producers to produce large amount only at high yield, this may hold at the macro-aggregated level.

From figure 2, 3 and 4 we can observe important time series properties which are important for our methodology. First, the production time series is reasonably smooth but with structural breaks and non-linearity which the area time series closely resemble. On the other hand, the yield series is much more volatile and vary from year-to-year but exihibiting a rather monotonic trend.

These figures suggests the behaviour of the yield time series is much more predictable since we know the year-to-year fluctuation is driven by climate condition and yield increases monotonically as technology improves. In contrast, breaks and non-linearity are much more difficult to model, and we also lack the information on the increasing/decreasing trend or discontinuity of production. This is the main motivation why we decided to model and impute the yield followed by the balancing of area harvested and production.

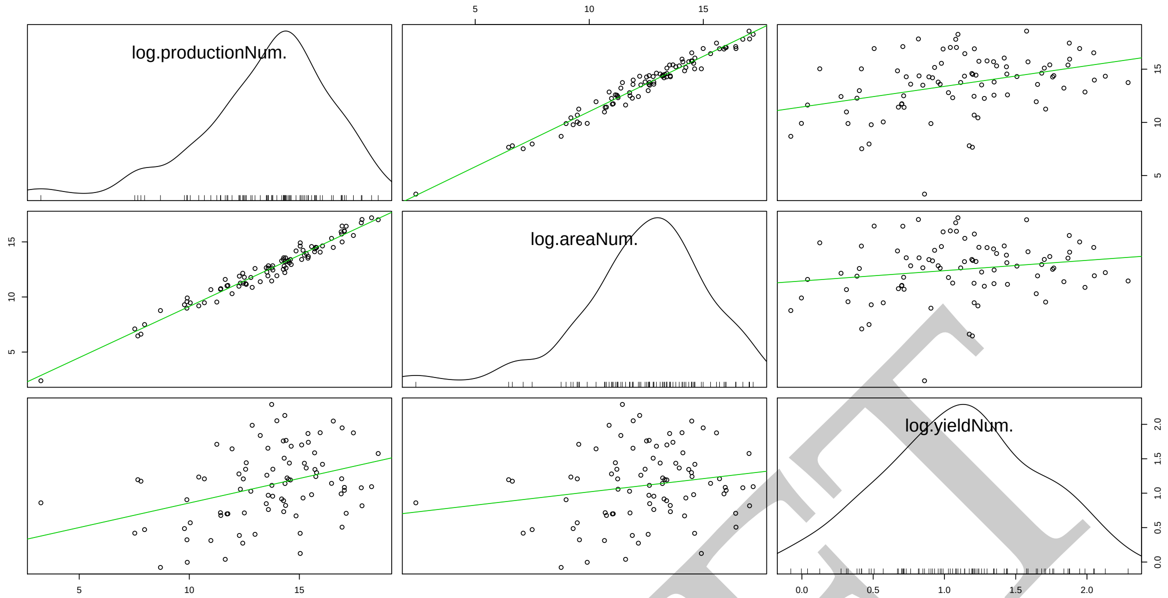


Figure 1: Relation between production, area harvested and yield

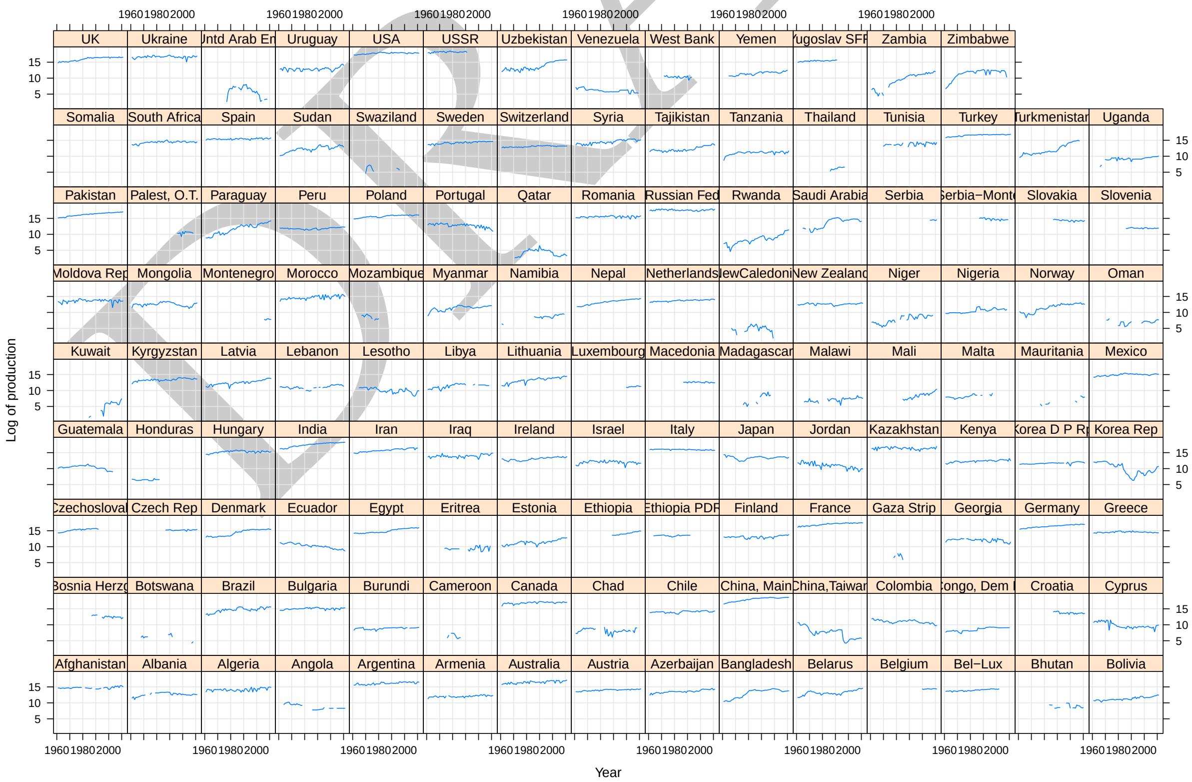


Figure 2: Logged production of wheat from 1961 to 2011 by country

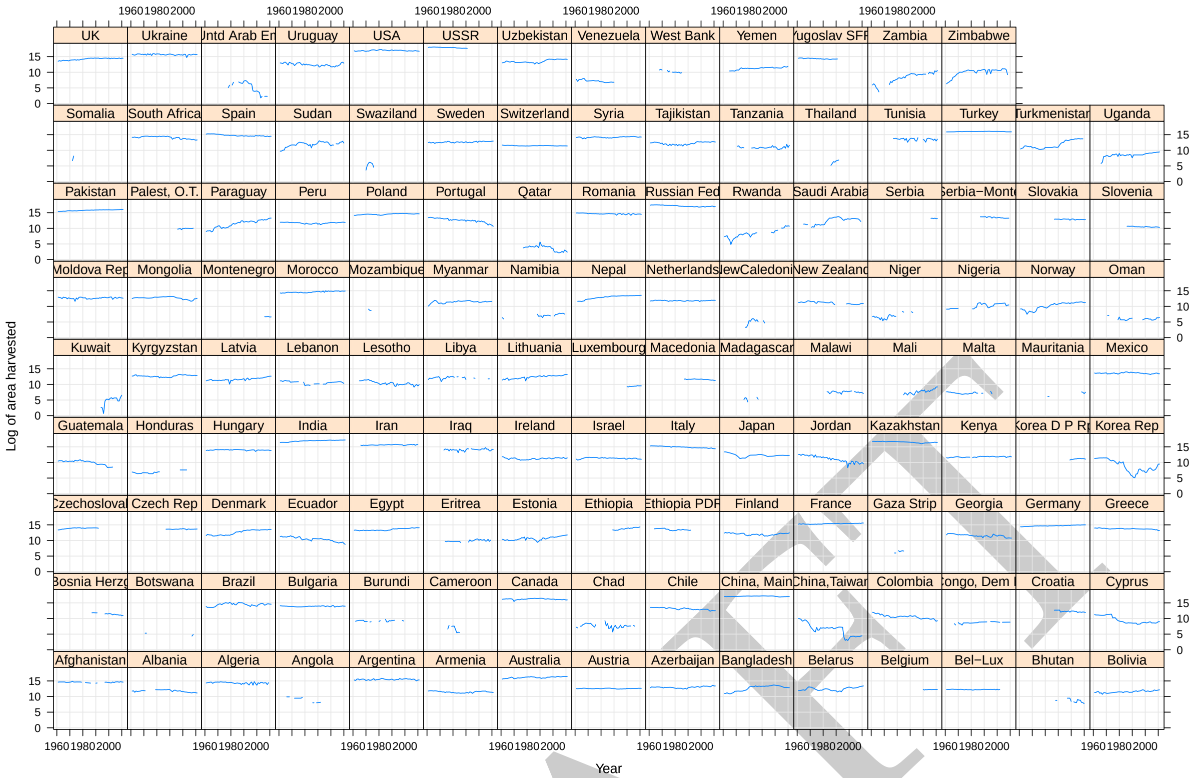


Figure 3: Logged area harvested of wheat from 1961 to 2011 by country

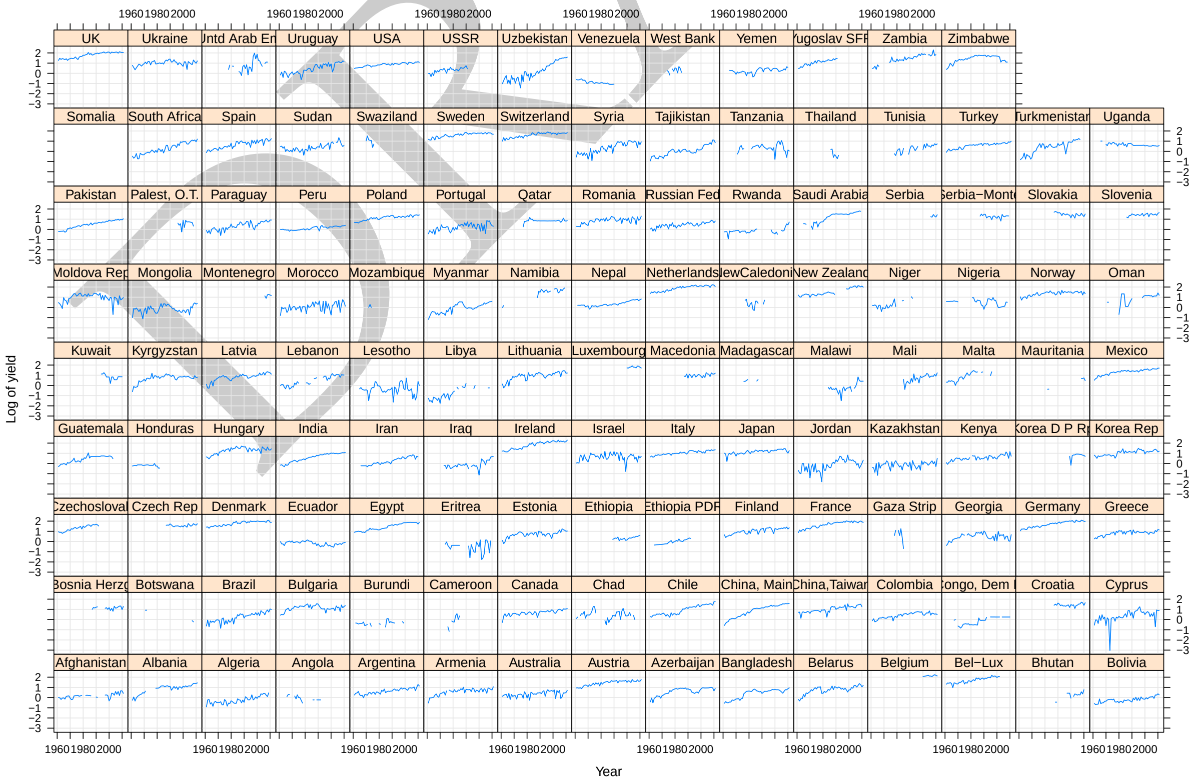


Figure 4: Logged yield series of wheat from 1961 to 2011 by country

3.2. Missing Data Mechanisms

Experience from domain experts in the past suggests that missing value are more common in countries which lack statistical capacity. In addition, commodity with low national priority for countries are often missing.

The following two graph illustrates this phenomenon, where countries are classified by their average value and the redder the cell the higher the production. We assume that the data is of Missing at Random (MAR) and that the distribution of missingness depends solely on the size of the production.

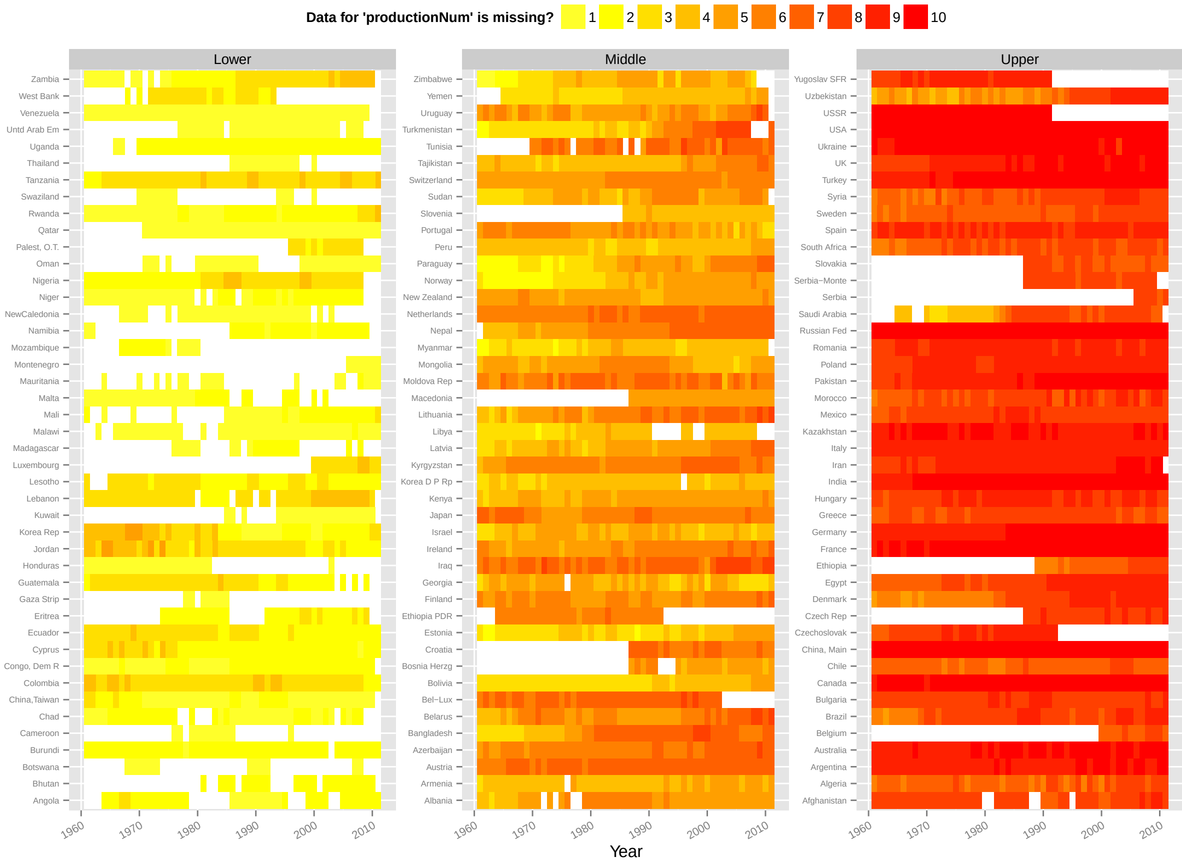


Figure 5: Missing pattern of wheat by value of production

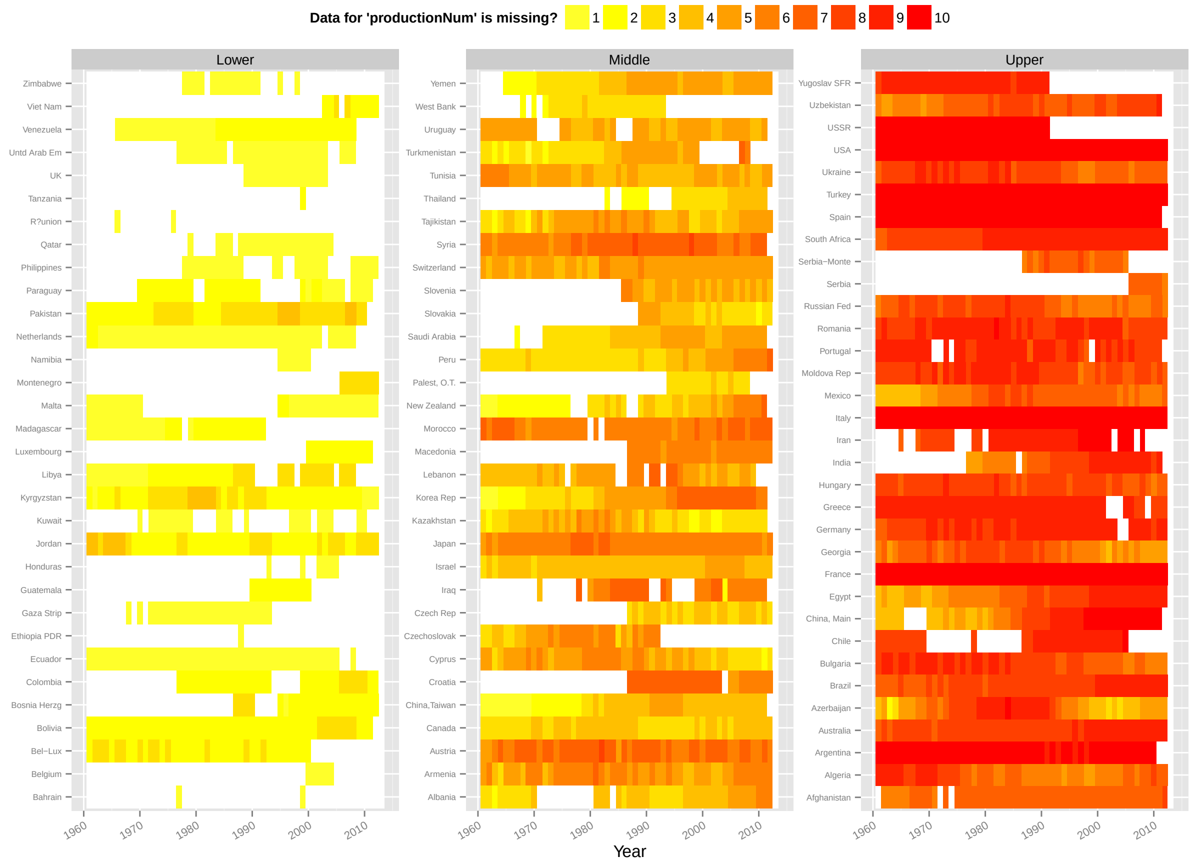


Figure 6: Missing pattern of grape by value of production

3.3. Data Quality Issues

During the development of the methodology, we have encountered several cases which required us to review and redefine our initial methodology. These are not exceptions, rather they are prevalent in the production domain and the analyst should bear in mind of these characteristics.

Sparse Data

Although missing values are expected given that the goal is to impute missing values, but one might be shocked at the sparsity of the data. Shown in figure 7 and 8 are just two out of several hundreds of cases where only a handful of points are observed over the past fifty years.

In particular, rice is the staple of Bhutan where accurate imputation of the missing value is critical in the estimation of the total food/energy supply and ultimately the undernourishment level of the country.

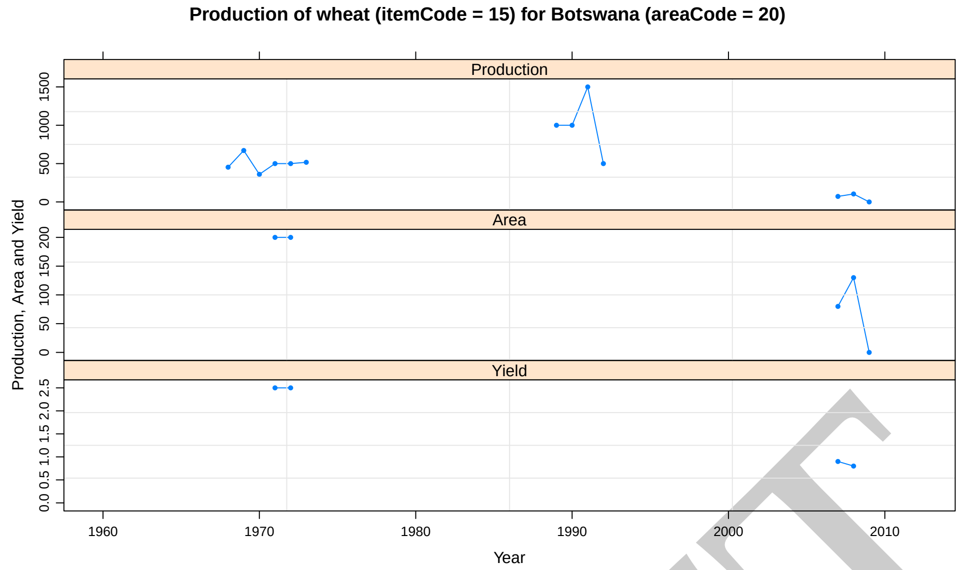


Figure 7: This example illustrates the extreme sparsity in some of the production domain, 13 observation for production while 4 for both area harvested and yield over a period of 50 years.

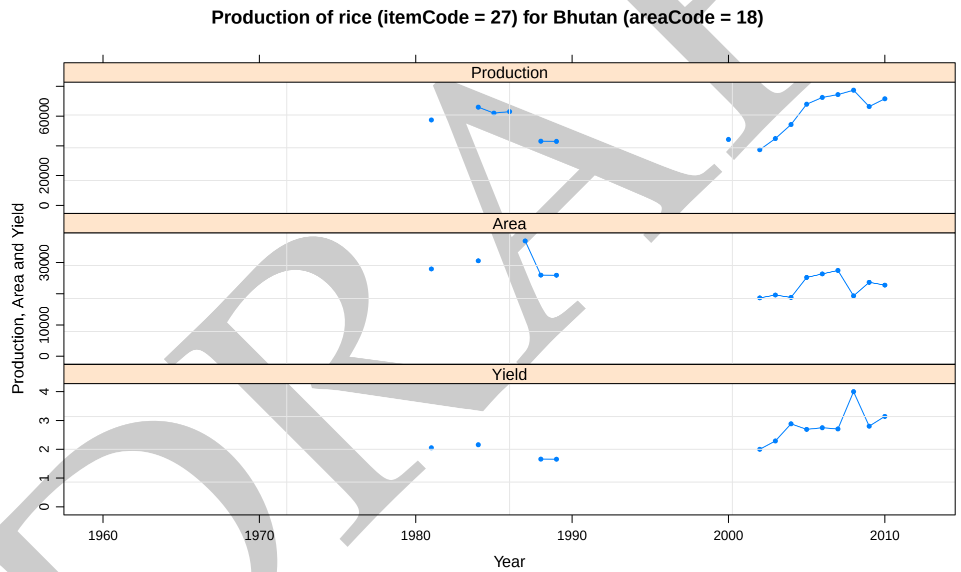


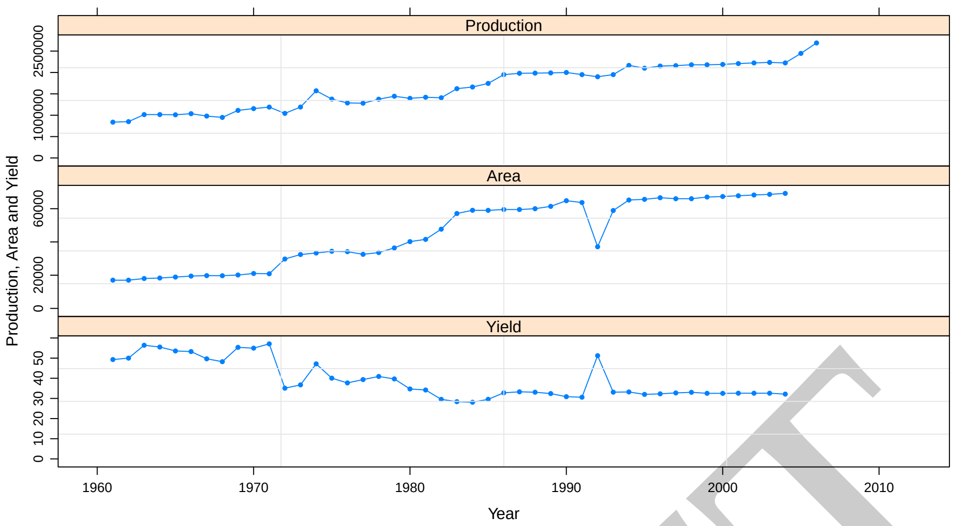
Figure 8: Given that rice is the main staple in Bhutan, it is extremely important to be able to impute accurately.

Peculiar/Divergin Trends

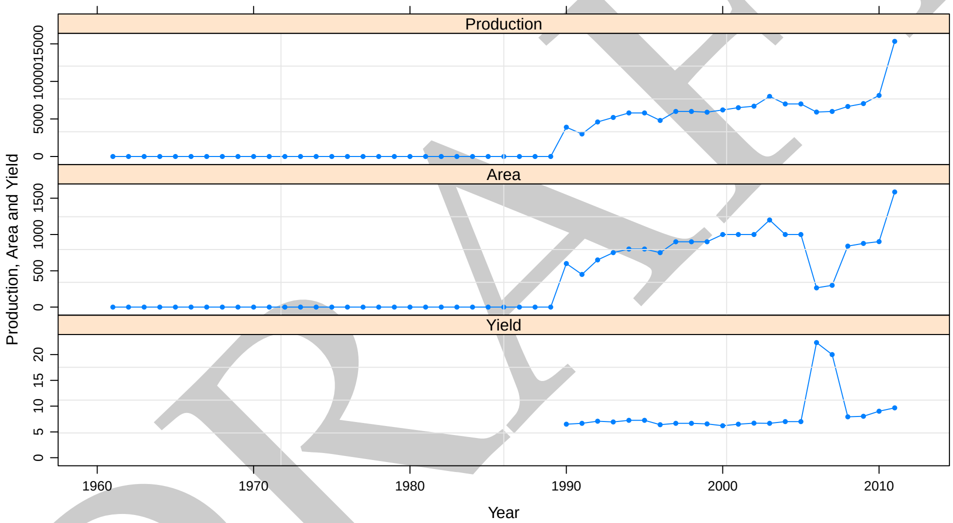
Another issue arose from the quality of the data collected reported and recorded. Here we observe suspicious trends or shocks of area harvested which results in unreasonable yield.

Take the sugar cane production in Madagascar for example, the yield suddenly shot up from a level around 30 to 50 while the area harvested dropped in 1992. We have no explanation of this other than poor data reporting. Similar phenomenon has been observed for garlic in Albania and bananas in Greece where the yield increased by as much as eight-fold from 2 to 16 in a single year.

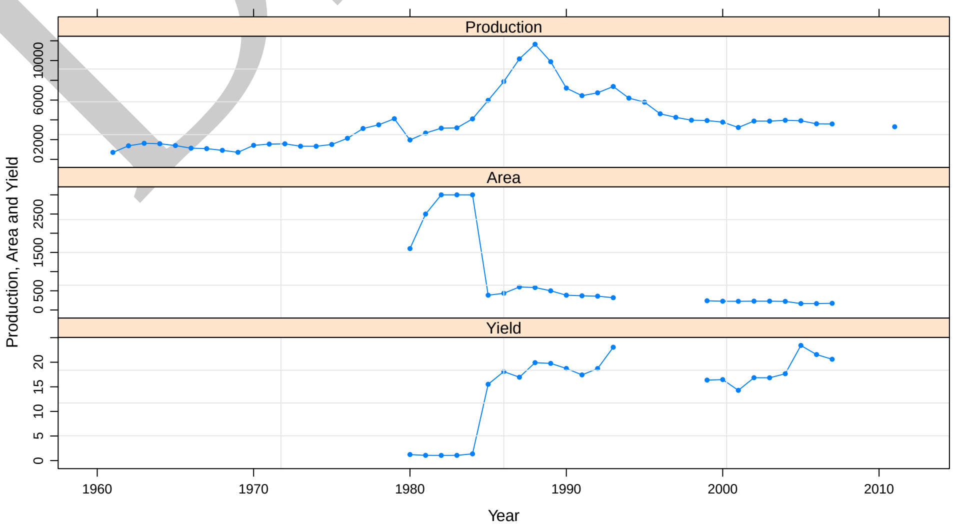
Production of sugarCane (itemCode = 156) for Madagascar (areaCode = 129)



Production of garlic (itemCode = 406) for Albania (areaCode = 3)



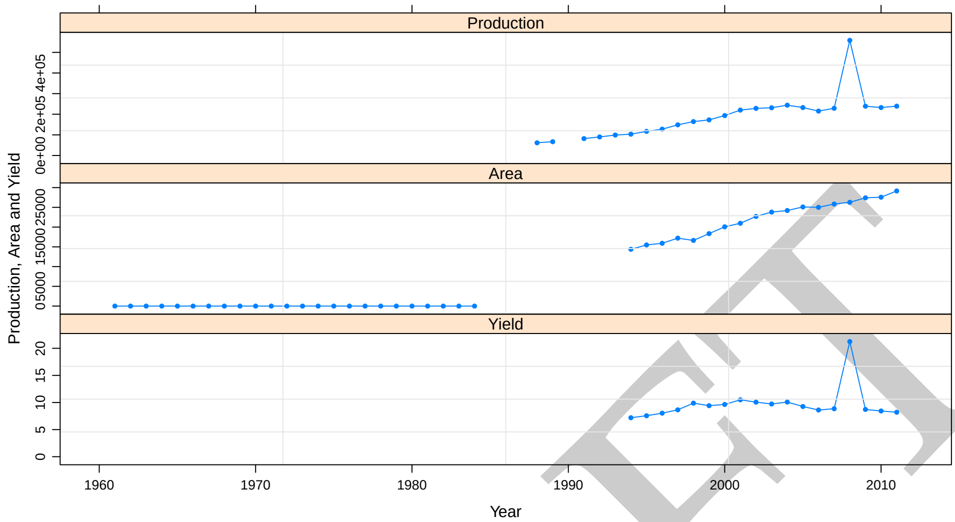
Production of banana (itemCode = 486) for Greece (areaCode = 84)



Shocks

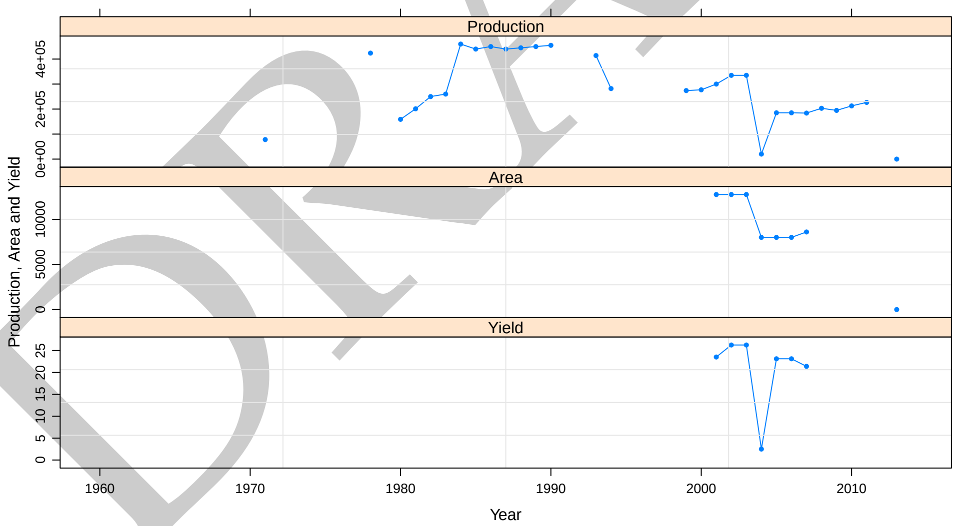
Finally, we present cases whether shocks are present. Sometimes, this may be a data quality issue in the case of coconut production in China where the wrong data was recorded in 2008.

Production of coconut (itemCode = 249) for China, Main (areaCode = 41)



On the other hand, the shock in Iraq is true in 2004 as a result of the Iraq war.

Production of grape (itemCode = 560) for Iraq (areaCode = 103)



These observations prompt us to use a robust method to safeguard ourself from non-sensical imputation.

4. Proposed Methodology

In order to capture the correlation of yield between countries as a result of climate conditions and avoid imputation of infeasible yield, we propose to impute the yield and area in contrast to production and area in the currently implemented methodology. The added advantage of this approach, with well designed validation, almost guarantees that the series will not diverge.

4.1. Imputation for Harvested area

From the exploratory analysis we can observe that the series is generally 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 and balance with production.

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. This is another reason why we decided to model the stable and low impact yield, take the coconut for China for example, if we model the area harvested process with a linear model, we might obtain an unrealistic yield.

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.

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

Then for values which we can not impute with linear interpolation, we impute with the closest observed 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 utilization of this model enables all information available both historical and cross-sectional to be incorporated. In addition, proposed indicators such as the vegetation index, CO_2 concentration and other drivers can be tested and incorporated if proven to improve predictive power.

Following the notation of Bates, the general form of the model can be expressed as:

$$\begin{aligned} (\mathscr{Y}|\mathscr{B} = \mathbf{b}) &\sim \mathscr{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2\mathbf{I}) \\ \mathscr{B} &\sim \mathscr{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\theta}) \end{aligned} \tag{7}$$

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

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

Where Y denotes yield with ΔY being the grouped change in 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 and temperature.

The grouped change in yield is computed as:

$$\Delta Y_{j,t} = \frac{1}{N_i} \sum_{i \in j} \text{sgn}(\Delta Y_{i,t}) \tag{9}$$

In essence, the imputation of the yield is based on the country specific level and historical trend while accounting for co-movements 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.

4.3. EM-algorithm

Since its introduction by Dempster, Laird and Rubin, the EM-algorithm has been used extensively in statistics for computing the maximum likelihood estimate from incomplete data.This section will give a brief account of the methodology, readers are to refer to the reference list for further details.

In breiveity, the algorithm consists of two repeated iterative steps, the expectation and the maximization step hence it's name.

We first use the naive imputation to populate the data in order to compute the estimate for both $\boldsymbol{\beta}^{(0)}$ and $\Delta\mathbf{Y}$ for initialization. and let $\mathcal{L}(\boldsymbol{\beta}|\mathbf{Y})$ be the likelihood function of equation 8.

Expectation-Step

In this step, we compute the expected value of the missing values given the initial parameters.

$$\hat{\mathbf{Y}}^M = E[\mathbf{Y}^M|\boldsymbol{\beta}^{(p)}] \tag{10}$$

Maximization-Step

The imputed value together with the observed value is then used to estimate the new set of parameter though the maximization of the likelihood.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\ \mathcal{L}(\boldsymbol{\beta}|\mathbf{Y}^{\mathcal{O}}, \hat{\mathbf{Y}}^{\mathcal{M}}) \tag{11}$$

The two steps are iterated until the likelihood function converges.

5. Case Studies

This section is devoted to illustrate how the imputation methodology works in practice. Agricultural products from several commodity groups were chosen to demonstrate the flexibility and robustness of the method.

5.1. Wheat

We begin the case study with the simple case, wheat production around the world from 1994 to 2011. Data for wheat are relatively complete and factors that determines the the production, area harvested and yield are well understood. From table we can see that only 15% of the production require imputation.

	Production Count	Production Percentage(%)	Area Count	Area Percentage (%)
	1917	80%	1833	77%
*	84	4%	57	2%
E	126	5%	168	7%
F	111	5%	176	7%
M	38	2%	56	2%
P	9	0%	11	0%
T	109	5%	93	4%
Missing	0	0%	0	0%
Total	2394	100%	2394	100%

In figure 9, we plot the official and semi-official yield collected from the country national statistics offices and the fitted value of the imputation model. The fit appears to be satistsfactory, and well capture the country specific trend and seasonal shocks where applicable.

Nevertheless, there are two cases Zambia and United Arab Emirates (UAE) where the fit does not appear to be satisfactory. In the case of UAE, this is a problem for extremely small producers where the variance of the ratio (yield) can be extremely volatile. On the other hand, further information are required to understand why the yield in Zambia increased for a single year in 2008.

The imputation exercise not only serve to assist teams to produce sound estimates of the missing value, but at the same time it helps us to re-examine the quality of the data and improve the data collection process accordingly.

In figure 10, the imputed production in blue is superimposed on the observed production depicted in pink. The two lines coincide when the production is observed, the important part is to examine whether the imputed value in blue shows peculiar trend or divergence from the observed data. This does not appear to be the case, all the imputed value appears to be reasonable. The same plot for area is shown in figure 11 and similar results are observed.

Overall, the fit of the yield appears to be fine and additional work are required to understand the observed bahaviour of Zambia nad United Arab Emirates. Furthermore, we see no cases of divergence between area and production associated with the previous imputation methodology.

Figure 9: Observed yied versus fitted yield. Here the pink line represents official or semi-official data while the blue line represents the fit to the data. Empty boxes implies that no yield was observed and imputed for that country.

Figure 10: Observed production versus imputed production. In order to make the countries comparable, we have log transformed the data.

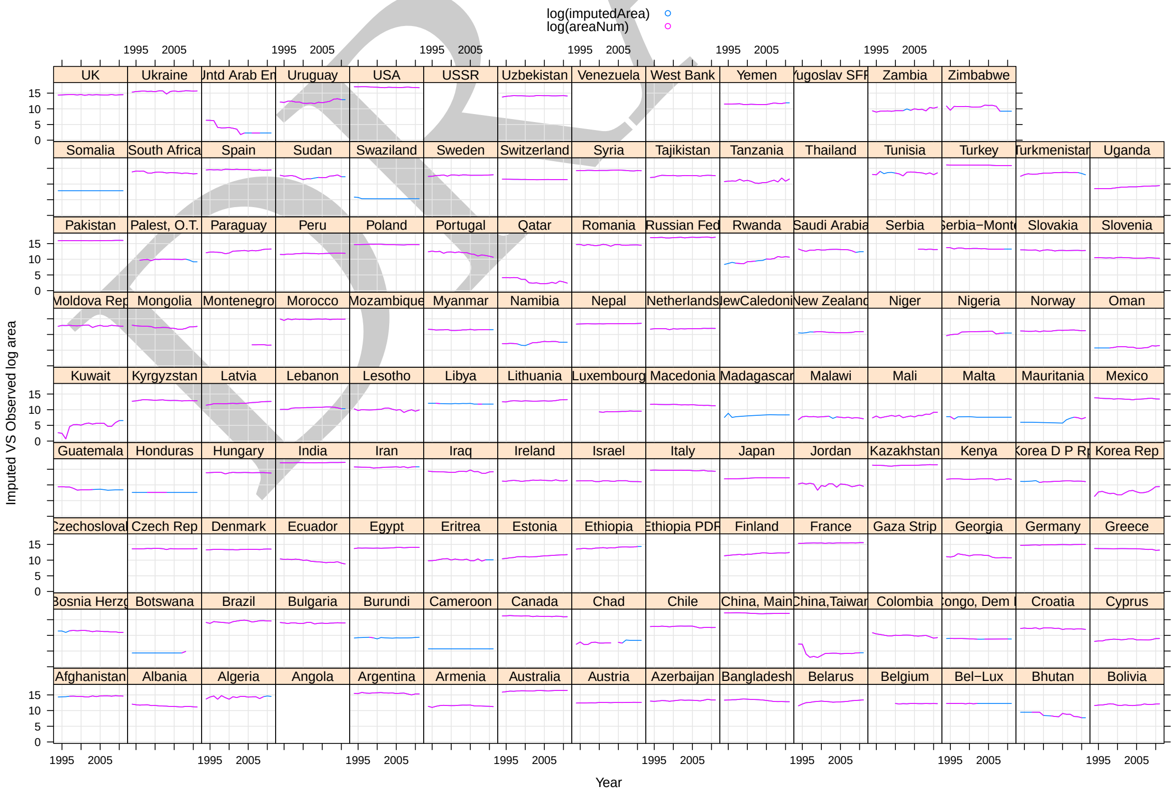


Figure 11: Observed area versus imputed area. Similar to the production, the area is log transformed.

5.2. Grape

In the second example, we take a slightly harder commodity to illustrate some of the power of linear mixed model and the the robustness of the current methodology. From the same flag table, we can see that around 22% of production requires imputations and we have a smaller number of producer.

	Production Count	Production Percentage(%)	Area Count	Area Percentage (%)
	1287	71%	1219	67%
*	52	3%	55	3%
E	135	7%	146	8%
F	171	9%	206	11%
M	59	3%	58	3%
T	114	6%	116	6%
Missing	0	0%	18	1%
Total	1818	100%	1818	100%

Figure 12 depict the same fitted yield super-imposed on observed yield like in the last case study. In general, again the observed fit is reasonable with no major departure. The major feature to notice is the fitted value of Vietnam and India, it illustrates the robustness of the linear mixed model oppose to the simple linear regression.

In the case of Vietnam, a much steeper trend would have been fitted if we were to use simple linear regression and negative yield would have been the result for earlier years. Nevertheless, given that the country may not depart drastically from the regional trend, the fit of the linear mixed model is much more resonable and feasible.

The same characteristics is illustrated for India, where a significant drop is observed in 2010 but the fitted value are only minorly affected by that single point which also suspect to be a recording error.

Again, both the imputation for both area production and area harvested appears to be feasible and does not shown any signs of model failure.

Figure 12: Observed yied versus fitted yield

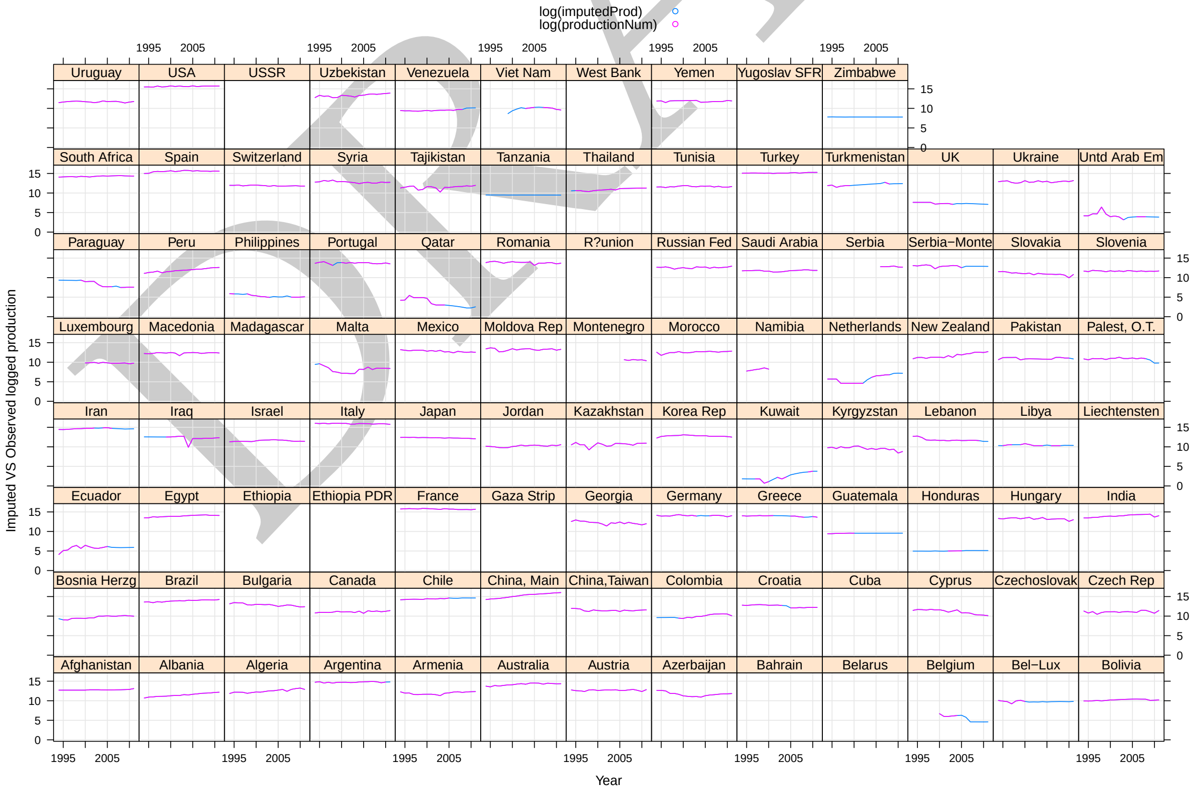


Figure 13: Observed production versus imputed production

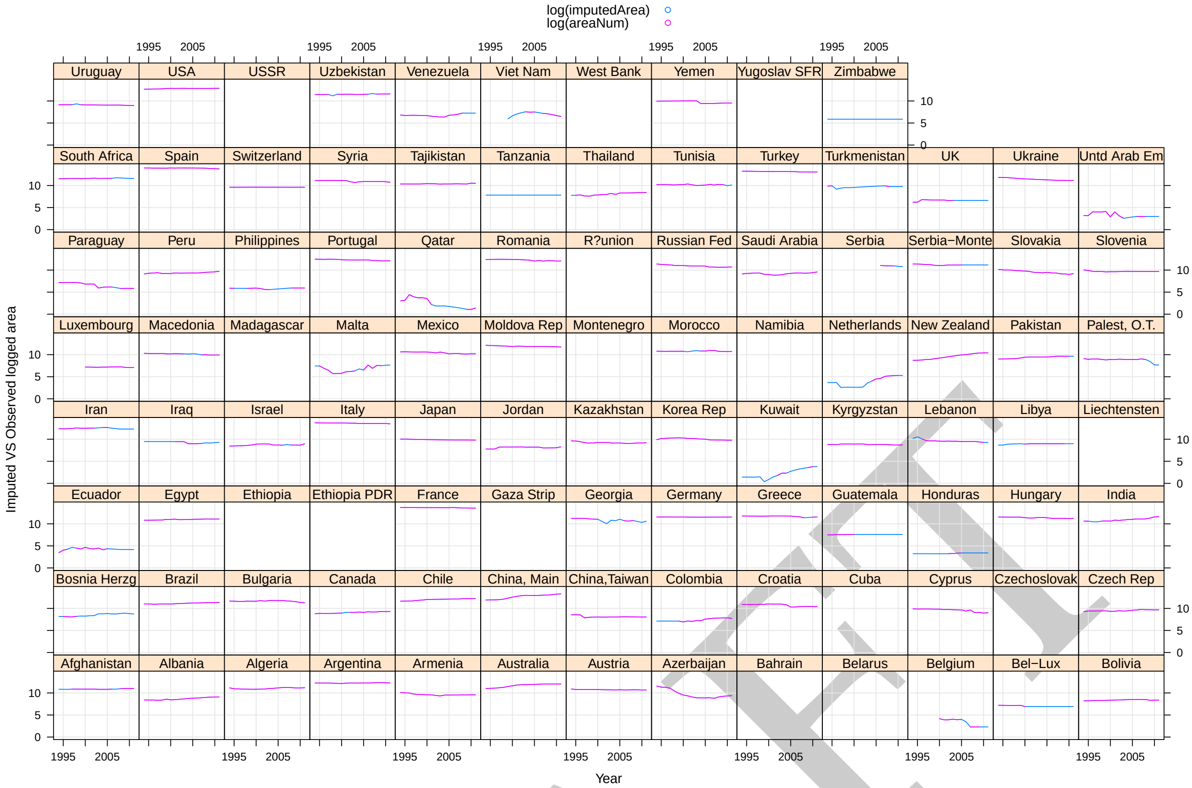


Figure 14: Observed area versus imputed area

5.3. Pepper

The final case study can be considered as one of the hardest and often problematic yet important commodity, pepper. In this case, we would need to impute 67% provided that only one-third of the data are present. From a statistical point of view, such a task should not be performed since the result is highly unreliable. Nevertheless, we will try and examine the result and performance of the imputation.

	Production Count	Production Percentage(%)	Area Count	Area Percentage (%)
	211	27%	198	26%
*	27	3%	4	1%
E	263	34%	228	29%
F	221	29%	237	31%
M	33	4%	60	8%
T	19	2%	11	1%
Missing	0	0%	36	5%
Total	774	100%	774	100%

Looking at figure 15, we can clearly see that the observed yield of Costa Rica in pink is displaying extremely erratic behaviour. Further analysis is required to understand the source of this volatility in order for us to improve the methodology. Nonetheless, most of other series observed are smooth and fitted well by the linear line.

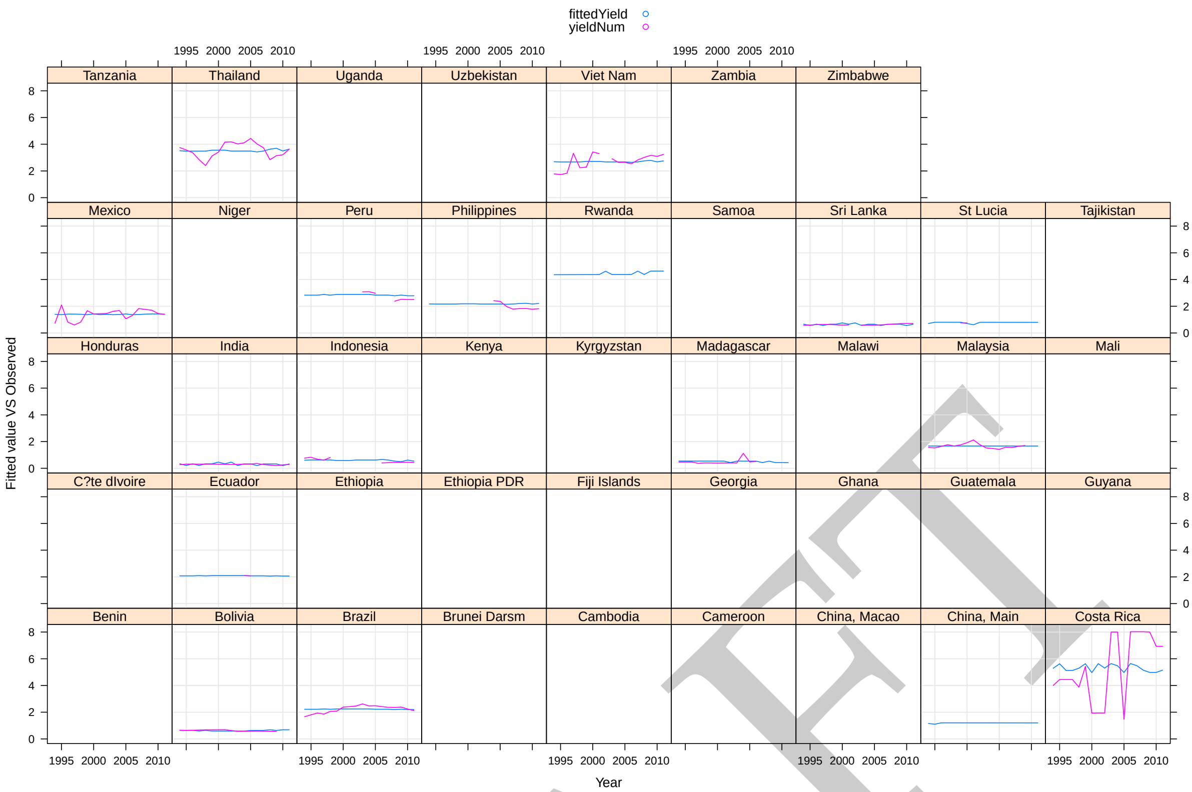


Figure 15: Observed yied versus fitted yield

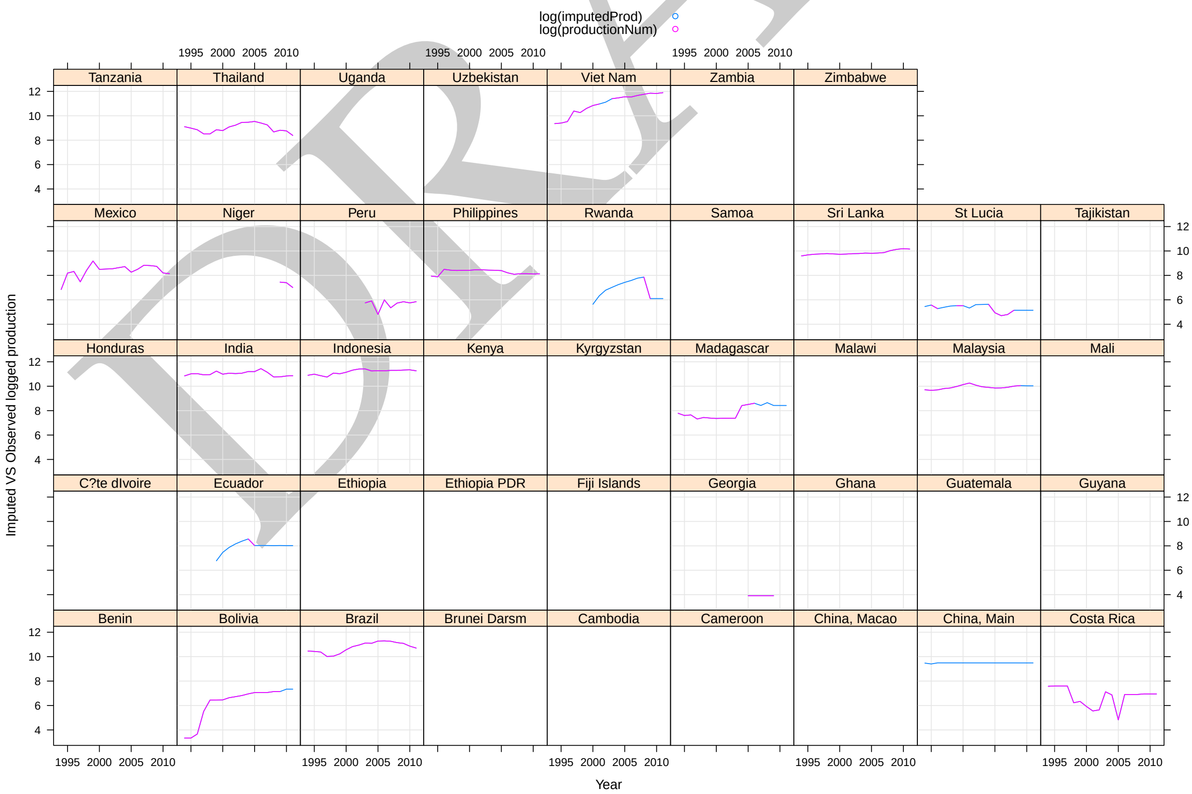


Figure 16: Observed production versus imputed production

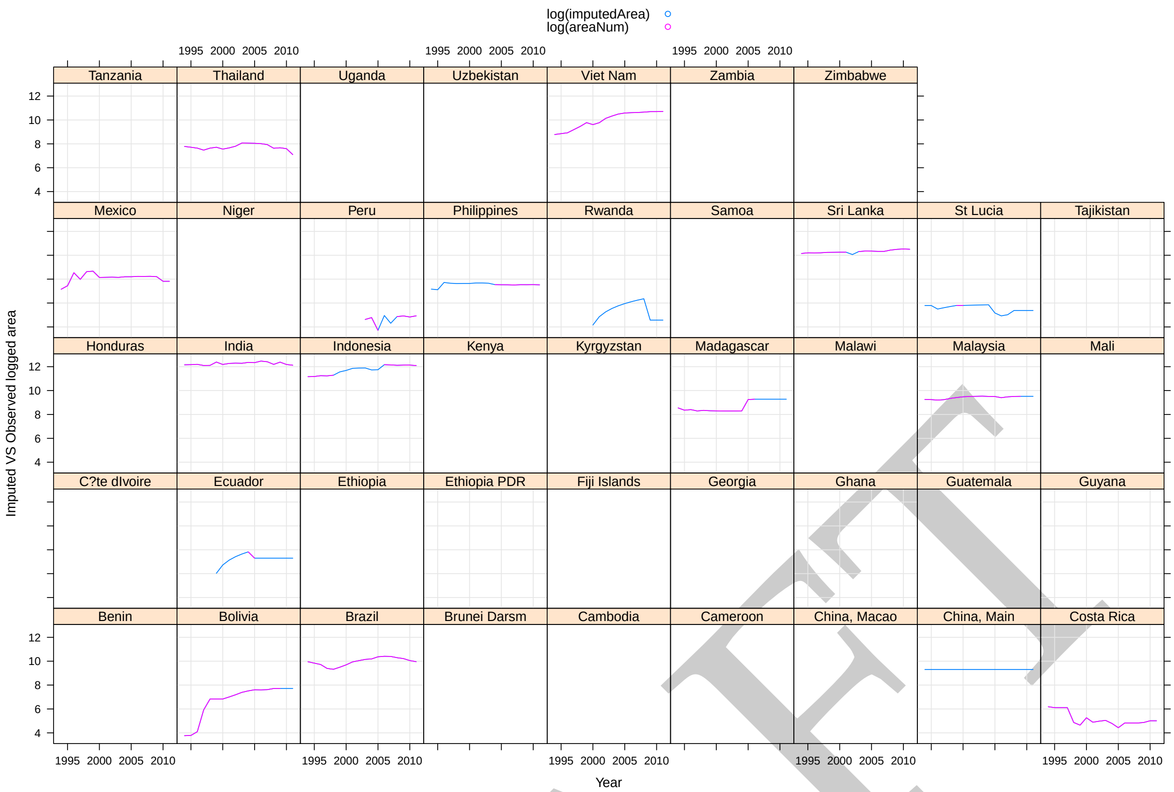


Figure 17: Observed area versus imputed area

6. Simulation Results

In order to understand the capability and performance of the imputation, we conduct bootstrap simulation to estimate the prediction error.

For each bootstrap, we take a sample containing only official and semi-official data and impute the values which are issing. The imputation is then benchmarked with the actual observed official and semi-official data. We use the Mean Absolute Percentage Error (MAPE) for assessing the accuracy of the imputed values; and we compute the coverage rate defined as the proportion of missing value imputed to examine the applicability of the method.

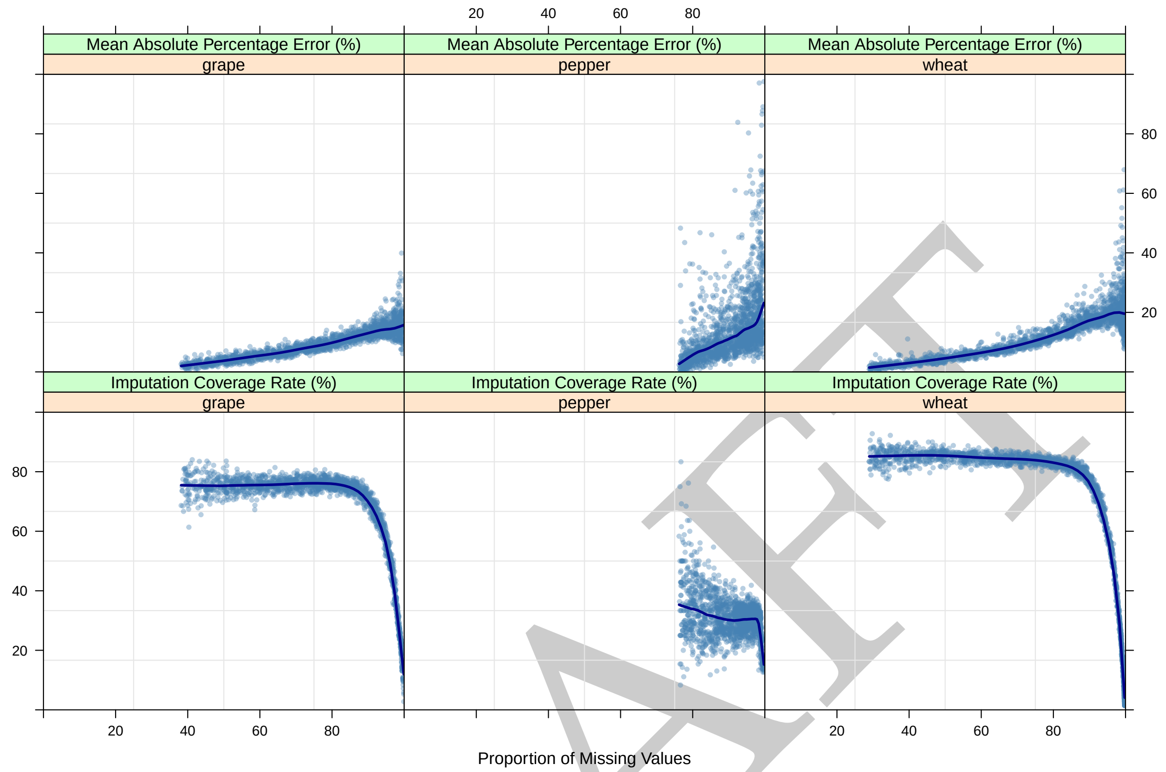
$$\text{MAPE} = \frac{1}{N}\sum |y_i - \hat{y}_i| \tag{12}$$

Each simulation draws a sample of varying missing proportion, this is to investigate the prediction error over different degree of missingess and at the same time to detect the breaking point of the method.

Since there are already missing value in the data, the benchmark is can only be computed on the available official and semi-official data which varies between commodity. We have over 70% availability of benchmark observations for wheat, while only slightly over 20% for pepper.

The simulation in particular for wheat and grape shows that the imputation has a high coverage rate and accuracy. For example, if half of our data are missing values, the simulation shows we can impute approximately 75% of the missing values with prediction error around 5%.

On the other hand, the coverage rate and the accuracy is much lower for pepper. The main reason for this is due to the fact that much of the robustness and accuracy comes from the regional and global information set. The number of countries producing wheat (133) and grape (101) is much larger than pepper (37) and we have less information to utilize and thus the observed result.



7. Conclusion and Further 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. Iterative imputation methodology such as full conditional specification are considered to imputed the area harvested, production and yield simultaneously. In addition, additive and robust mixed models are under-investigation for country non-linearity departures and shocks.

Acknowledgement

This work is supervised by Adam Prakash with assistance from Nicolas Sakoff, Onno Hoffmeister, Luigi Castaldi, 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: 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 November 20, 2013and is subject to changes and updates.

Annex 2: 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 3: Diagnostic graph of case studies

Wheat

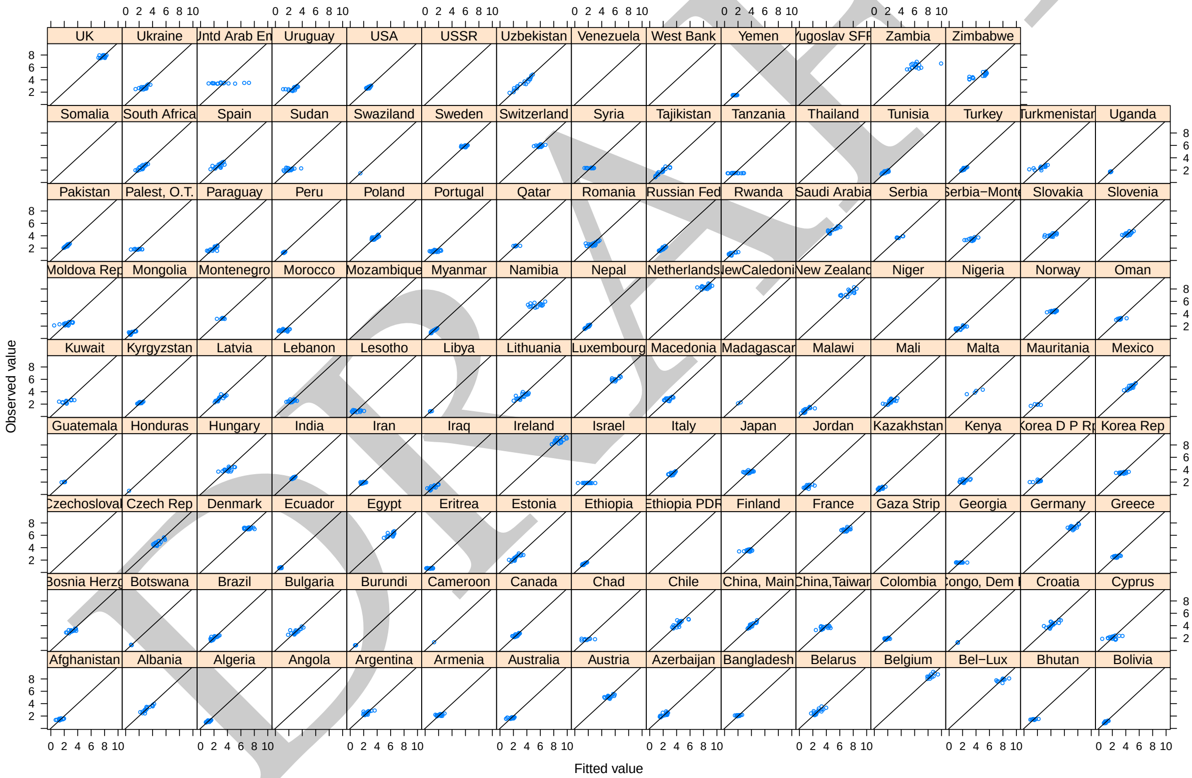


Figure 18: Observed versus fitted value

Grape

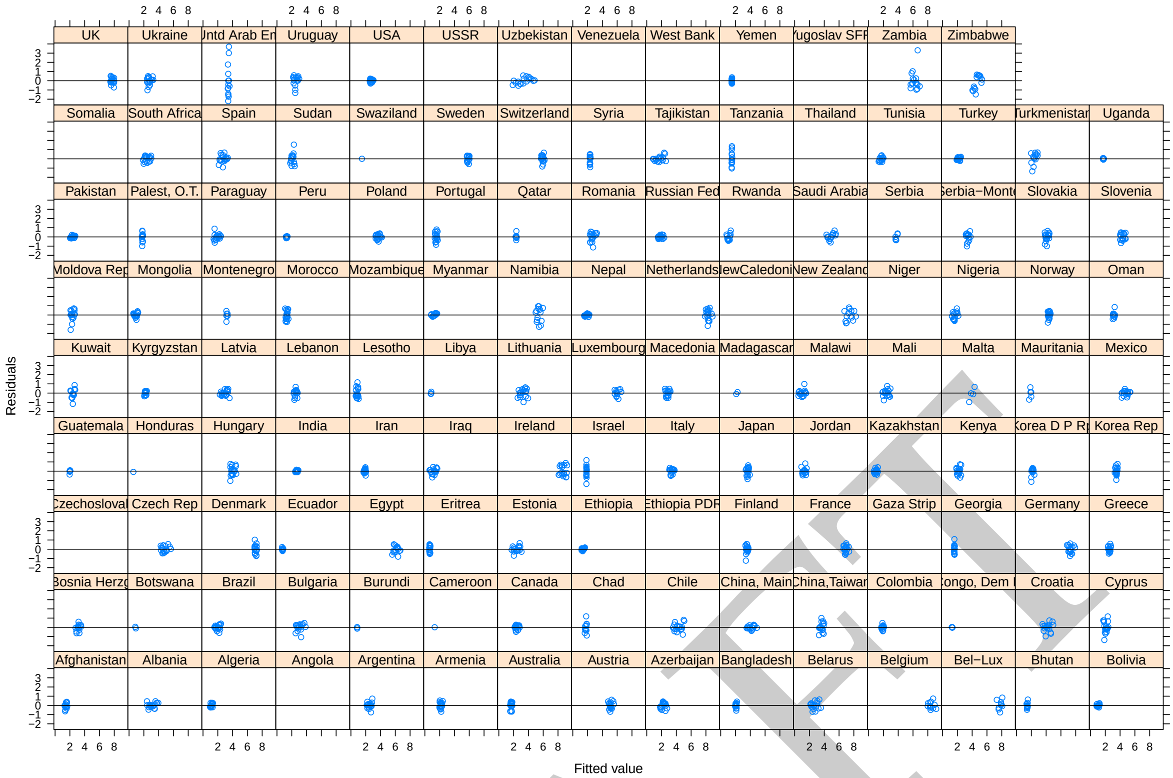


Figure 19: Residuals versus fitted value by country

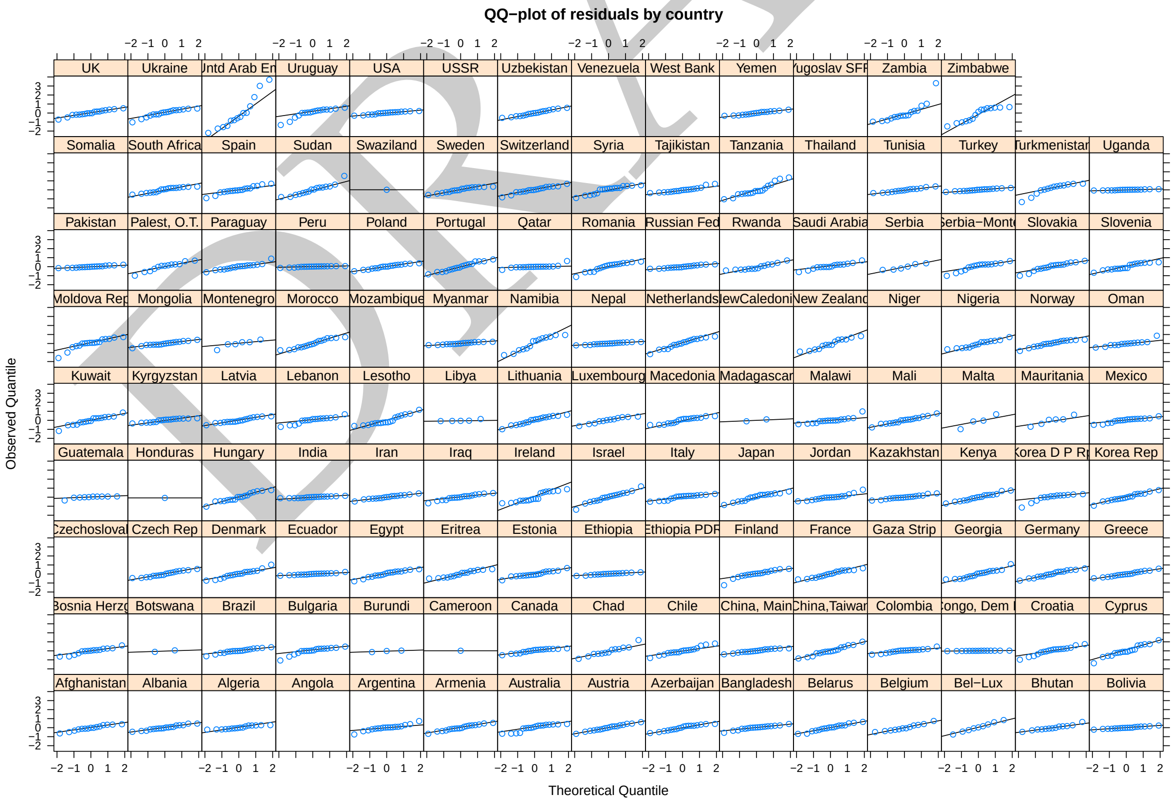


Figure 20: Normality of residuals by country

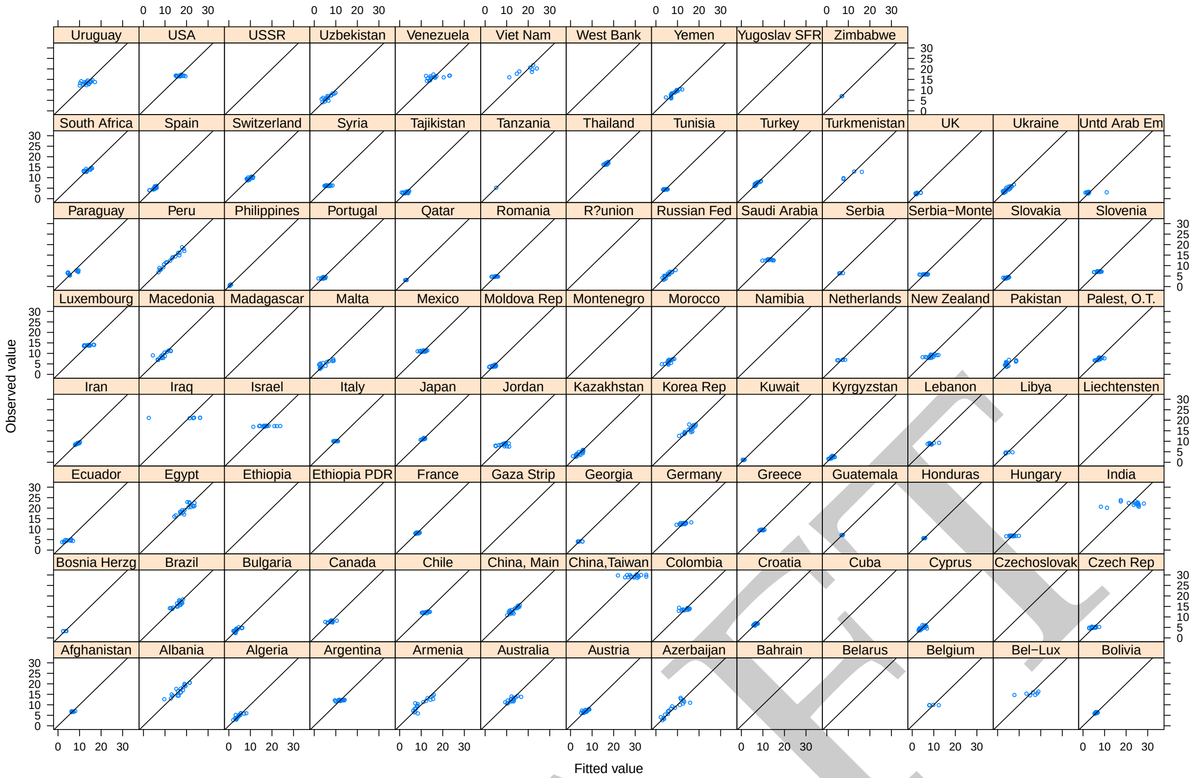


Figure 21: Observed versus fitted value

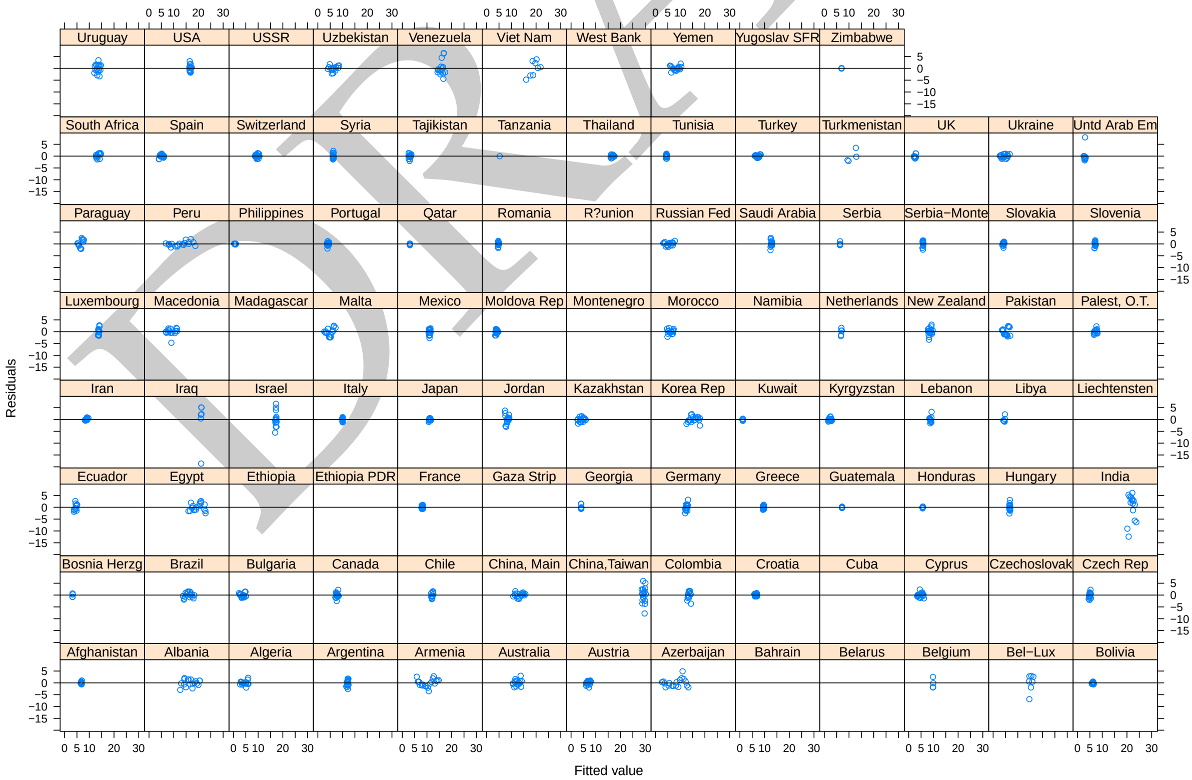


Figure 22: Residuals versus fitted value by country

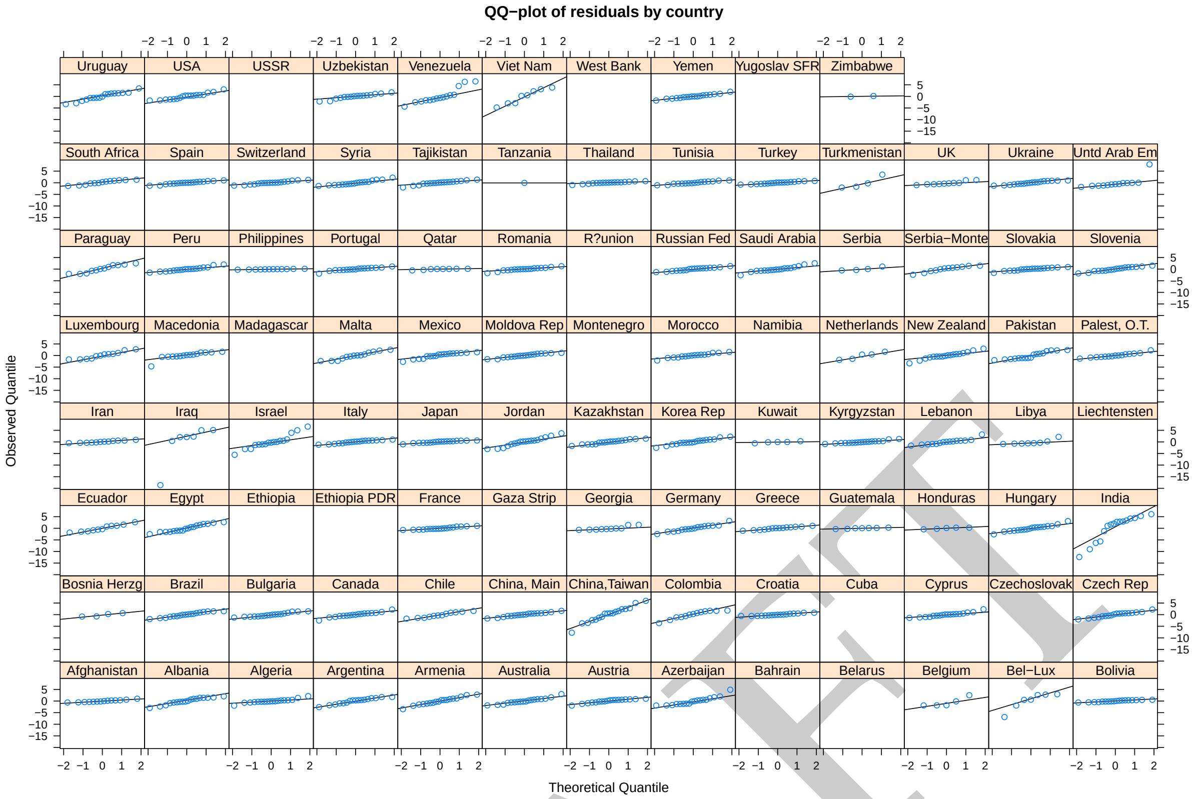


Figure 23: Normality of residuals by country

Annex 4: Pseudo Codes

Algorithm 1: EM-Algorithm for Imputation - function *shocklme4*

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**
    E-step: Compute the expected value of the missing values;
    $\hat{\mathbf{Y}}^{\mathbf{M}} \leftarrow f(\mathbf{Y}^{\mathbf{O}}, \hat{\boldsymbol{\beta}}^{(i)})$
    M-step: Estimate the Linear Mix Model in 8;
    if $\mathcal{L}_{new} - \mathcal{L}_{old} \geq \mathcal{E}$ **then**
      $\hat{\boldsymbol{\beta}}^{(i+1)} \leftarrow \hat{\boldsymbol{\beta}}$;
    end
    else
      break
    end
  end
end

Algorithm 2: Imputation Procedure - function *swsImputation*

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

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 Statis-

tical 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, *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.0-4. http://CRAN.R-project.org/package=lme4, 2013.

[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