Statistical Working Paper on Imputation 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 ensemble learning.

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 begin with an exploration of three selected variables in the production domain and an attempt to pin point their potential drivers. These provide supports and explainations to the newly proposed methodology.

A detailed account of the methodologies is provided. The linear mixed model demonstrates ability to capture cross-country and cross-commodity information which are present in the yield series. On the other hand, the ensemble learning display flexible yet robust characteristics for the imputation of the production where traditional method of applying a single model will fail.

Keywords: Imputation, Linear Mixed Model, Agricultural Production, Ensemble Learning.

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.

This paper is dynamically generated on September 10, 2014 and is subject to changes and updates.

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 im-

putation.

The relationship between the variables in the production domain by definition can be expressed as follow,

$$P_t := A_t \times Y_t \qquad P_t \geq 0,\ A_t \geq 0,\ Y_t > 0 \tag{1}$$

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.

It is important to recognize that when area harvested and production are both zero, yield is undefined. Yield is a derived statistic, and when we can not observe it through calculation it does not imply it is zero. In fact, it is a missing value by the very nature of the definition being unobservable yet a value does exist.

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)

1.1. Scope of the project

A total of 169 commodities just in the crop domain and 19 primary livestock and 59 processed livestock. There are in total of 245 countries including obsolete classifications and territories which result in more than 180,000+ potential times series to impute.

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.

3. Exploratory Data Analysis

We first take a visualization tour of the data to grasp an understanding of the underlying pattern of the production domain and the relationship between the variables.

3.1. Yield

The next three graph depicts the yield of three selected commodity from different commodity group, wheat, grape and beef.

From the graphs, we can first observe that there is a general increasing trend in the yield accross all countries and commodities illustrated. Similar stories are observed in almost all commodities that has been studied during the development of the methodology. This is a result of continuous advancement in both technology and agricultural practice driven by research & development. Improved irrigation provides crop with sufficient and uninterrupted water source, while tailored compound feed provide the precise nutrient requirements ensuring the livestocks consume the optimal diet for growth. Regardless whether these practice are sustainable or beneficial, there are strong evidence of increased productivity over time.

Nevertheless, just like all available technology such as internet, the distribution is far from perfect. The adoption of technology depends on the access which may be hindered by the presence of patents, or it may be restricted by service providers. Imperfect information and limit financial resources are also major obstacles for embracing the new developments, this is particularly true for countries where the majority of the producers are smallholders or rural farming.

Furthermore, producers faces different constraints and cost. Countries such as Brazil and Russia which has a large amount of arable land does not bear the same cost for land acqusition in comparison to small states such as the Netherlands. The cost translates to different pressure to improve productivity and yield. Inovations required are also different for countries, wheat breeding for the development of drought and diease resistent variety were crucial to withold Australia's dry climate.

Despite the differences among the countries branching from various combination of technological advancement and economic condition, these factors all contribute towards a positive improvement in productitive which can be estimated as an aggregated mixture effect.

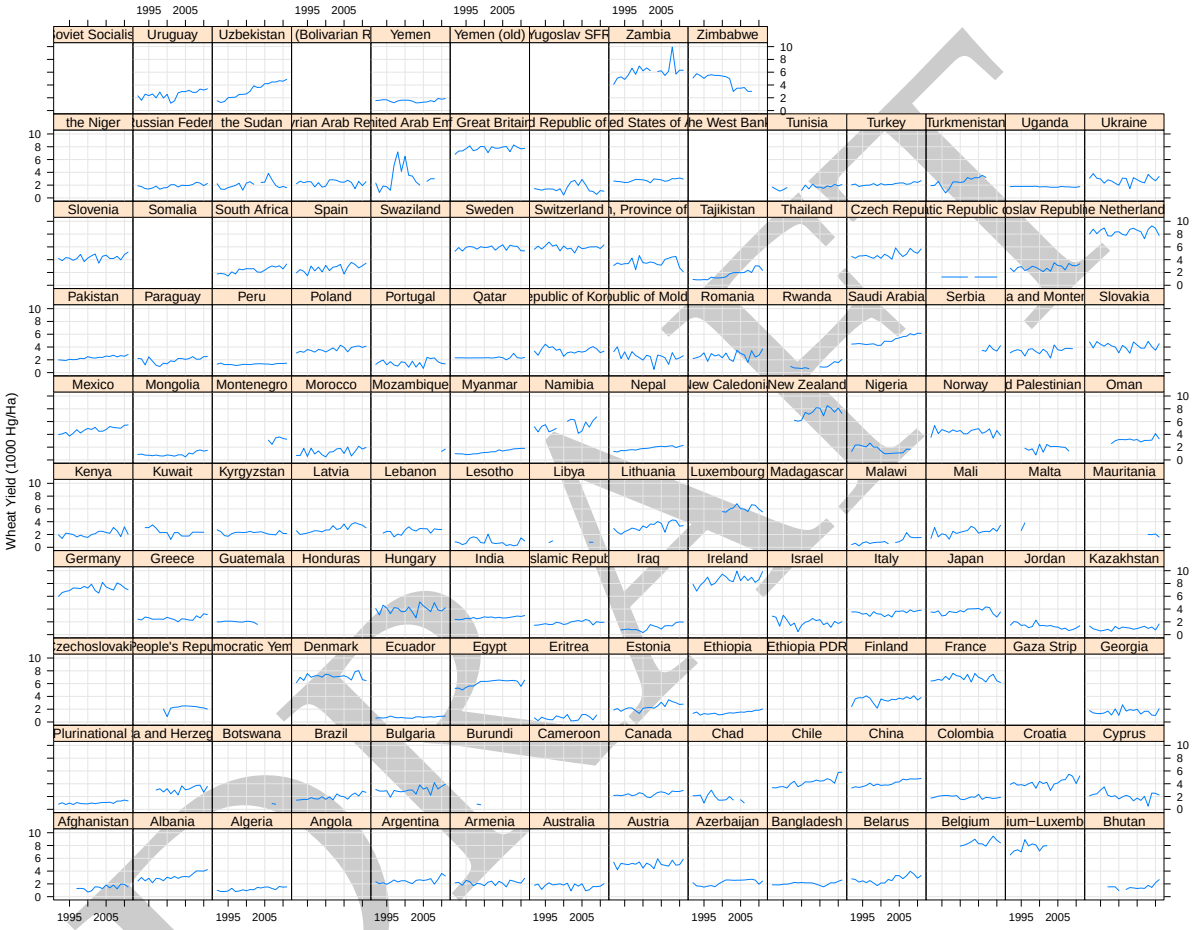


Figure 1: This figure illustrates the yield of wheat accross all countries, it provide strong support to the facts previously mentioned. First of all, we can observe the concordant increasing trend accross all country where technological innovation such as improved seed, and synthetic nitrogen fertilizer contributed to the increase in productivity. Yet at the same time, we can also observe that the rate of growth differ between countries.

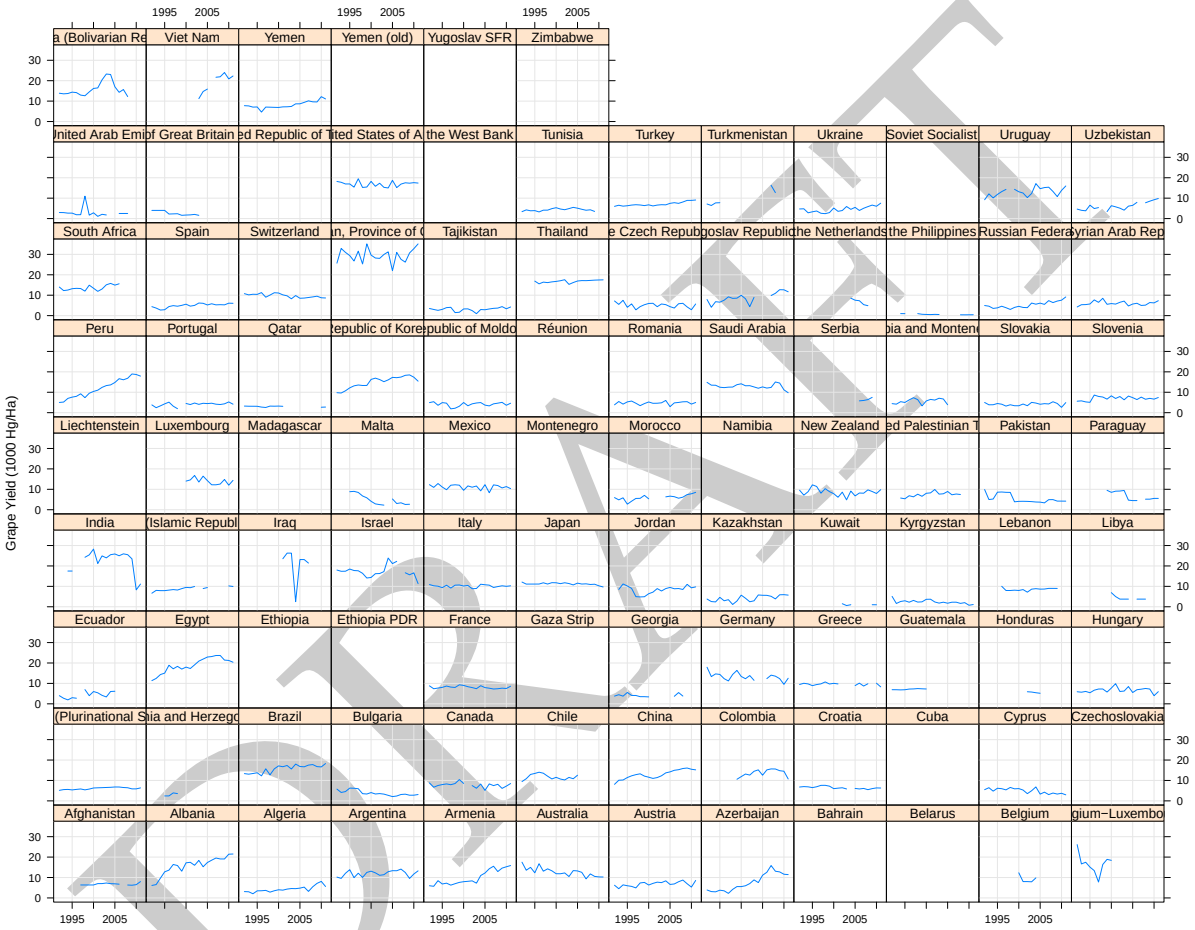


Figure 2: Unlike the yield of wheat, the yield for grape has remain rather constant over time except a few selective country such as Peru and Azerbaijian. There are a few spikes observed, namely Iraq, the invasion of Iraq may have contributed to the negative shock.

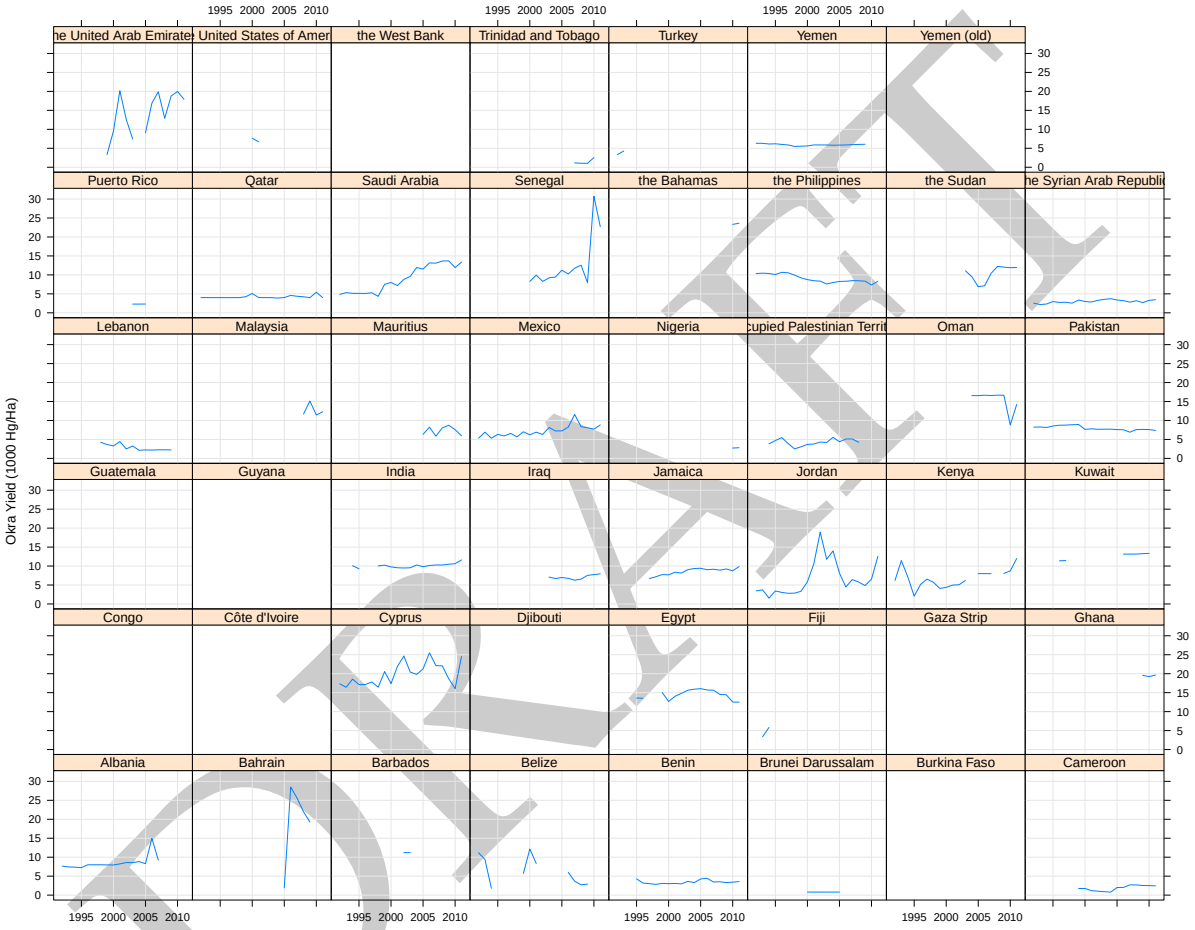


Figure 3: Shown in this graph are the yield of Okra over time. We can observe that the data is extremely sparse, further the quality of the data is questionable. Yield growth from less than 10 Hg/Ha to greater than 30 Hg/Ha in a single year for both Bahrain and Senegal is deemed suspicious.

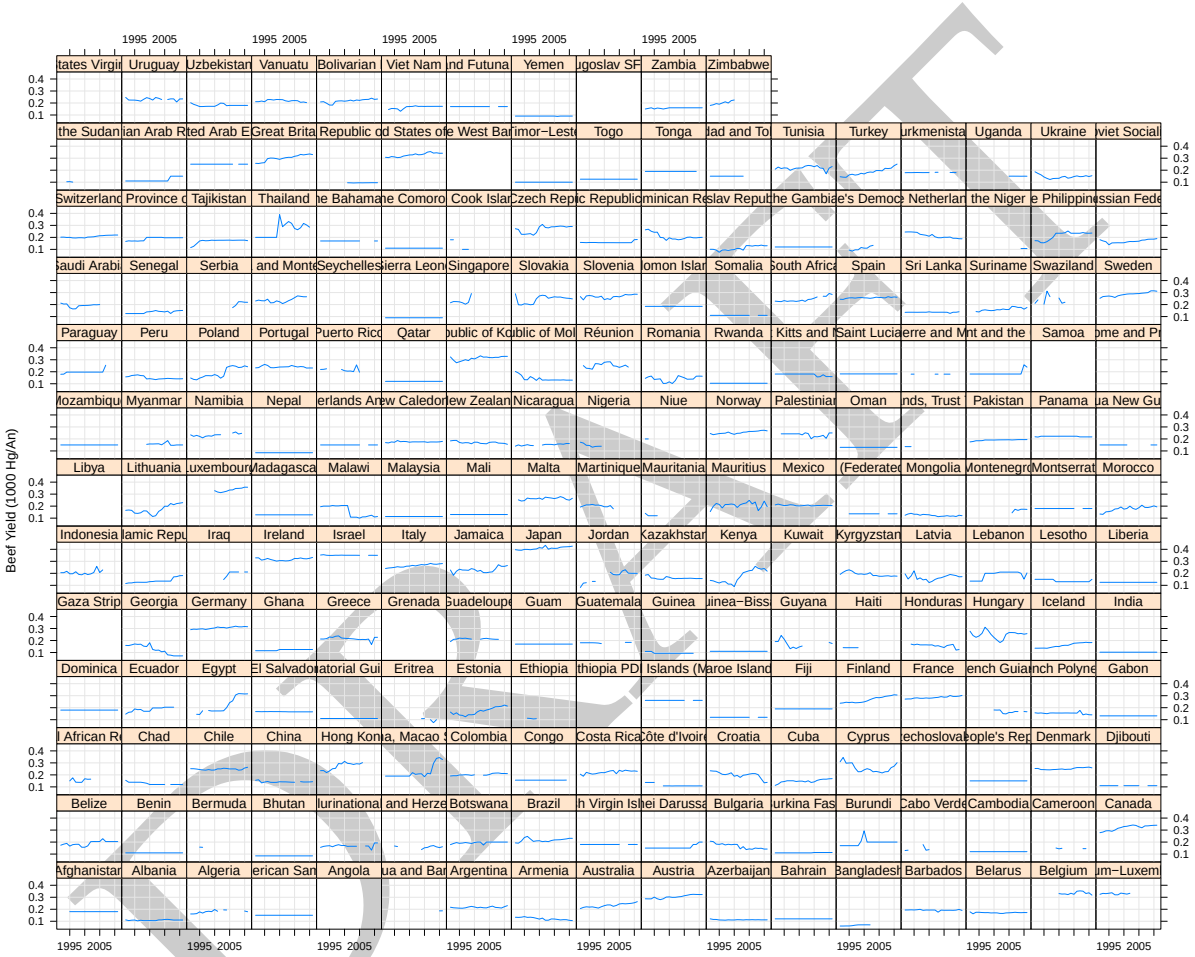


Figure 4: Similar to grape, the carcass weight per animal of beef also shows the same pattern of a very modest improvement in yield with a handful of exceptions.

3.2. Production and Area Harvested

Although yield plays a vital role in the production process, the actual quantity of production is usually dictated by the area sown and harvested. The illustrations in this section shows that the production series is usually dominated by how much area was planted and harvested.

In contrast to the simple mechanism of yield where all dominant factors contribute towards improving the productivity, the mechanism of production is much more unpredictable.

Production is determine by area harvested and hence area sown in the previous period by the farmer. Which ultimately depends on the perception of information and subjective judgement of the producer. Production can increase or decrease production as a response to the state of the market, wheat field can be substitue to harvest sorghum if prices are expected to be high. Further, individual entities faces different risk profile, even under the assumption of all producers are profit-seeking the risk profile may alter the portfolio of products held by the producer. Markets has been known to be difficult to forecast, let alone the prediction of human judgement is just shy of impossible.

Only in cases where the commodity is a major staple or exporting item, we can observe simple trend explained by the continuous increase in demand. On the other hand, commodities which are of relative lesser importance, the pattern of the production may display unpredictable erratic behaviour.

One worthnoting point is that production which display simple pattern are typically commodities in countries which are highly commercialized or important for consumption as staple or trade for financial resources. These item are rarely the one we need to impute as they are generally complete, it is commodities of lesser importance that are required to be imputed and as a result that are often harder due to their unstable demand and nature.

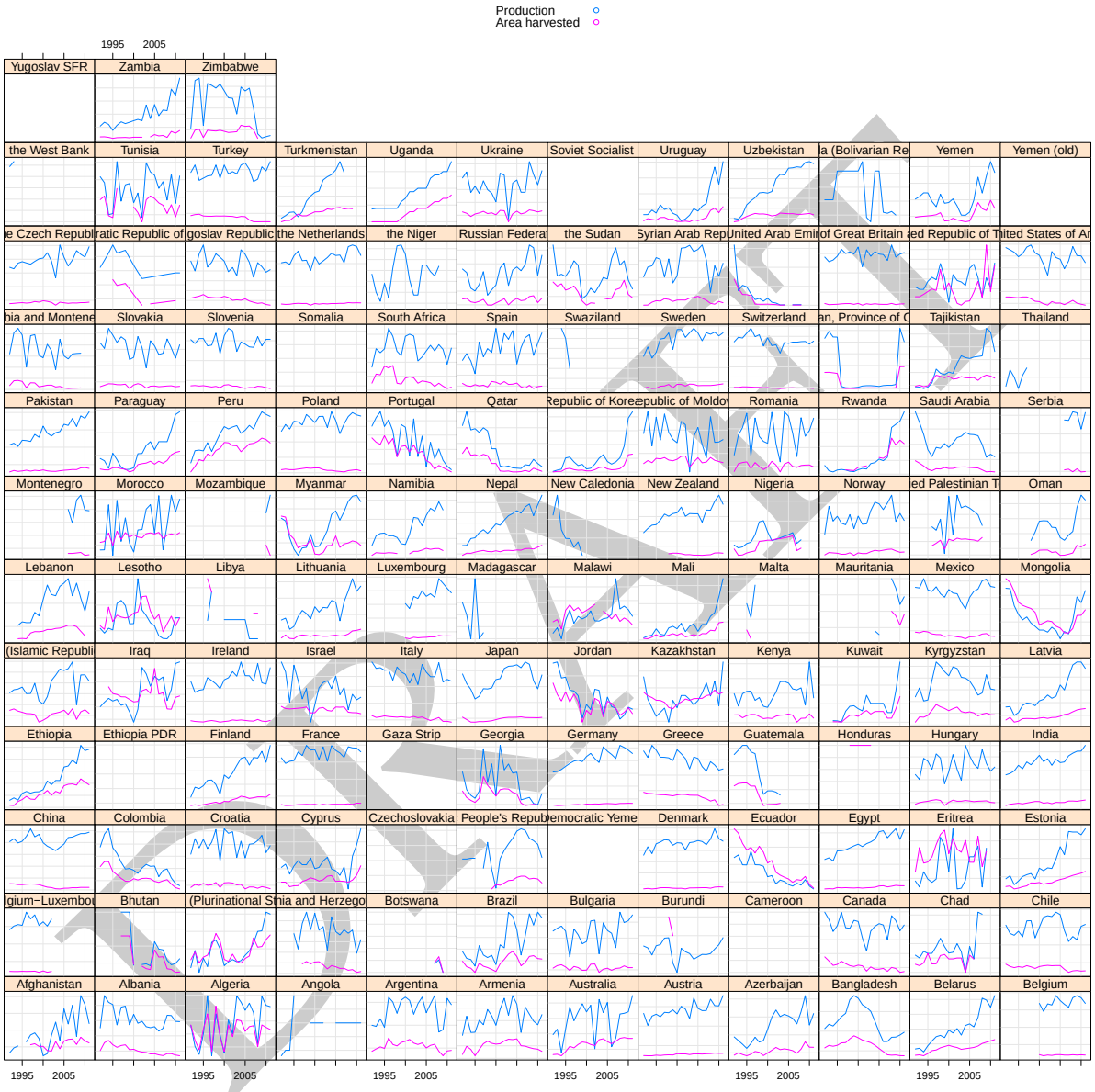


Figure 5: Wheat production and area harvested by country. The figure shows that excluding several producers such as Turkemenistan, Nepal, and Pakistan which has a stable trend in production, both the production and area display erratic behavior.

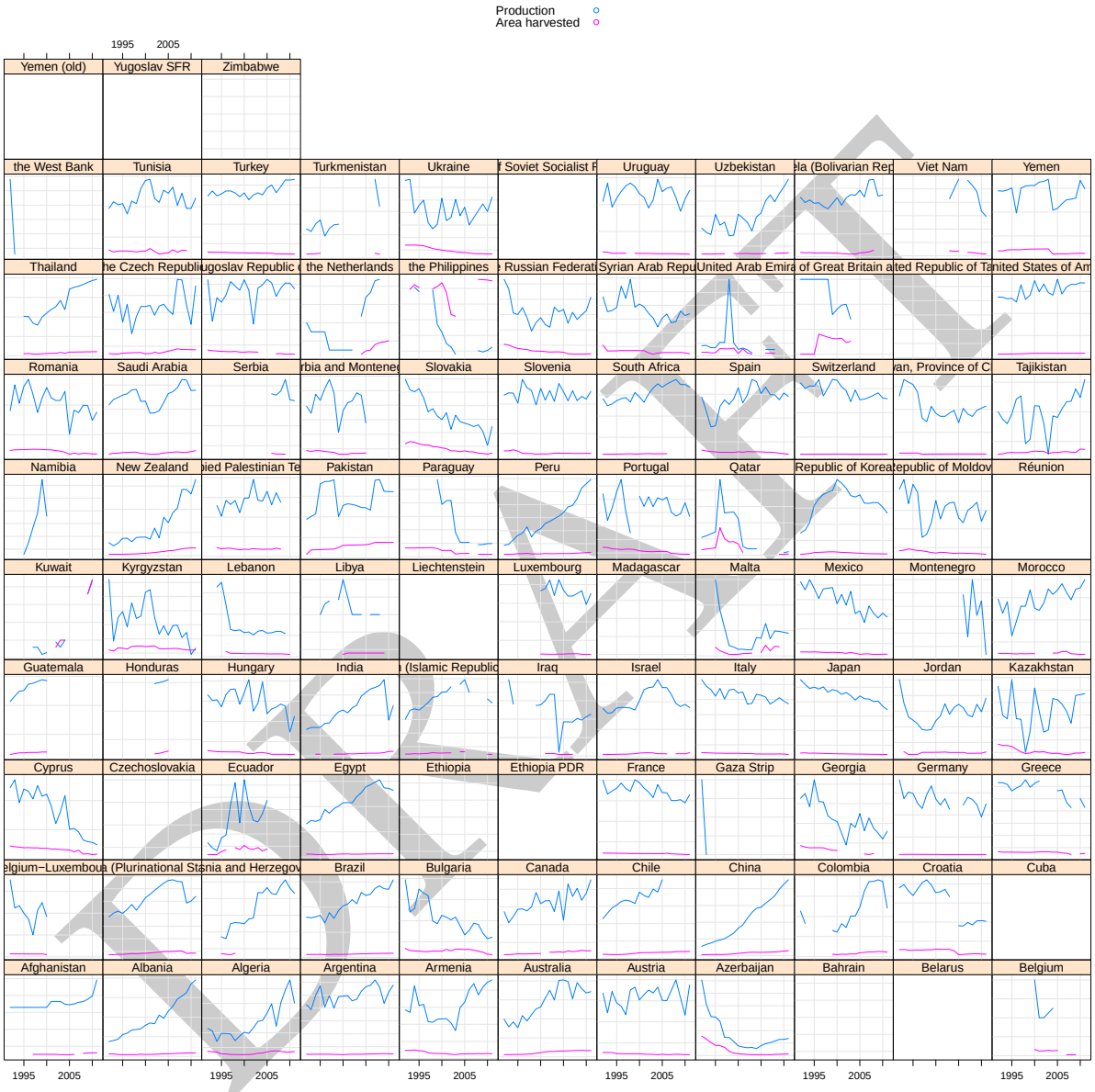


Figure 6: In contrast to wheat, the area for grape is much more stable, a character of tree which takes year to plant and nuture and the alteration of the land use is much more difficult. Nonetheless, the production also display different trends over different time period

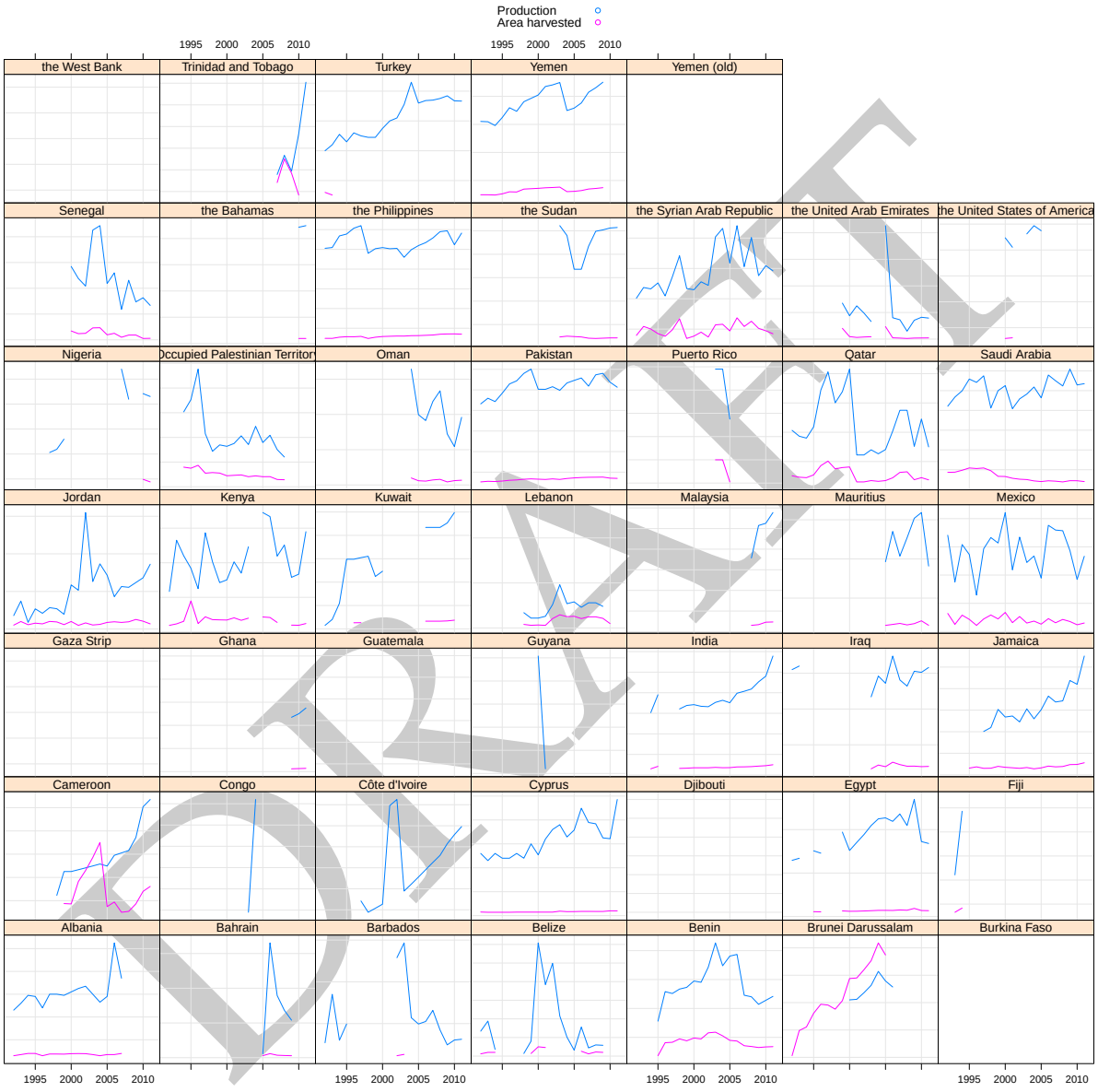


Figure 7: Even more so than both wheat and grape, the production appears to demonstrate unpredictable trends and shocks.

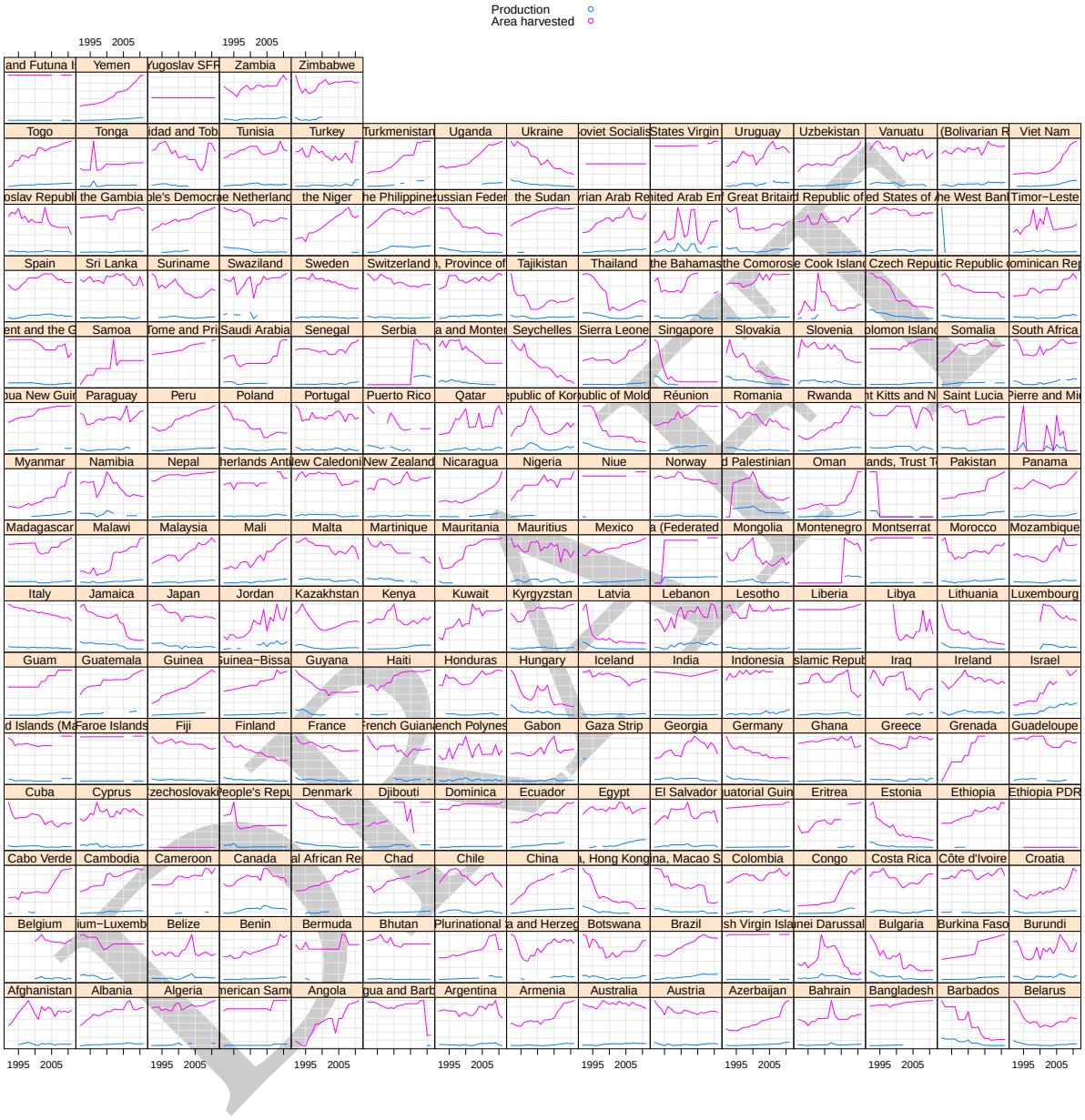


Figure 8: Of all the production, livestock meat such as beef and veal may have been the easiest to predict and impute. There is continuing demand around the world for meat, while shift in production is usually difficult due to the high expenditure in machinery capital. With this being said, we can likewise observe shocks and or period of contraction or expansion over time.

3.3. Data Quality Issues

During the development of the methodology, we have encountered several data quality issues 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.

Extremely High Sparsity

Although missing values are expected provided the goal of the task is to impute missing values, but one may be stunned at the sparsity of the data. For commodity such as pepper, merely 20% of the data are observed which raises the question whether imputation remain valid.

Diverging Trends and shocks

Another issue arose from the quality of the data reported and recorded. It is not uncommon to observe unexplainable diverging trends or shocks of production and area harvested which resulted in exploding yield. The yield of Okra for Bahrain and Senegal in figure 3 are prime examples. Coconut production of China in 2008 is another example, a change in classification resulted in large escalation of production while the area harvested remained similar to the previous year resulted in a three-fold increase in the yield solely for that particular year.

No indicator distinguish zero, missing values and not application

The analyst should be made aware of the fact that although a framework does exist to distinguish zero and missing values in the database, in practice this may not be the case.

These observations prompt us to devise a robust method to safeguard ourself from nonsensical imputation.

4. Proposed Methodology

In this section we will provide a detailed explanation of the proposed methodology with illustration.

4.1. Imputation for Yield

The determination of yield can be categorise into three components, a trend which reflects improvement in technoglogical innovation and fluctuation as a result of climate related factors with adhoc events such as war and diseases causing shocks to the series.

Climate effect can be broadly grouped into two category, year-to-year variation and cyclical-catastrophical. Year-to-year changes in temperature, precipitation will result in small variations around the trend while cyclical-catastrophical phenomenon such El Niño will appear as shocks. Both are difficult to employ in practice since data such as temperature and precipitation suffer problem from period matching; using annual data will reduce correlation and create spurious noises, while monthly data will have to be matched crop by crop by identify the relevant months corresponding to the production cycle.

Shocks such as war or diseases are difficult to model due to the number of events observed in the history is small and the effect of each event may be different.

On the other hand, the continuous advacement of technology and innovation and increasing productivity can be capture by a simple linear trend.

The proposed methodology 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. This allows us to capture not only the technological advancements in a specific commodity, but at the same time the relative speed of technological adoption of the country. 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.

The model

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

$$\begin{aligned}(\mathscr{Y}|\mathscr{B}=\mathbf{b}) &\sim \mathscr{N}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b},\sigma^2\mathbf{I})\\ \mathscr{B} &\sim \mathscr{N}(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\end{aligned} \tag{4}$$

Where the fixed component $\mathbf{X}$ models the effect of exogenous variables when supplied, while the random component of $\mathbf{Zb}$ captures the aggregated effect of innovation and improved practice of each specific country and commodity.

More specifically, the model implemented has the following expression:

$$Y_{i,t} = \overbrace{\beta_0 + \sum_{j=1}^{p}\beta_j x_j}^{\text{Fixed effect}} + \overbrace{b_0 + \sum_{k=1}^{df} b_{k,i}B_k(t)}^{\text{Random effect}} + \epsilon_{i,t} \tag{5}$$

Where Y denotes yield, i for country, t for time, and k the degree of freedom for the B-spline. The fixed effect is left for external drivers such as precipitation and temperature.

To test the number of degree of freedom for spline required to capture the presence of possible non-linearity, the algorithm proceed by estimating the model with one degree of freedom then continue the process by testing the hypothesis whether additional degree of freedom is necessary.

Imputation is predictive in nature, therefore the test is designed to select the model which provides optimum predictive performance. The testing is performed by drawing bootstrapped samples and re-fitting the model to obtain b sets of prediction errors. Then we test whether the newly proposed model has a lower out-of-sample prediction error when compared to the benchmark model. The algorithm will test iteratively until the proposed model has higher prediction error or when the maximum degree of freedom is reached.

$$\begin{aligned}H_0 &: d \leq 0,\\ H_1 &: d > 0.\end{aligned}$$

Where d is defined as follow:

$$d = \sum_{b=1}^{B} \left(\frac{1}{N_1} \sum_{i=1}^{N_1} (\hat{Y}_{b,df} - Y)^2 - \frac{1}{N_2} \sum_{i=1}^{N_2} (\hat{Y}_{b,df+1} - Y)^2 \right)$$

The B-splines implemented is first degree linear, from the analysis we believe that higher order polynomial are not required. The non-linearity can already be captured by having multiple degree of freedom, increasing the order would only over-fit the data and can display erratic imputation in the tails.

In essense, the imputation of the yield is based on the overall country specific improvements in innovation while also capturing the relative advancements in the designated commodity, both accross time. Since more information are utilized and pooled together, imputed values display stable characteristics while reflecting changes resulting from different aspect.

4.2. Imputation for Production

From the exploratory analysis, we can see that the trend and shape of production and area harvested are closely related. Nonetheless, both the series and missing mechanism can behave very different depending on the country and the commodity. Furthermore, there does not seem to be any dependable information for us to model.

This is partly inherint in the nature of the size of the data which compose vast variation of commodity and countries while at the same time the relationship maybe too complex to be expressed in a suncinct way. This is in contrast to the nature of yield where cross-country and cross-commodity information can be pooled together for a better informed imputation.

Land can be expanded or contracted between different commodities depending on the price and expectation of the market, single block of land can also be used for multiple commodity. Both market forces and personal decisions are relatively unpredicatable which resulted in the difficulty of imputation of area harvested and production.

This motivates us to employ what is known as ensemble learning to impute the production and to deal with the multi-hypothesis problem.

Given the strong correlation between area harvested and production, we have decided to impute production and leave area harvested to balance. This is based on expert advice that the production data are often much more reliable and comprehensive.

The algorithm

Ensemble learning in its simplest sense, is to build a collection of simpler base models or learners which are later combine to obtain the composite model or prediction. One of the most famous application was the prediction of movie rating competition held by Netflix, which the top two performer both used a form of ensemble.

The method consist of two steps:

1. Building multiple models/learners.

2. Combine the models or predictions.

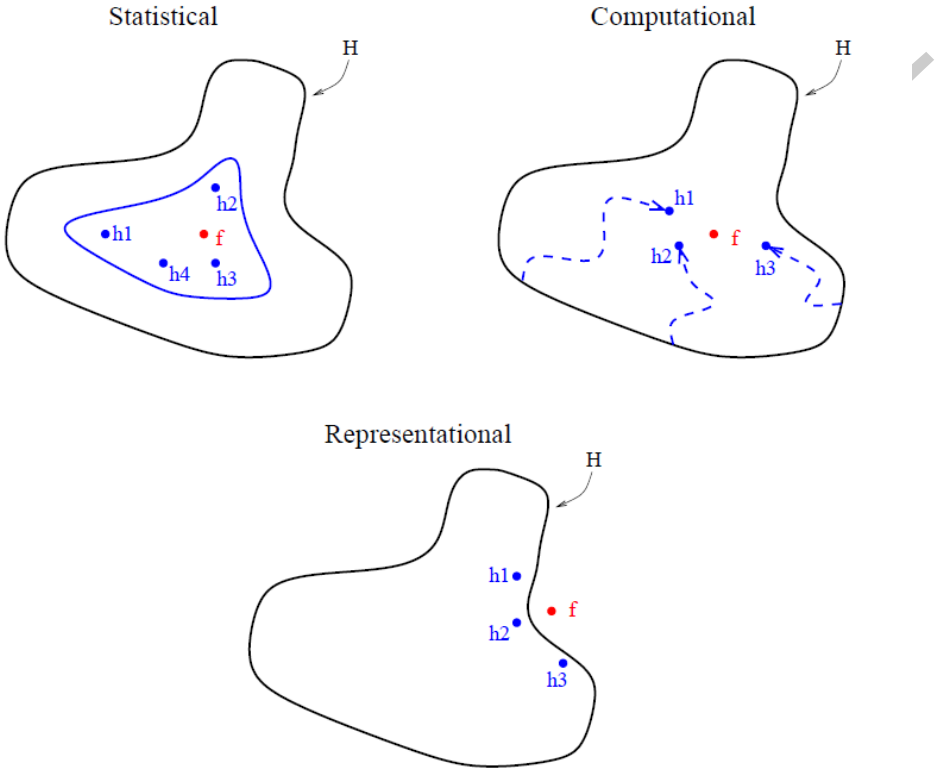
The ensemble method reduces the risk of choosing a poor model, this resembles putting the eggs in different baskets. Thus we reduce the risk of implementing a single model which may produce nonsensical imputation for a certain subset of data.

Ensemble as described by Dietterich can mitigate the following three issues.

Statistical: Lack of data to identify an unique solution.

Computational: Optimization

Representational: Complex model



The statistical problem refers to the lack of data to support a particular hypothesis. The problem can be formulated as finding the best hypothesis among competing models in the space $\mathcal{H}$. In the top left graph of the following depicion from Dietterich we can see that hypothesis within the blue boundary will give the same fit, and there is insufficient information to determine which one is better. By combining the models, we may reduce the risk of choosing a terrible model. If we only observe two data points for a country, then fitting a linear line or a log curve can both give the same fit and we may have no information to distinguish the two.

The second problem which is lesser of an issue for our task is the computational problem which applies to model which employs greedy algorithm such as step-wise regression and classification tree. The search within the hypothesis space at each step is local and thus has a high probability of failing to achieve global maximum.

The final problem, representational, refers to the fact that the true function f can not be represented by any of the model. However, by combining the models we may expand the space of representable functions and potentially approximate the true function f if it exist. Take the production of wheat in Uzebekistan and Rwanda for example, if the production of a country has been growing at a linear rate for the past four decades, but expands rapidly in the last two decade, neither linear or exponential model will provide a satisfactory result. However, an ensemble combining a linear and exponential model will provide a good solution by capturing different characteristics of the data.

From an integration point of view, the algorithm is adaptive in the sense that if the data generating mechanism changes in the future, the method will shift weights to models which better present the data and thus reducing the need of constant monitoring and updating of methodologies.

The model can also be seen as a mixture of different expectations. From prior discussion, we know that the area and production depends on the perception and cognition of information leading to optimal judgement. However, even the same information can lead to various decision as the information can be interpreted differently and the model attempts to capture the difference in the judgement by combining various expectation.

The details of the ensemble implemented is describe here, the base learner are listed in increaseing order of complexity.

An effective ensemble will have base models as diverse as possible. If there are no diversity and all model generates similar result, then a correction and the variance reduction property of the ensemble model will be poor.

- **Base learners:**

Mean: Mean of all observations

Linear: Linear Regression

Exponential: Exponential function

Logistic: Logistic function

Naive: Linear interpolation followed by last observation carried forward and backward.

ARIMA: Autoregressive Integrated Moving Average model selected based on the AICC, and imputation via Kalman Filter.

LOESS: Local regression with first degree local polynomial and sample size variant window.

Splines: Cubic spline interpolation.

MARS: Multivariate Adaptive Regression Spline

- **Combiner:** non-trainable algebraic combiner - Weighted sum rule

$$p_n(x) = \sum_{i=1}^{K} w_i f_{n,i}(x)$$

$$w_i = \frac{\sum_{i=1}^{N} (1/|f_{n,i}(x) - x|)^2}{\sum_K \sum_{i=1}^{n} (1/|f_{n,i}(x) - x|)^2}$$

Where the weights (w_i) of model i depends on its fit ($f_{n,i}(x)$) on the available data with a ceiling set at 70%. The naive imputation which does not have a fitted model always take uniform weights. A much more favourable method is to compute weights based on the prediction error obtained from bootstrap or jackknife, yet the missing values restrict us to draw any reasonable size for sub-sample fitting. This is further penalized by the time series nature of the data where boostrap with replacement is not possible.

In this section, we take a sample of production from the exploratory analysis section and illustrate the imputation methodology. The selection of the sample is intended to illustrate the flexibility and robustness of the methodology, rather than based on the importance nor the quantity produced.

The black points in the graph represents observed data while the thicker blue line represents the final ensemble model with the points being the estimation of missing values. Other lines represents the fit of the base learner, the weight of each model is displayed in the legend.

Component models which failed to fit the data has a weight of 0.

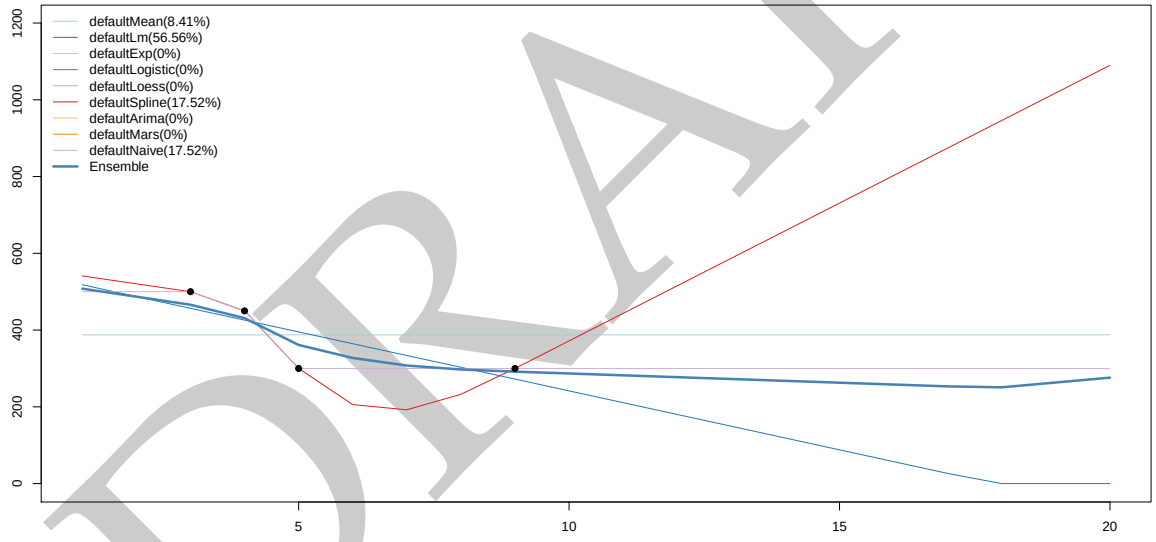


Figure 9: This selected series is the production of wheat in Swaziland, from the black observed production we can see that no data has been observed in the paste 10 years. Yet, the ensemble fits several possible extrapolation which all seem reasonable. Yet, we can see that although the linear regression obtained the highest weight the ensemble is safe guarded by naive interpolation and spline so it does not go below zero.

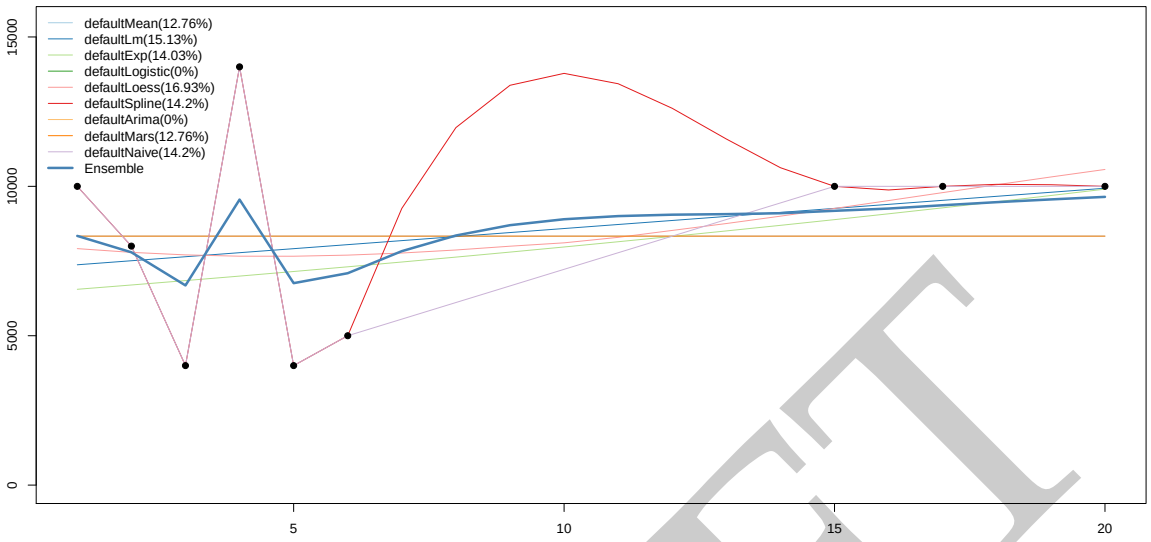


Figure 10: In contrast to swaziland, the wheat production in Mauritania exhibits a simple shape. Almost equal weights were allocated to model which did not fail.

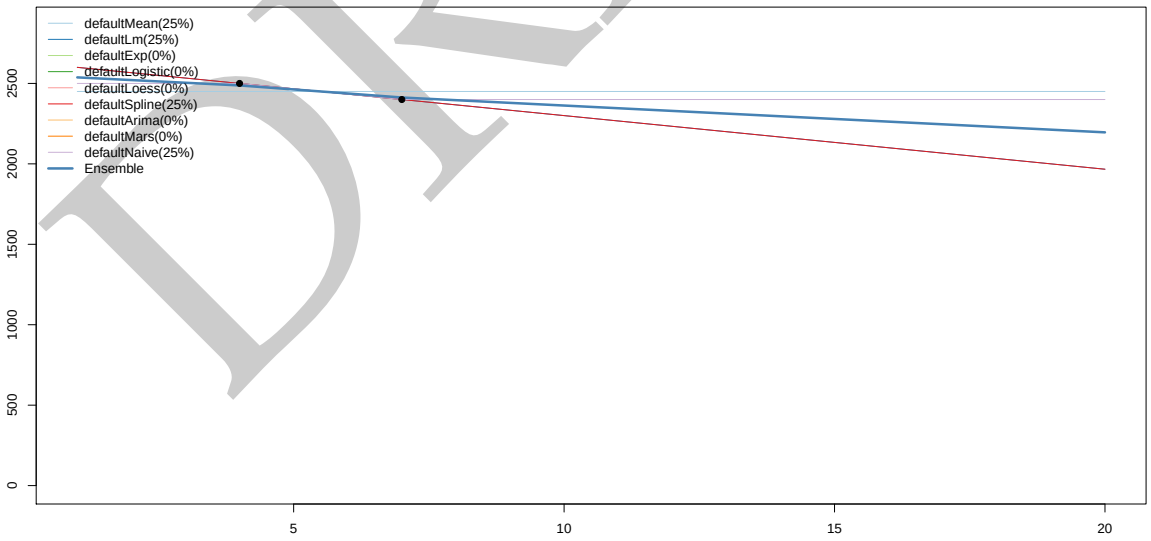


Figure 11: In this depiction, the grape of Zimbabwe is used for illustration. In the case where there is not sufficient amount of data, the ensemble will collapse to form a simple model with strong agreements between the models.

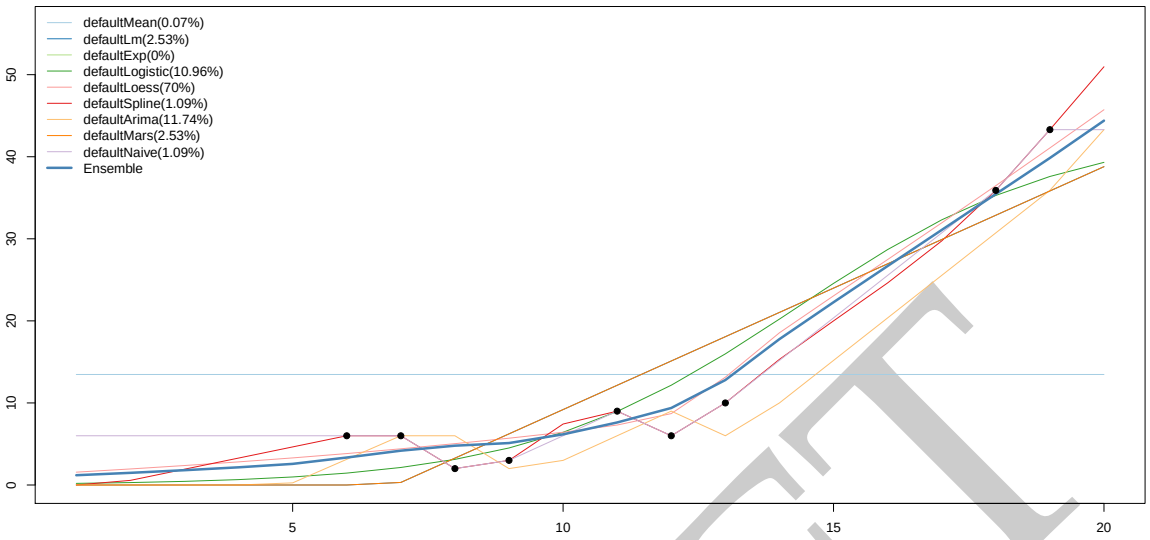


Figure 12: This time series for the grape produciton in Kuwait also displays simple exponential like growth. Here, the smooth LOESS model obtained the highest weight and in fact reached the ceiling of 70 percent. The exponential failed because the default model requires that at least one point need to be observed at the first and the last 5 time point. This is to prevent the exponential model imputing values which are not supported by the data.

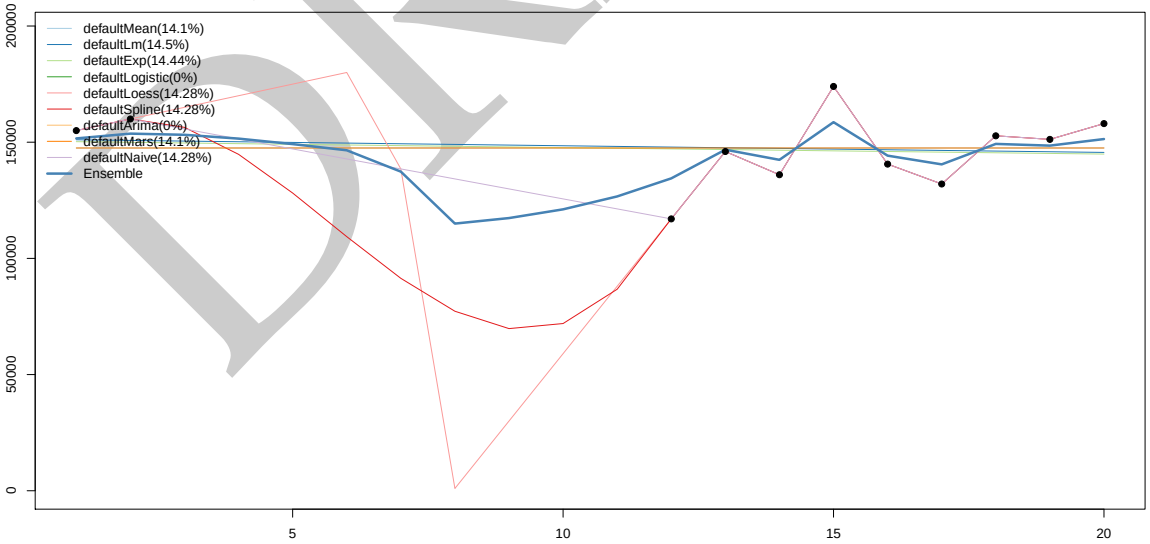


Figure 13: There appears to be a gap in the series of the production of Okra for Iraq. Most model collapse to a straight line close to the mean of the series, but both spline and loess captured the possible downward trend which is reflected in the dent in the final ensemble.

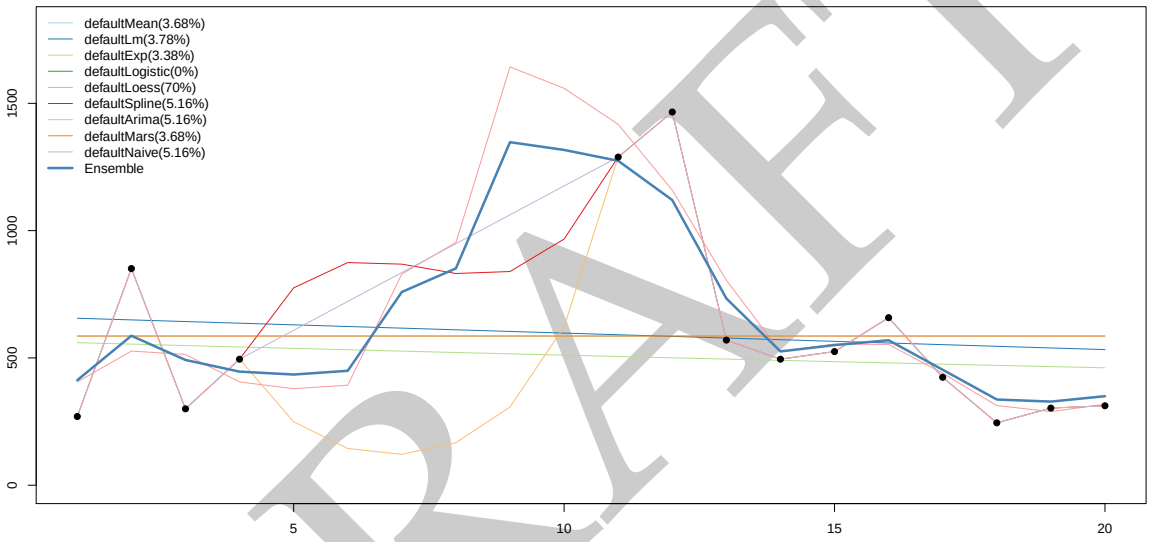


Figure 14: Finally, a more complex example is shown here, the series is the production of Okra in Barbados. There were a few observation in the beginning of the series, then in the center of the time series the production doubled before falling to prior level. A relative complex model is required to fit such a time series, loess took most of the weight with the remaining weight assigned to other model.

The examples demonstrates that the model is flexible, able to capture from the simplest linear trend to more complex behaviours without the need for model selection. Further, with carefully selected and fine tuned component models, the ensemble will exhibit extremely robust characteristics.

4.3. Imputation for Area Harvested

After the imputation of production and yield, area harvested is left to balance to satisfiy the identity equation.

4.4. Use of external variables

The proposed methodology does not employ any other variables for several reasons. First, the external variables may also contain missing value and often they may actually in fact be more sparse than the dataset we are trying to impute. Secondly, the selection of variables may be difficult. Difference in commodity nature, market structure, commercialization and other conditions calls for different information set for prediction. Finally, even if a set of data can be selected, the use of external variables will require large resources to maintain and specific design set up to be used under various imputation setting.

5. Case Studies

The imputation of the three commodities are presented in this section. Due to the limitation that this is a dynamic generated paper we have only include the cross-country information for the imputation of the yield.

5.1. Wheat

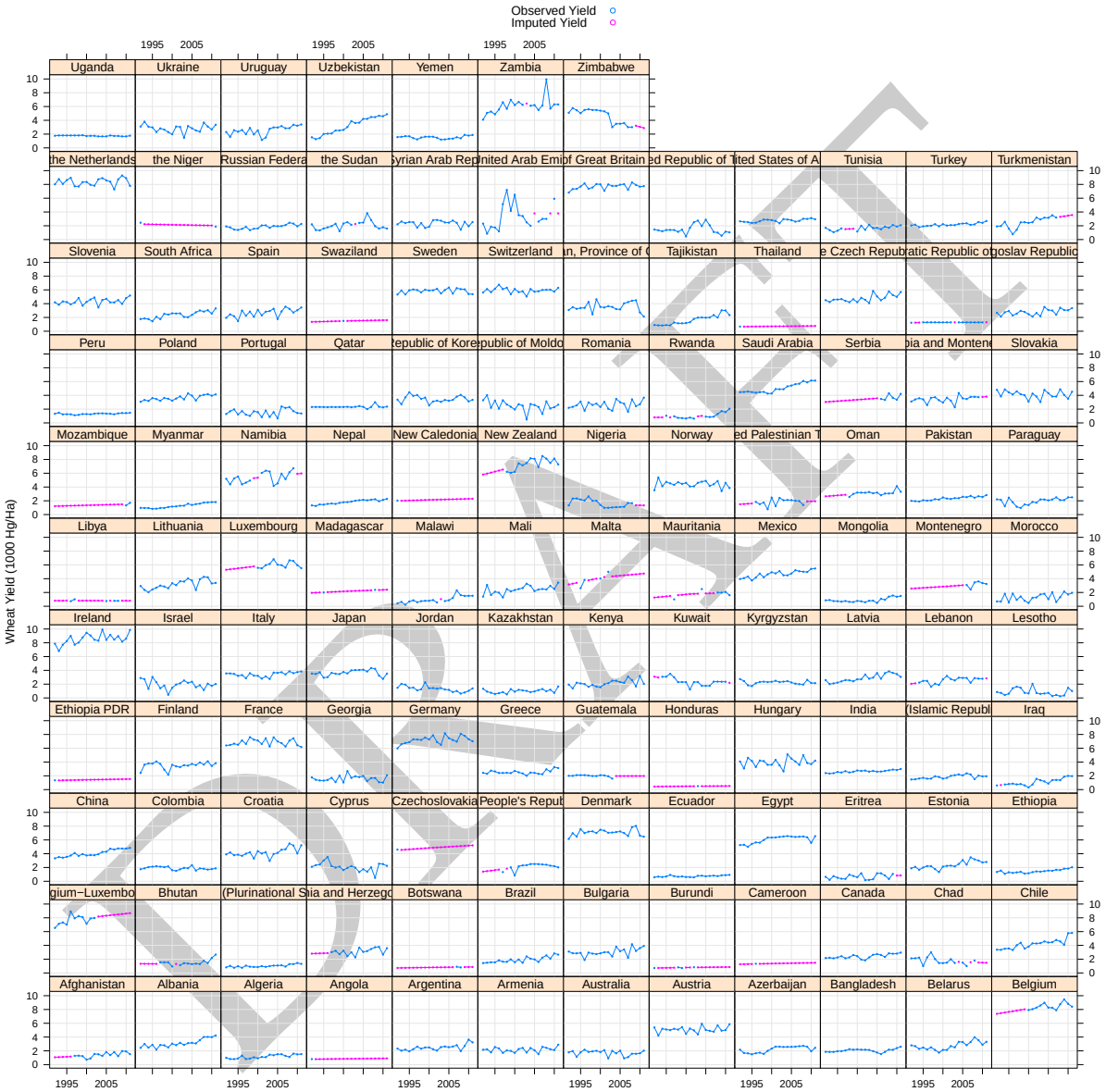


Figure 15: We can see the imputation of yield in purple seems to produce reasonable values.

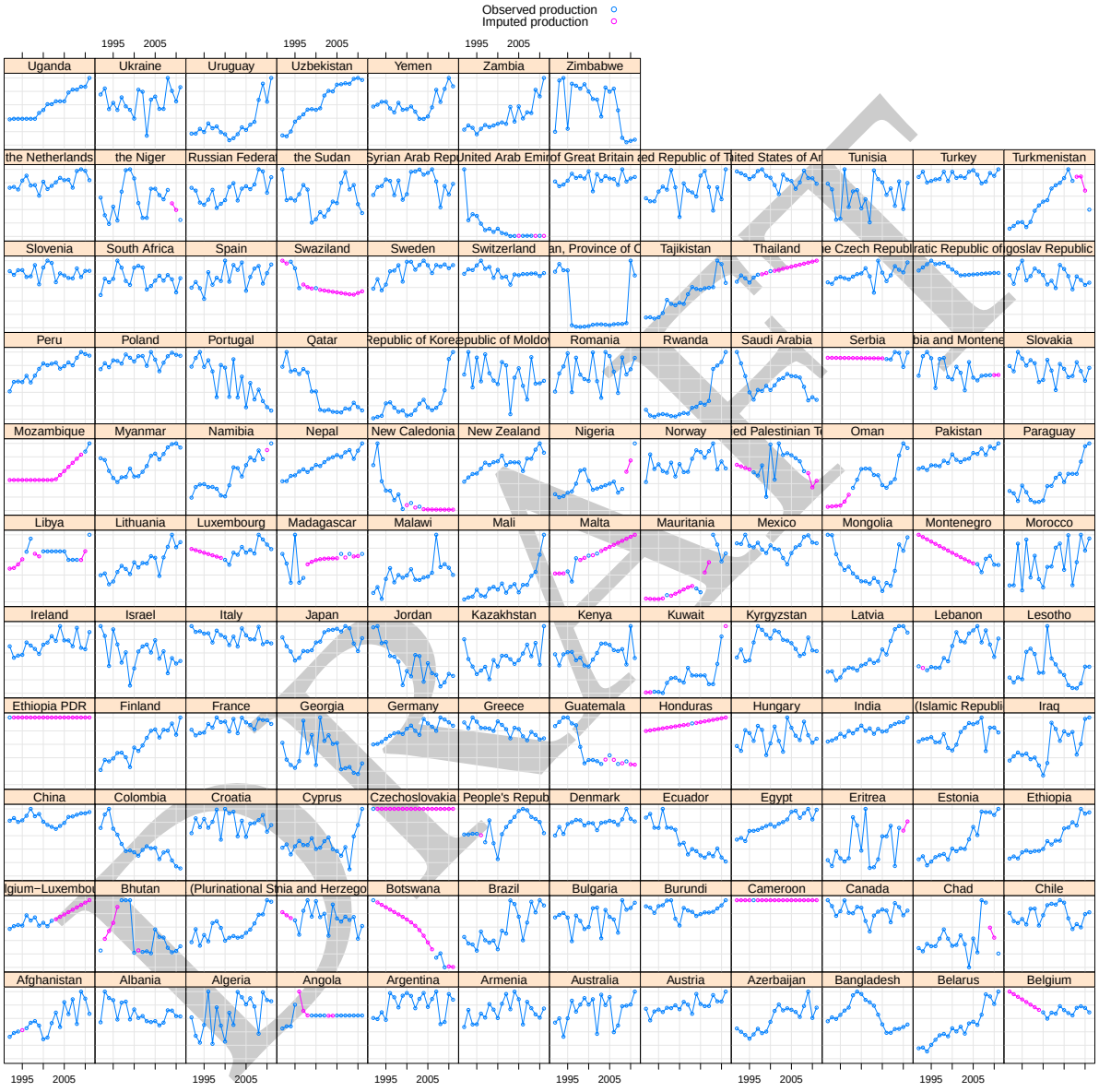


Figure 16: There does not appear to be any pecularly in any of the imputation of the production of wheat.

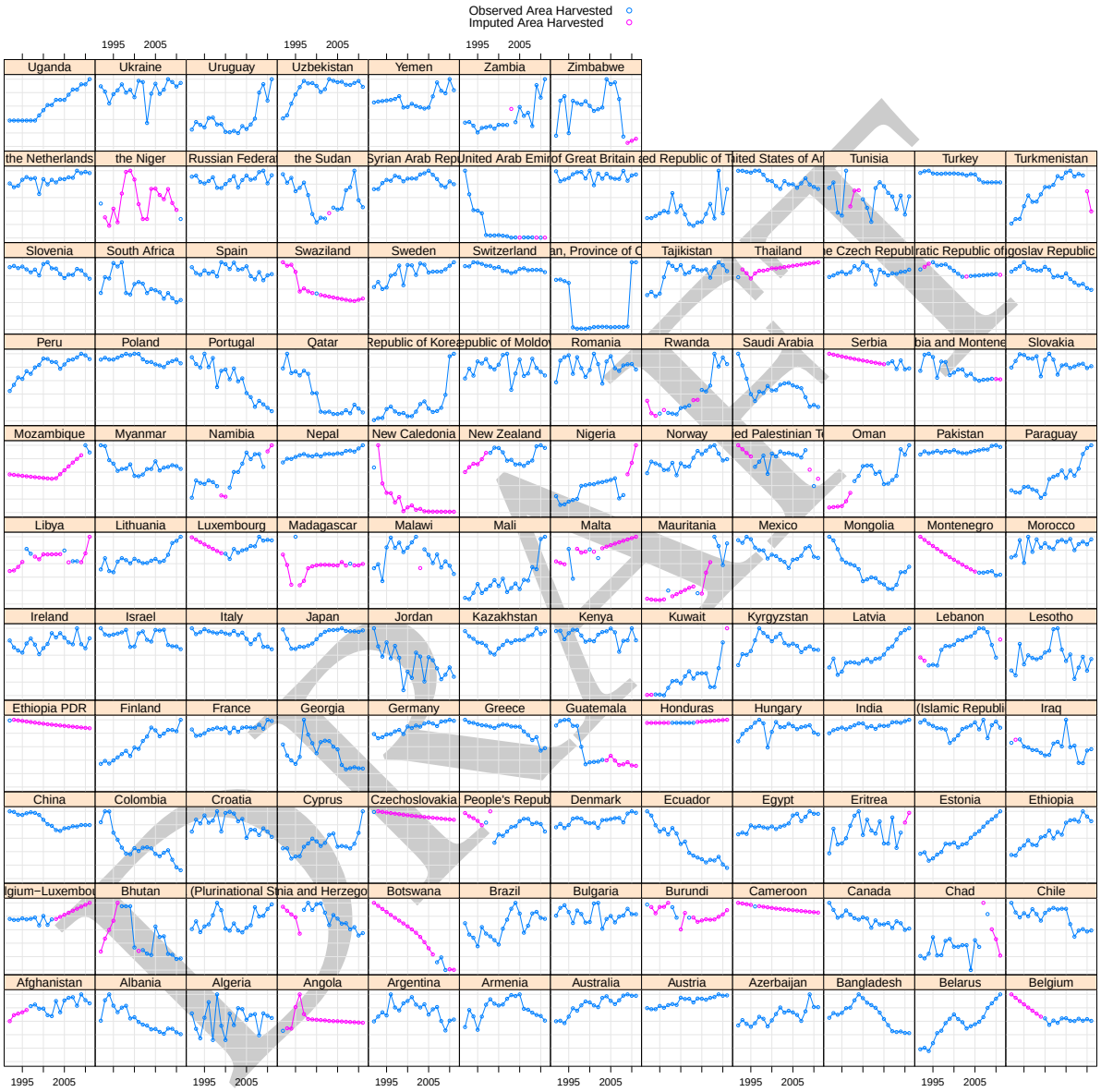


Figure 17: The balance of area harvested also does not show any sign of divergence or problematic symptoms.

5.2. Grape

Figure 18: The imputation of yield also appears to be reasonable, even in the case for Iraq the linear mixed model is not severely influenced by the one off event.

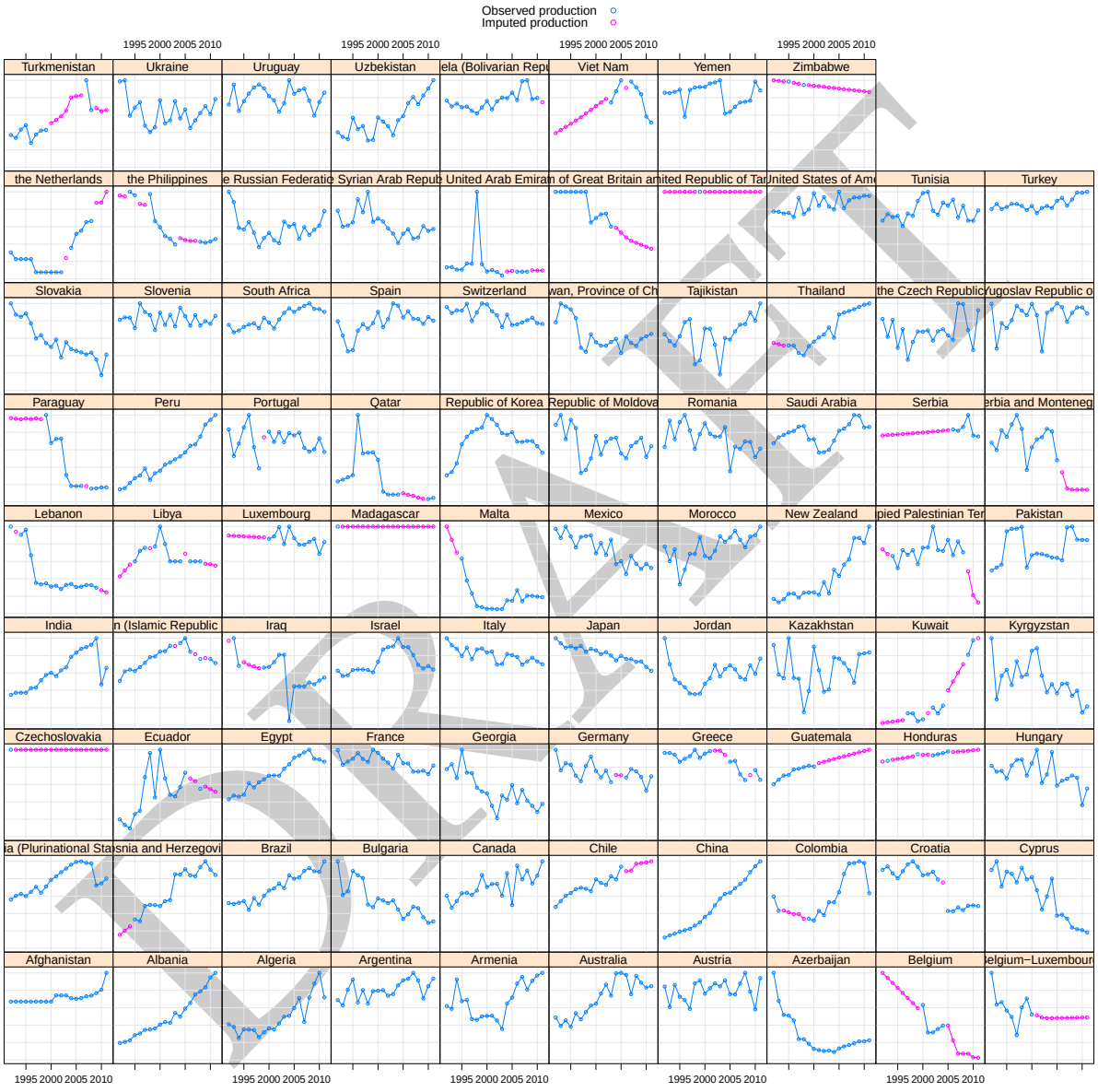


Figure 19: The imputation of grape prodution.

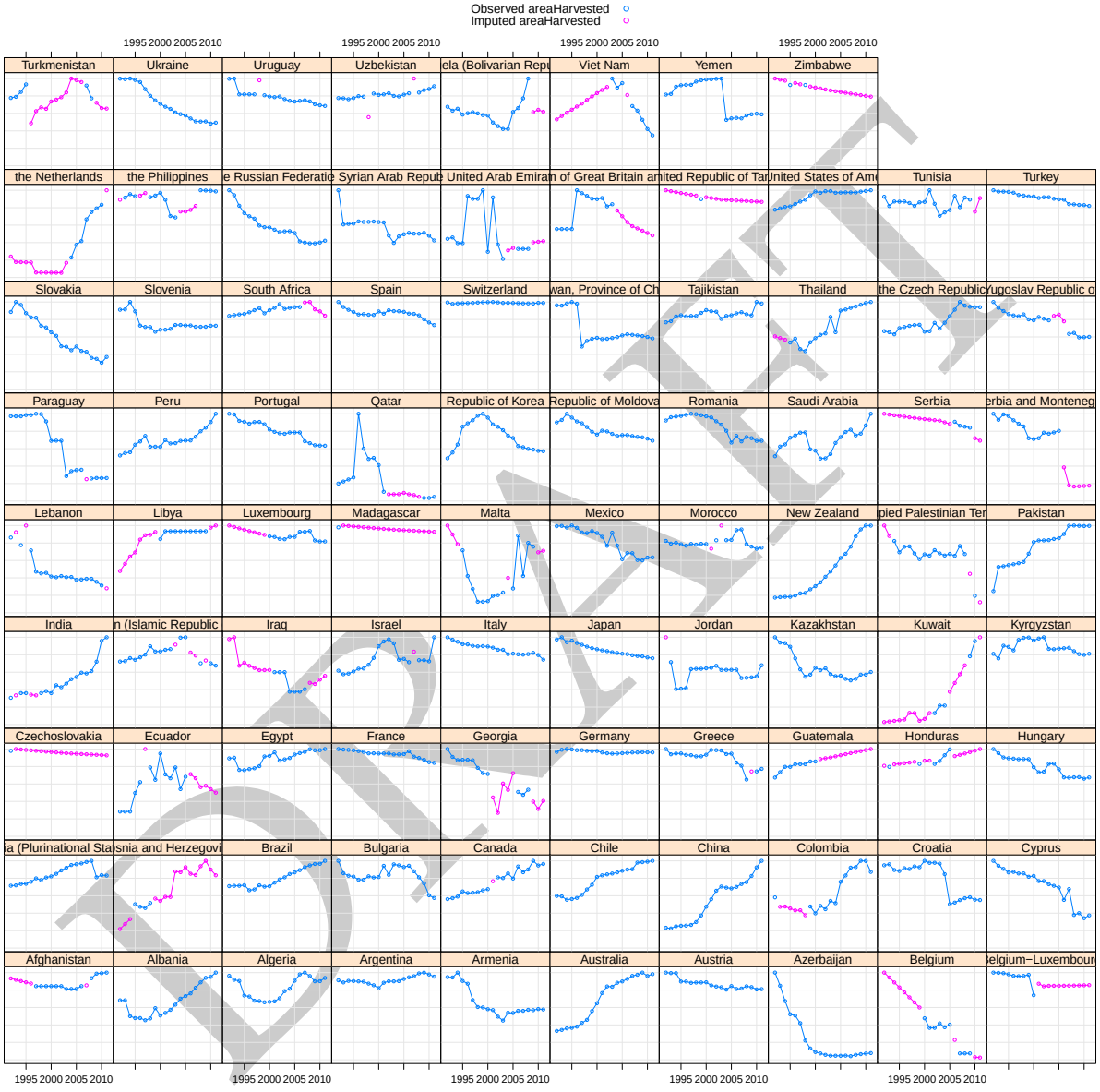


Figure 20: Imputation of area harvested of grape. The area harvested does appear to be problematic, mostly resembling the trend of the production.

5.3. Okra

Figure 21: The imputation of yield also appears to be reasonable, here once again we see the robustness of linear mixed model not being influenced by the bad data quality of both Senegal and Bahrain.

Figure 22: The imputation of okra prodution.

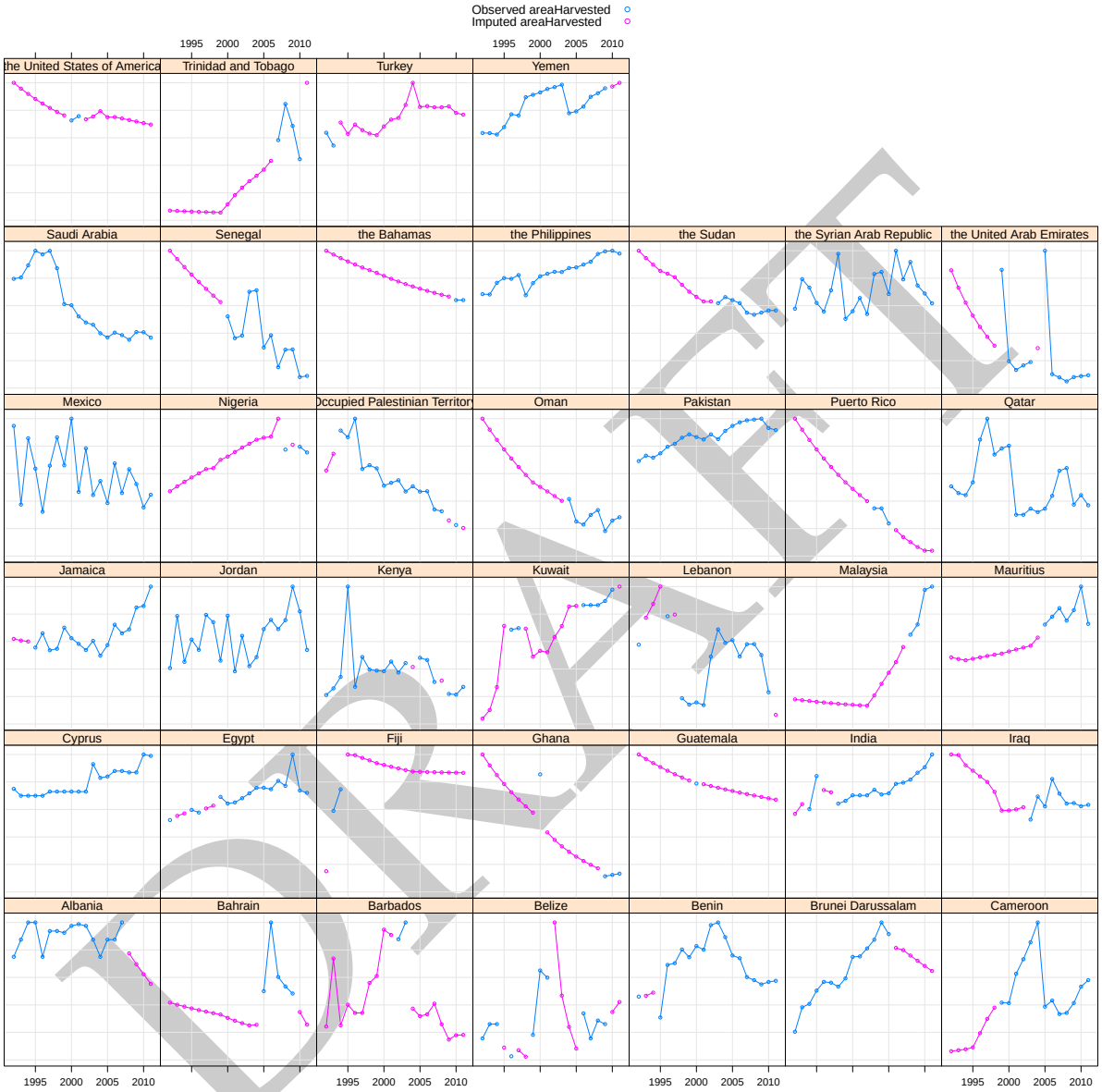


Figure 23: Imputation of area harvested of okra. The area harvested of Viet Nam does appear to be problematic, the unexpectantly high area harvested is a result of an extremely low yield in the earlier years. This is due to the fact that long extrapolation for over 40 years with only approximately 10 years of data is unreasonable. We can eliminate this problem if we restrict the working data to the past 3 decades.

5.4. Beef

Figure 24: Imputation of beef carcass weight. .

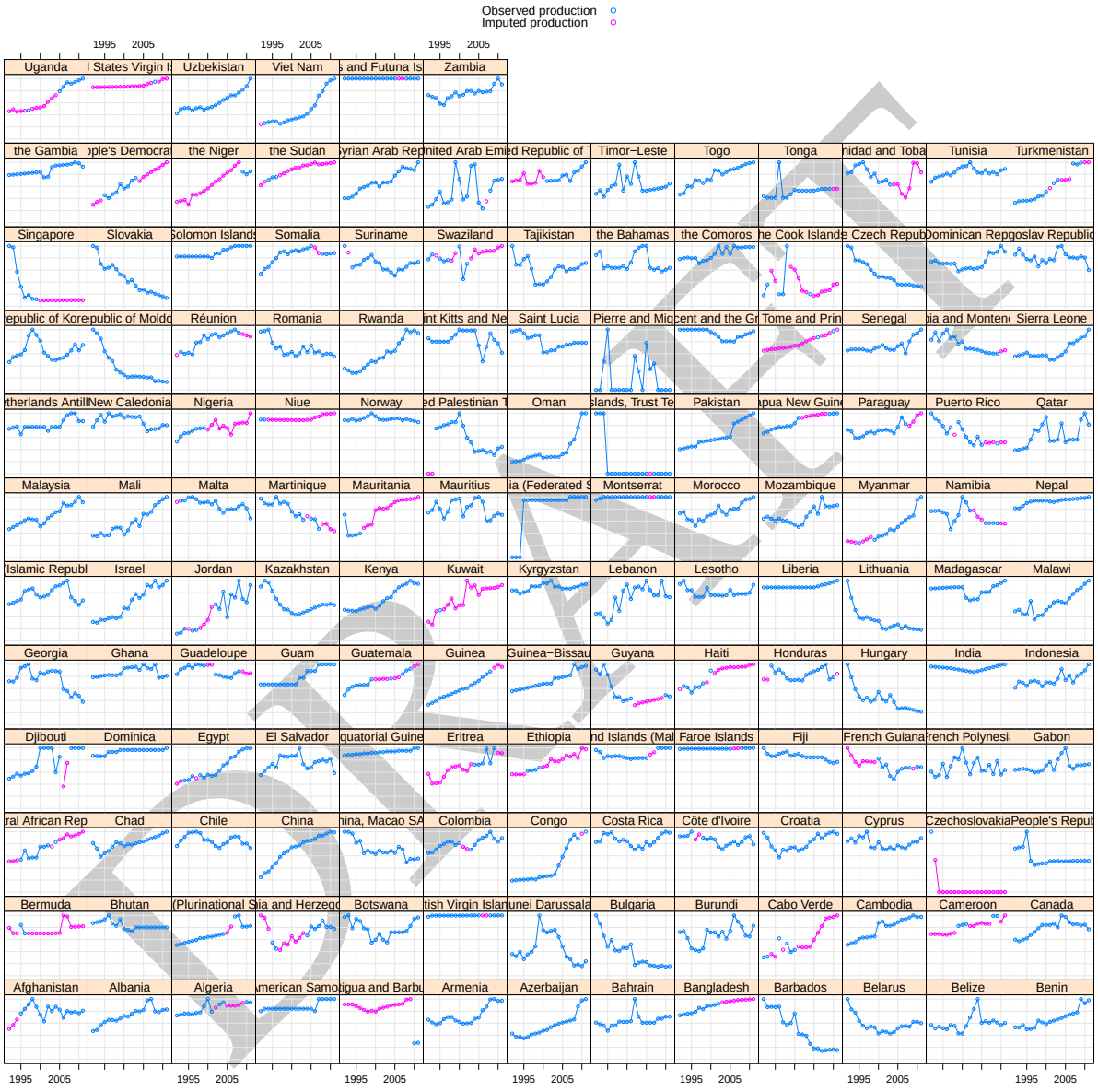


Figure 25: Imputation of Beef production.

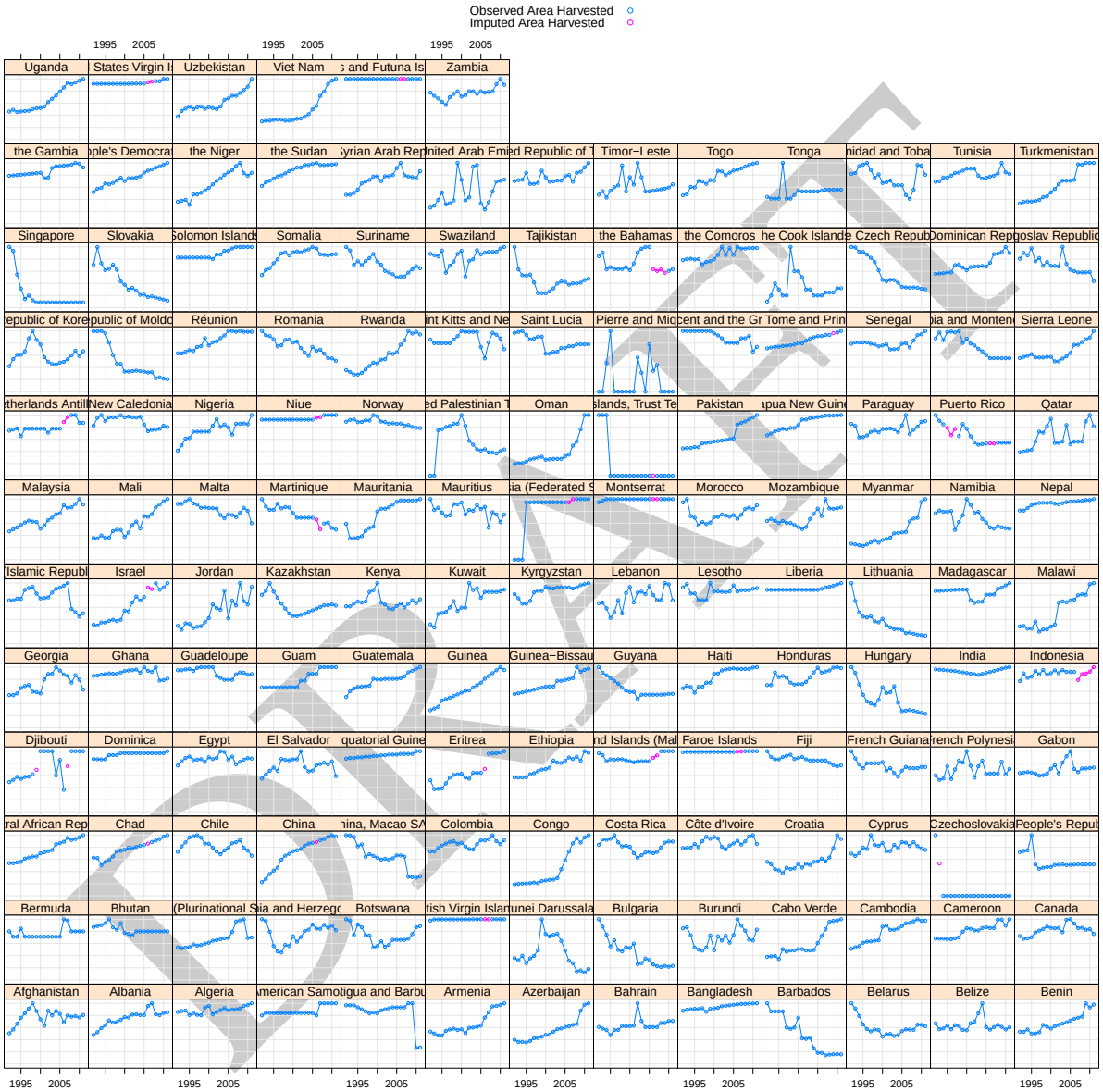


Figure 26: Imputation of number of animals slaughtered.

From the case studies, we can see that we have an extremely flexible model with almost no possibility of over-fitting since none of the model are itself complex. The model is robust as a result of the variance reduction property of ensemble and even in cases where we have extremely sparse data we can still fall back to simpler models with fewer assumption.

The model is not only flexible in adapting to change in the data generating mechanism, but at the same time flexible to accommodate further models in which we consider appropriate. Any model which offers diversity and additional predictive power can be included to further enhance the ensemble.

6. Simulation Styudy

In order to understand the performance and characteristics of the imputation, we have conducted simulation to estimate the prediction error of the imputation.

For each bootstrap, we take a sample containing only official and semi-official data and impute the values which are missing. 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.

$$\text{MAPE} = \frac{1}{N} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{6}$$

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.

Here is the table for the result of the simulation, 50 boot strap samples were drawn for each scenario. The final percentage error is calculated as the average of all the bootstrapped error within the group. The missing percentage for grapes and beef were higher than 20% already and thus a simulation was not possible for the group of 20% missing proportion.

Table 2: Simulation result for the imputation methodology.

Commodity	Total Missing Proportion	Effective Missing Proportion	MAPE (%)
Wheat	30.9%	20%	3.68
	48.1%	40%	7.84
	65.4%	60%	13.01
	82.7%	80%	25.19
Grapes	39.5%	20%	4.49
	54.6%	40%	9.31
	69.7%	60%	14.63
	84.9%	80%	21.69
Beef	57.4%	20%	1.04
	68.1%	40%	2.32
	78.7%	60%	4.15
	89.4%	80%	7.67

7. Conclusion and Further Improvements

This paper demonstrated the flexibility and robustness of the imputation methodology. The robust methodology is capable of capturing cross-country and cross-commodity improvements in the producticity, while providing flexibility to extend the models wherever possible.

An ongoing project is dedicated to the designation of weights reflecting the information content of the observations. Data collected from official and semi-official sources may be deems as more reliable while observation from expert judgement may be less reliable. A weighting scheme will allow the models to better assess the quality of information and improve the performance.

Further, a pilot experienment for utilizing remote sensing data is tested. The potentials remain unknown, but it may provide a better estimation of the area sown and harvested.

Acknowledgement

This work is supervised by Adam Prakash with assistance from Josef Schumidhuber, 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 previous discussions providing valuable feedbacks.

Annex 1: Supplementary Resources

The data, source code and documentation can all be found and downloaded from https://github.com/mkao006/sws_imputation, the package can also be installed by following the instruction.

Annex 2: Pseudo Codes

Algorithm 1: Imputation Procedure - function *swsProductionImputation*

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 yield algorithm ;
    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 production ($P_{i,t}$) with ensemble;
    forall the *imputed production* $\hat{P}_{i,t}$ **do**
      if $\hat{Y}_{i,t} \neq \emptyset$ **then**
        $\hat{A}_{i,t} \leftarrow \hat{P}_{i,t} / \hat{Y}_{i,t}$;
      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 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, *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