Statistical Working Paper on Imputation Methodology for the FAOSTAT Production Domain

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

Abstract

This paper proposes a new imputation method for the FAOSTAT domains based on linear mixed model and ensemble learning.

The proposal provides a resolution to many of the shortcomings of the current approach, and offers a flexible and robust framework to incorporate further information to improve performance.

A detailed account of the methodologies is provided. The linear mixed model demonstrates an ability to capture cross-country and cross-commodity information; meanwhile, the ensemble learning approach provides flexiblity and robustness where traditional imputation methods of applying a single model may have failed.

Keywords: Imputation, Linear Mixed Model, 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(s) 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 than as a replacement for data collection. Imputation itself does not create information; it merely create observations based on assumptions.

This paper is dynamically generated on April 6, 2015 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 imputation.

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, estimated and previously imputed data are marked as either **E** or **I** and 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
I	Imputed figure
T	Extrapolated/interpolated
M	Not reported by country
E	Expert sources from FAO (including other divisions)

2. 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. The series commence in 1992 and continues to 2011.

2.1. Yield

The next three graph depicts the yield of four selected commodity from different commodity group, wheat, grape, okra 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 restriction imposed 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.

Forces of nature also play a vital role in the determination of crop productivity. However, unlike technological advancement, precipitation and temperature are associated more closely to year-to-year variation.

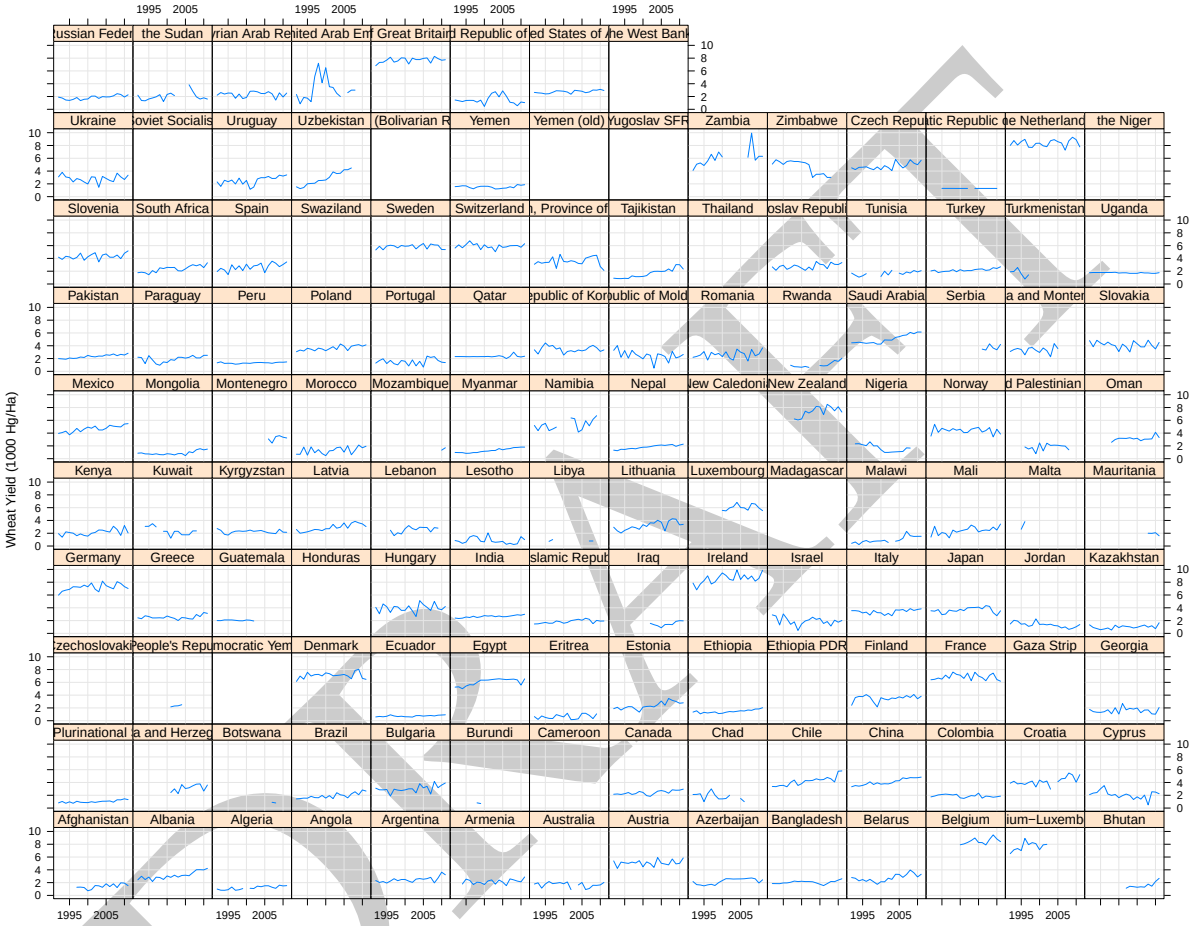


Figure 1: This figure illustrates the yield of wheat accross all countries, it provide strong support to the facts stated. First of all, we can observe the concordant increasing trend accross all country where technological innovation such as improved seed, and synthetic nitrogen fertilizer contributed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

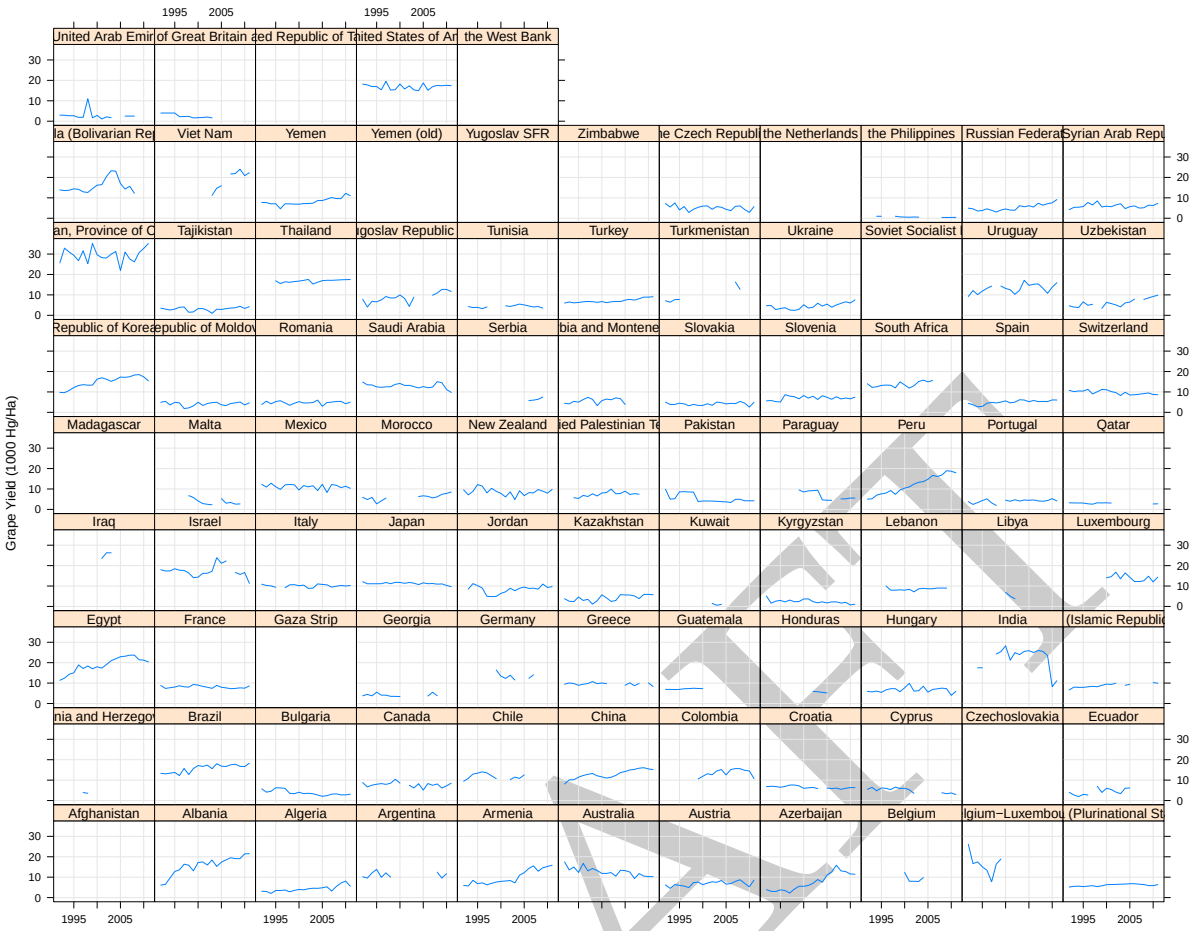


Figure 2: Unlike the yield of wheat, the yield for grape has remain rather constant over time except a few selective country such as Peru and Azerbaijian. There are a few spikes observed, namely Iraq, the invasion of Iraq may have contributed to the negative shock. The considerable fall in the yield for India is of unknown cause, and potentially a data entry error.

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.

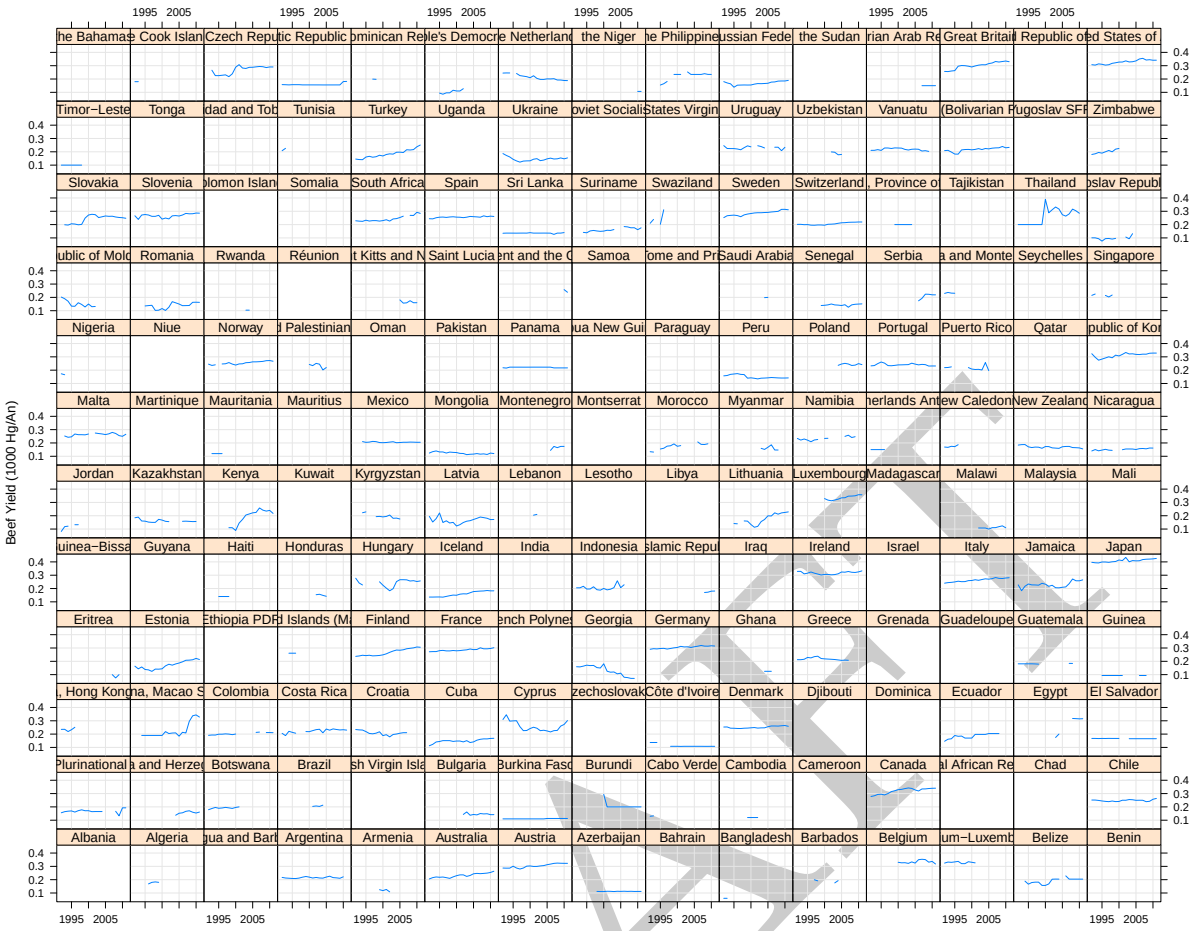


Figure 4: Similar to grape, the carcass weight per animal of beef also shows the same pattern of a very modest improvement in yield with a handful of exceptions.

2.2. 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 under curren state of understanding.

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.

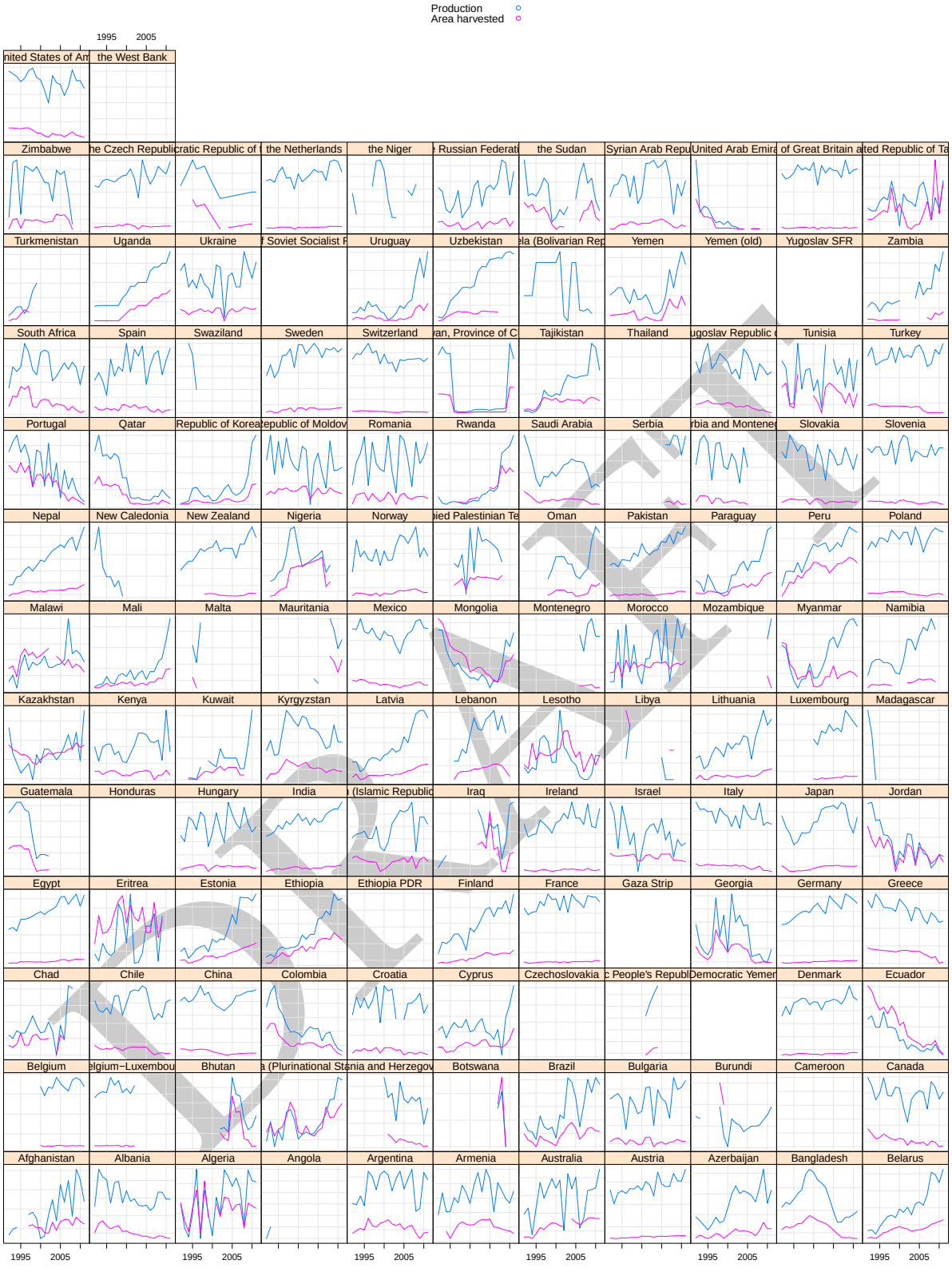


Figure 5: Wheat production and area harvested by country. The figure shows that excluding several producers such as India, Nepal, and Pakistan which has a stable trend in production, both the production and area of most countries display erratic behavior.

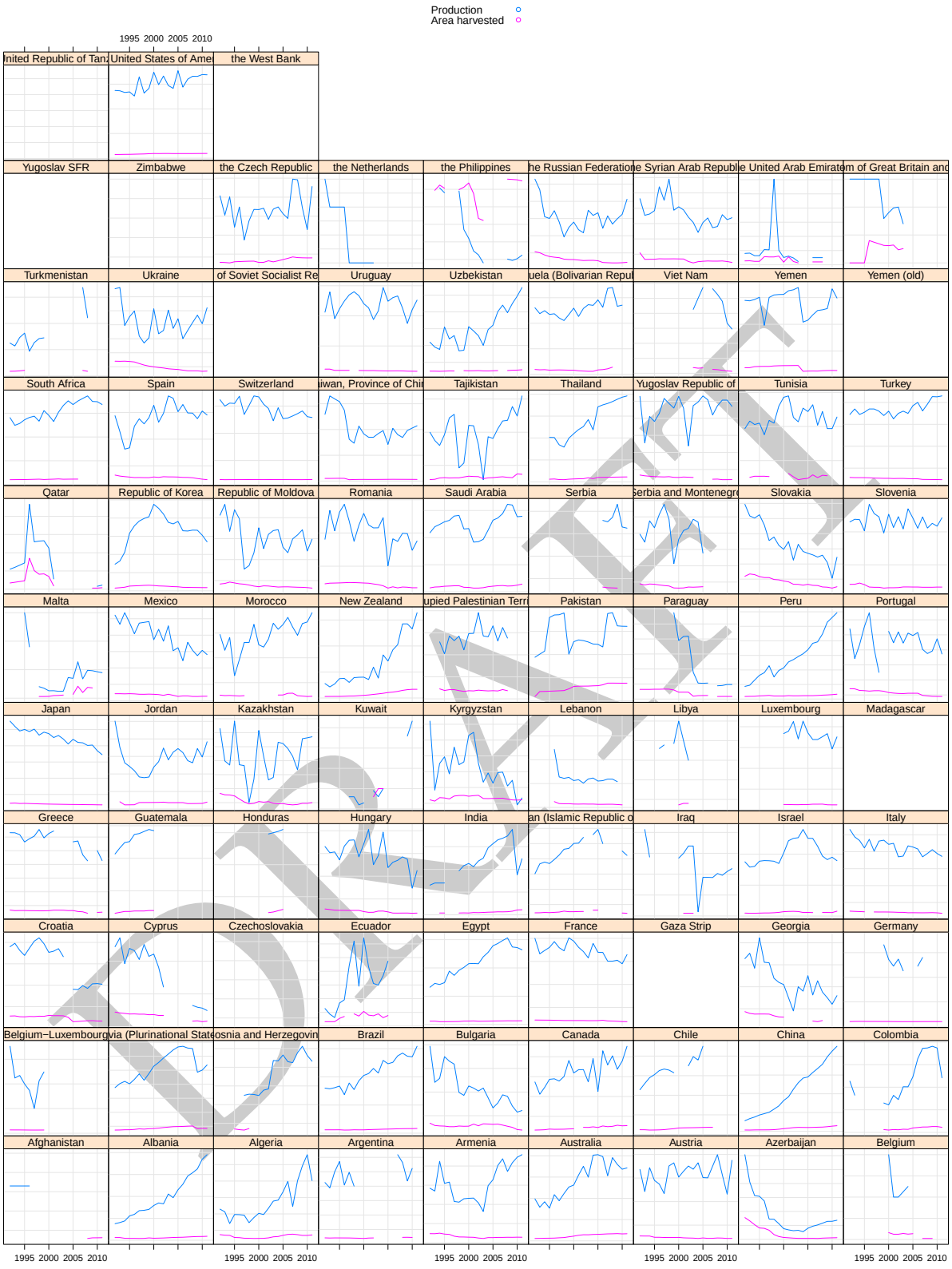


Figure 6: In contrast to wheat, the area for grape is much more stable, a character of tree which takes year to plant and nuture and the alteration of the land use is much more difficult. Nonetheless, the production also display different trends over different time period

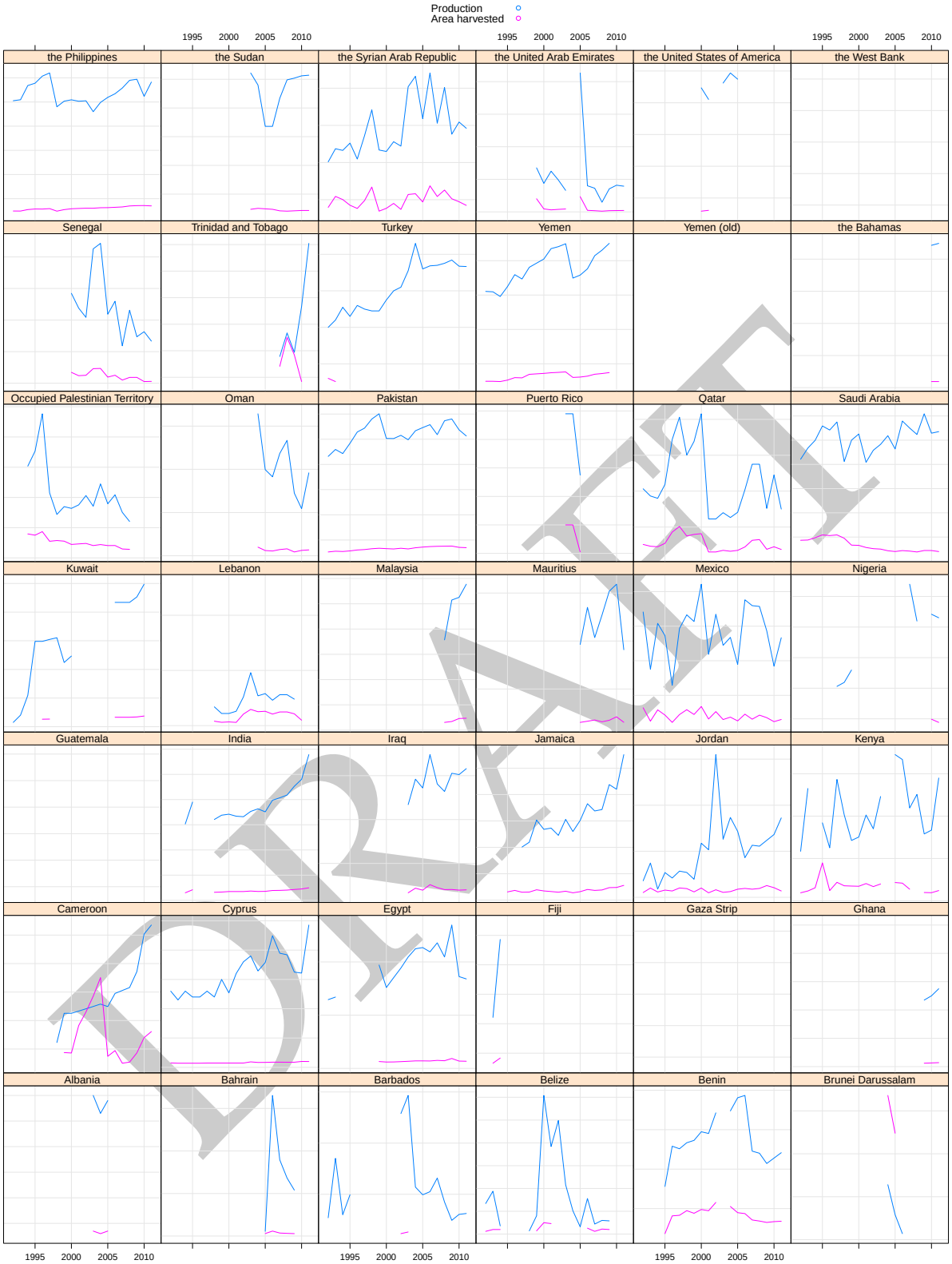


Figure 7: Even more so than both wheat and grape, the production of okra appears to demonstrate unpredictable trends and shocks.

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. Proposed Methodology

The imputation of missing observations is traditionally done via a model. For example, we may consider a simple global mean model where we compute the mean of all available observations and use that value to impute missing observations. Alternatively, we could use a more complex model (i.e. linear/exponential/logistic regression, spline, etc.), fit the model to the available data, and then estimate missing values using this model. However, this approach has two problems. First, we may choose a poor model and thus obtain poor estimates. Second, we have to specify which model to use for each set of data, and this could be very tedious if we have many time-series to impute. To avoid this problems, we consider ensemble models.

3.1. Ensemble Imputation

Ensemble learning refers to the process of building a collection of simple base models or learners which are later combined to obtain a composite model or prediction. One of the most famous applications of ensemble learning was the prediction of movie ratings held by Netflix in which the top two performers both used an ensemble of different models. Ensembles are very popular in the data-mining community because of their ability to combine multiple models and come up with an estimate that is better than any of the individual models.

The method consist of two steps:

1. Building multiple models/learners.
2. Combining the models or predictions.

The ensemble method reduces the risk of choosing a poor model as we are averaging multiple models. Thus we reduce the risk of implementing a single model which may produce poor imputations for a certain subset of data. Moreover, model selection is unecessary, since all model are included in the final ensemble.

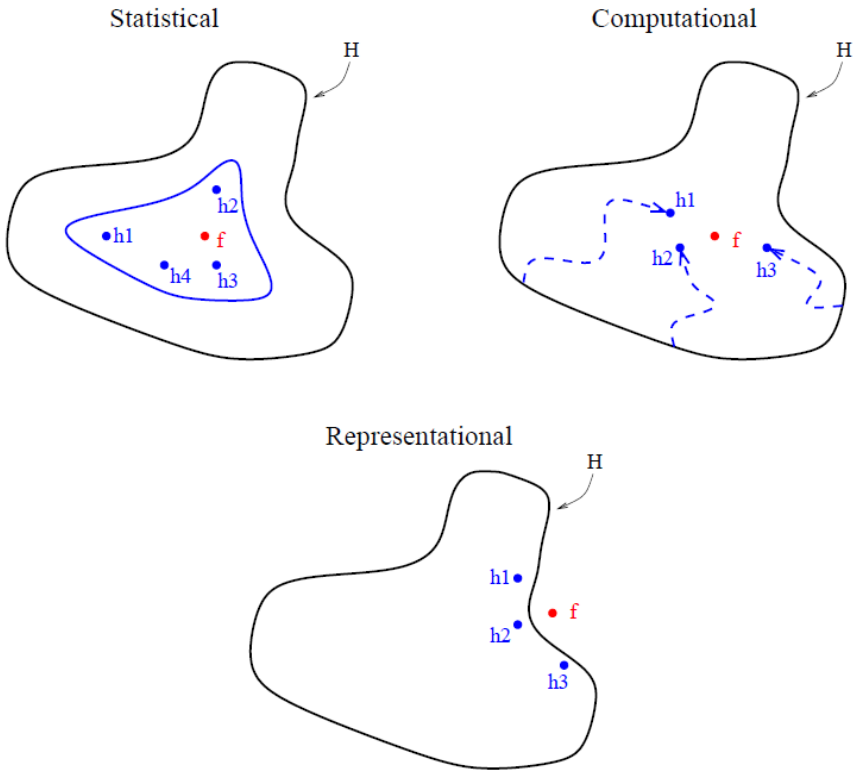
Thomas Dietterich describes several problems in machine learning in his paper "Ensemble Methods in Machine Learning" (see http://www.eecs.wsu.edu/~holder/courses/CptS570/fall07/papers/Dietterich00.pdf), and he also discusses how using an ensemble can reduce the errors from the following three issues.

Statistical: A lack of data may allow multiple models to fit the training set well.

Computational: Optimization procedures occasionally converge to local solutions instead of the global solution.

Representational: It may not be possible to model the true phenomenon with a known 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 depiction from Dietterich we see a blue boundary, and the idea is that all models within this boundary will give the same fit to the training data. Thus, there is insufficient information to determine which one is better. By combining the models, we reduce the risk of choosing a terrible model. For example, if we only observe two data points for a country, then fitting a linear line or a log curve can both give the same accuracy on the



training data and we may have no information to distinguish between the two.

The second problem is that some models are fit by optimizing some cost function. These numerical algorithms can often converge to local solutions instead of the global solution. The top right graph from Dietterich represents this problem, with points h1, h2, and h3 representing the local solutions and f the true global solution. Thus, combining the multiple fits should get us closer to the true optimum f. At the time of writing this vignette, no models which use this numerical optimization are present in the default methodology; however, we could introduce such models in the future (for example, a neural network).

The final problem, representational, refers to the fact that the true function f can not be represented by any of the individual models. However, by combining the models we may expand the space of representable functions and more closely approximate the true function f. For example, if the production of a country has been growing at a linear rate in the distant past but has expanded rapidly recently, then neither a linear or exponential model will provide a satisfactory result. However, an ensemble combining a linear and exponential model will provide a better solution by capturing different characteristics of the data.

From an implementation point of view, the algorithm is adaptive and will not need constant updating. For example, If the data generating mechanism changes in the future, the next fit of the ensemble will shift weights to models which better represent the data and thus it will not be necessary to constantly monitor and update the methodologies/models manually.

3.2. Description of Models

This section describes the different base learners for the ensemble methodology, and they are listed in increasing order of complexity. An effective ensemble will have base models as di-

verse as possible. If there is no diversity and all models generate similar results, then little is gained by combining these models and the ensemble model will not be much of an improvement from an individual model.

Mean: Mean of all observations

Median: Median of all observations

Linear: Linear Regression

Exponential: Exponential function

Logistic: Logistic function

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

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

LOESS: Local regression with linear models and model window varying based on sample size.

Splines: Cubic spline interpolation.

MARS: Multivariate Adaptive Regression Spline

Mixed Model: Linear mixed model with time as a fixed effect and country as the random effect.

3.3. Model "Level"

We wish to be able to construct models of varying levels of complexity, and in doing so we'd like to be able to build very localized models as well as more general/global models. One way of doing this is by restricting the training dataset for each model constructed. For example, if we have fairly good data availability for a particular country/commodity, then we may wish to use only that country's data when building a model. However, if one country/commodity only has a few valid observations, then we may need to use global trends for that commodity to model that country/commodity more accurately.

In the current code, there are two "levels" for constructing a model: "countryCommodity" and "commodity." The first, "countryCommodity," means that the model will only use data for that specific country/commodity pair in the fitting of the model. Thus, we will construct a different model for each different country/commodity time series. Most of the implemented models fall under this methodology: mean, linear, exponential, logistic, naive, loess, splines, and MARS.

The "commodity" level means that the model uses data for that commodity but considers all countries. The mixed model follows this approach in that all data for one commodity is used to fit a mixed effects model (where year is considered a fixed effect and country a random effect).

3.4. Extrapolation

The ensembling process makes use of many different models, but we must be careful in considering what kinds of models to use in which scenarios. As a simple example, suppose a country has production values of 1, 2, and 4 in three consecutive years and then twenty years

of missing data. If we fit an exponential model to this data, we'll be estimating production values over 4 million at the end of the twenty year period! In addition, other models (such as splines and LOESS) don't extrapolate well.

Thus, for each model, we have an extrapolation parameter. This value allows the user to control how far outside the range of the data a particular model can be used. Using this functionality allows us to prevent extrapolating with models that clearly shouldn't be used outside the range of the data while still making use of these models for interpolation purposes.

3.5. Computation of Weights

To construct an ensemble, we use a weighted average of all of the input models. However, we must determine a meaningful way for computing weights, as models which perform poorly shouldn't receive as much weight as models which fit the data well. A simple approach is to compare, for each valid observation, the model estimate and the true value. If we average the error between these two values across all valid observations, we get an estimate for the error of the model. Then, we can use this error to compute the model weights:

$$w_i = 1/e_i / \sum_{j=1}^{n} 1/e_j$$

where w_i is the weight of the ith model, e_i is the error of the ith model, and n is the total number of models. Thus, models with smaller errors receive more weight in the final ensemble, and the summation on the bottom of the above formula ensures that the weights sum up to one (ensuring that our weighted sum is in fact a weighted mean). This approach is possible in the current code by setting the errorType to "raw" in the imputation parameters list.

However, the above approach is not ideal. The problem lies in the fact that complex models generally will fit the training data better because they are more complex. In reality, we want to know if these models are more accurate at predicting unknown values, and thus we need a way to measure how effectively these models can predict on new observations. To accomplish this, we use cross-validation.

With cross-validation, the observed data are split into k different groups (often $k = 10$, and this is the default for this package as well). Then, for each group i, we build a model using all of the observed data except for those in group i and we measure how well this model estimates the data in group i. If we average this error across all k groups, we get a measure for how well this model predicts on our particular dataset. We perform this cross-validation error estimation for each of our different models, and then we compute model weights via

$$w_i = 1/e_i / \sum_{j=1}^{n} 1/e_j$$

The formula here is the same as the one above, but now the errors are the average cross-validation errors instead of the errors on the training set.

Affiliation:

Joshua M. Browning and 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: joshua.browning@fao.org, michael.kao@fao.org
URL: https://svn.fao.org/projects/SWS/RModules/faoswsImputation/