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 March 18, 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 **??** 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 **T** and are the target values to be imputed.

FIX THIS TABLE!!!

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. Exploratory Data Analysis

2.1. Yield Data

2.2. Production Data

2.3. Seed Data

2.4. 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 these characteristics.

Extremely High Sparsity

Missing values are expected given that the goal of the task is to impute missing values, but one may be stunned at the sparsity of the data. For commodities such as pepper, merely 20% of the data are available; this raises the question whether imputation approaches are even valid at all.

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 **??** 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. This resulted in a three-fold increase in the yield solely for that particular year.

Distinguishing between zeroes, missing values and N/As

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 prompted us to devise a robust methodology to safeguard ourselves from non-sensical imputation.

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.

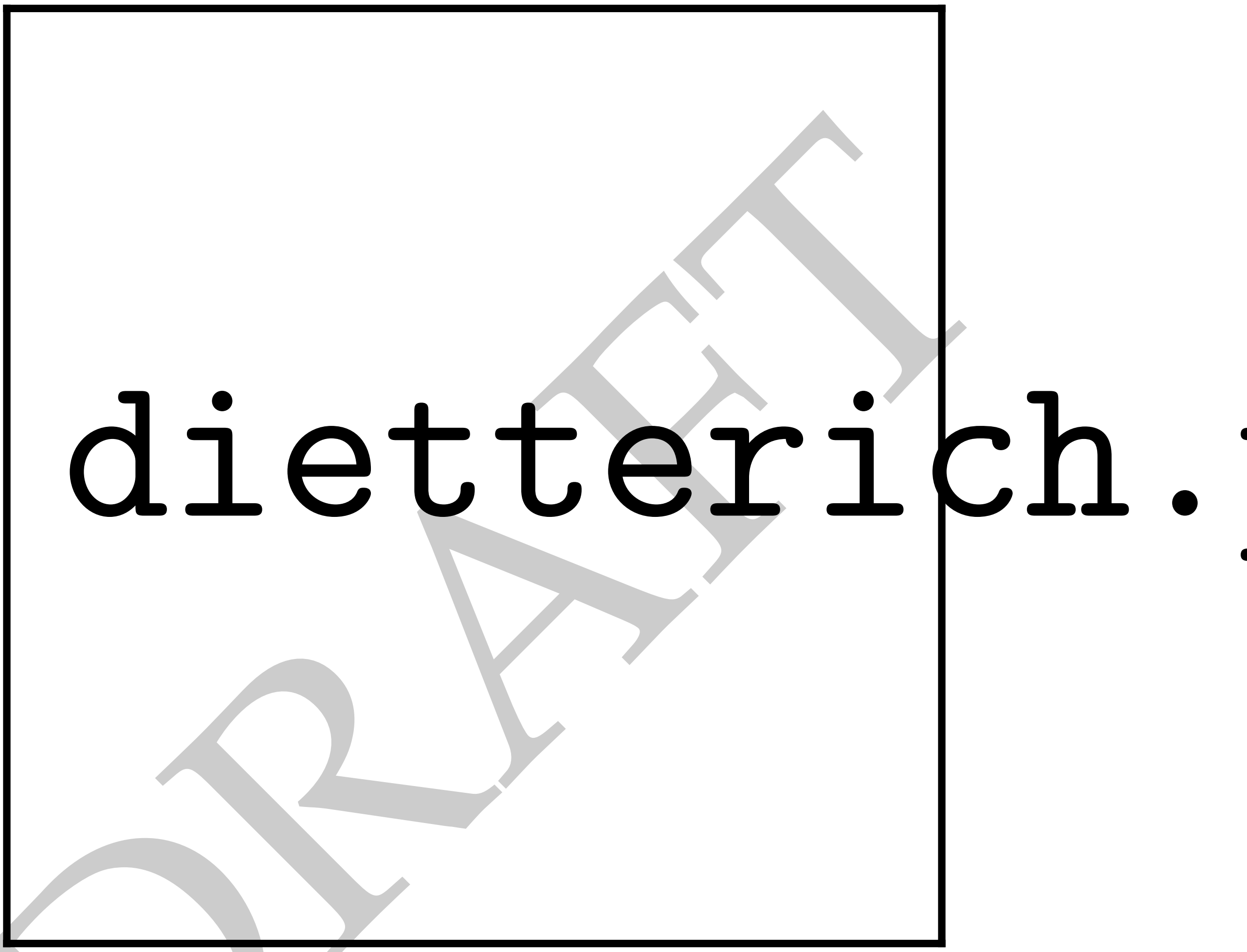
The ensemble method reduces the risk of choosing a poor model as we are averaging multiple models. Thus we reduce the risk of implementing a single model which may produce poor imputations for a certain subset of data. Moreover, model selection is unecessary, since all model are included in the final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

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

3.3. Extrapolation

Describe the purpose and show examples of extrapolation weights.

3.4. Computation of Weights

Describe the leave-one-out cross-validation procedure.

4. Case Studies

Show a bunch of examples where certain products were imputed.

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/