faoswsProduction: A sub-module for the imputation of missing time series data in the Statistical Working System - Seed

Francesca Rosa
Food and Agriculture Organization
of the United Nations

Abstract

This document illustrates in detail the approach used in the seed module. It describes the most updated version of the module and it provides to the reader the state of the art in terms of methodology reflected into functionalities already implemented though R routines.

Keywords: Imputation, Linear Mixed Model.

1. Introduction

Seed use is defined to include *all amounts of the commodity in question used during the reference period for reproductive purposes, such as seed, sugar cane planted, eggs for hatching and fish for bait, whether domestically produced or imported*. This definition includes double or successive sowing or planting, whenever it occurs. Seed use also includes the quantities needed for sowing or planting of crops for use as fresh fodder or food (e.g. green peas, green beans, maize for forage), at least when this information is available. On average, the amount of seed needed per hectare planted in a given country and for a given crop does not vary greatly from year to year, while the overall *seed quantities* depend on the *area sown* that may be subject to greater variations over time.

The module has been developed to impute the Seed quantities referring to crops.

FAO collects data for seed and area sown through the FAO production questionnaire. However, while overall response rates to the questionnaire have been rising, not all countries provide estimates for all commodities.

Where no official seed use information is available, seed use has to be imputed. The *seed use* component is not particularly important for dissemination purposes while it has a fundamental role in the Food Balance Sheet (FBS) compilation, since it is one of the main utilization for primary crop items.

The current approach has been developed in order to properly exploit all the source of information to produce a coherent and consistent output. That is why the imputations methodology starts from an evaluation of the *Area sown*.

The overall process can be summarised by two important steps:

- Impute *area sown* starting from *area harvested*
- Estimate *seed use* through a hierarchical linear model using as covariates **area sown**, **time** and **temperature**. The rationale behind the inclusion of these covariates will be introduced and described in the next paragraphs.

2. Impute Area Sown

To impute the actual area sown, the following approach is taken:

- If values of the area sown and the area harvested are available, then an average ratio of the area sown to the area harvested is computed:

$$\overline{AverageRatio} = \frac{\sum_t \frac{AreaSown_t}{AreaHarvested_t}}{t} \tag{1}$$

  Then, if the area sown is unavailable in one year, it is imputed by multiplying the area harvested in the following year by the average ratio.

$$AreaSown_t = AreaHarvested_t * \overline{AverageRatio} \tag{2}$$

  This process occurs after the run of the *non-livestock imputation module*[1]. Since *area harvested* is one of the components of the so-called *crop-production triplet*, official or imputed data about area harvested will be always available at this stage of the process.

  The link between *area sown* and *area harvested* ensure that the final *seed use* estimations are implicitly linked to *crop-production* and *yield*.

  The idea is that the *area sown* is always equal to/or greater than the *area harvested*. The ratio represents the factor to be multiplied to the available *area harvested* in order to obtain the *area sown*. The average is computed by country-commodity combinations. The discrepancy between *area sown* and *area harvested* is strictly dependent on the commodity took into account and some environmental factors (climate, kind of soil ...)

- If no prior information on the area sown is available, or the ratio is erroneosly lower that 1, the corrective factors $\overline{AverageRatio}$ are imposed equal to 1 and consequently, the *area sown* is assumed to be equal to the *area harvested*. This is a reasonable approximation that leads to feasible final imputations for the *seed use* component.

3. Climate data

As already introduced, the *Hierarchical Linear Model* to impute the *seed use* includes the *country annual average temperature*. This information is available in the *World Bank* domain of the `SWS` and in particular, in the *Climate* dataset.

The model accounts for changes over time and differences among countries which are supposed to be captured by the annual temperature variable. The assumption is that seed rates need to be higher where production conditions are difficult where there is potential for late and frequent frosts (Russian Federation), and can be lower where production conditions are more favorable.

Eventual missing observations for temperature are extrapolated through the average temperature in the country.

4. Seed use: the model

The estimation of the *seed component* is performed via a hierarchical linear model. The rationale for this model is that it is capable of capturing and modeling complicated trends when data is

[1] The overall methodology is described in the paper "Statistical Working Paper on Imputation Methodology for the FAOSTAT Production Domain "

available. Moreover, the hierarchy of the model allows accurate imputation on countries with very sparse data by pooling together global data. The mathematical model can be written as follows:

$$log(Se_{i,j,k}) = \beta_0 + \beta_1 Temp_i + \beta_2 Time + \beta_{3,j,k}(AreaSown_{i,j,k} + \epsilon_{i,j,k} \tag{3}$$

$$\beta_{3,j,k} = \gamma 3,0 + \gamma_{3,1,j,k}(Item : Country)_{j,k} + \delta j,k \tag{4}$$

$$\gamma_{3,1,j} = k_{3,1,0} + k + 3,1,1(ItemGroup)_j + \varsigma j \tag{5}$$

Where β, γ, K are coeffients to be estimated from the data, ϵ, δ, ς are erroe estimats. $Temp_i$ is the average annual temperature of country i (provided by the World Bank), and *Time* is measured in years and is included to capture linear trends over time. The i indices run over all countries, the j indices over all CPC groups, and the k indices over all unique country/CPC code combinations. Thus, the model estimates seed use proportional to the area sown. As altready introduced the model also accounts for changes over time and differences among countries; the latter are captured by the annual temperature variable.

If data for a particular country and commodity are sparse, then $k_{3,1,0}$ and $k_{3,1,1}$ will likely be stimated as close to 0. Thius $\gamma_{3,1,j,k}$ will be close to its mean value, and the model will account only for availability within commodity groups. However, if data are available for a country or commodity, the estimates of $k_{3,1,0}$ and/or $k_{3,1,1}$ will be far from 0, and thus the model enables adaptation to the individual characteristics of a particular scenario.

5. The utilization table

Using the *Hierarchical linear model* as methodological basis to produce imputations for *seed use*, it is sufficient to dispose of a series of *area harvested* to estimate the seed use component.

This means that, even if the *seed use* component time-series did not exist at all, before the launch of the module, it is possible to obtain imputations, from scratch, for the *seed use* series.

This is not correct: for many crops, despite the existence of an official series of observation for *area harvested*, the *seed use* imputation is inappropriate.

Fortunately, in the `SWS` it is available the so-called *utilization* datatable containing for each CPC item the list of elements that have to be populated. The list of seed-commodities contains a subset of primary items (such as cereals) which request to annually allocate an amount (of the item itself) for reproductive purposes.

Affiliation:

Firstname Lastname
Affiliation
Address, Country
E-mail: name@address
URL: http://link/to/webpage/