Flag Aggregation and Much More ...

Michael C. J. Kao

Food and Agriculture Organization
of the United Nations

Outline

Outline for section 1

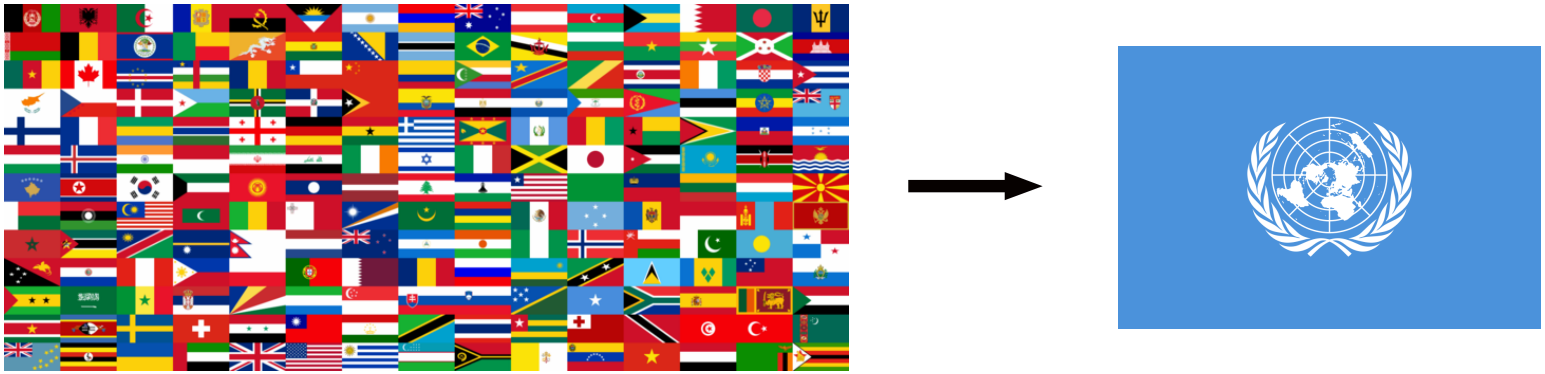


Figure : Flag Aggregation

Background

Since the introduction of the new Statistical Working System (SWS), many innovative approaches has been devised to improve the current status and to accomodate for future needs of the fast changing world of statistics.

One subtle yet fundamental change is the separation of a single symbol into two separate flag. A symbol or flag is a meta data which indicate the collection or methodological procedure that generates the value.

The aim of this presentation is to introduce the change and how the observation status flag can impact future works.

Mixing of Status and Method

Historically, a symbol can represent how it was calculated, or where it was is collected. However, this mixed approach has created some confusion and loss of information.

For example, when yield are calculated based on production and area harvested, a flag "C" representing "calculated" is assigned. However, it does not show the observation status of yield.

A value for yield calculated based on official data has the same equivalent meaning to those calculated on estimates.

This mixing of information results in loss of information and potential bias analysis.

Separation of Status and Method

To better represent these information, the decision was made to split the symbol into two separate flag reflecting different piece of information.

One for the **observation status** which represents how the data was observed, whether collected from official or semi-official data source, or it maybe estimated or imputed.

The second flag would denote the **methodology** it was obtained. Official figure can be obtained from questionnaire, or it can be from database or publications. Value which are estimated can be manually derived based or algorithm driven.

Problem

Nevertheless, it presents a problem. How do we assign observation flag when a value is calculated? Take the yield for example again, we would assign a method flag indicating the value was calculated; but what is its observation status?

If we had production value collected from unofficial source (T) while area harvested were imputed (I), what is the observation flag for yield?

Outline for section 2

What should the correct observation flag?

One approach is to quantify the quality of information for each flag and treating them as ordinal variables. Then the aggregation can be proceeded by taking the lower bounds of the set.

Flag Aggregation

$$F = \min_{\mathcal{S}} \{f_1, f_2, \cdots, f_n\}$$

Where the set $\mathcal{S}$ is the collection of flag which are used in the calculation.

The rational is that the aggregated flag should reflect the maximum amount of uncertainty.

The Flag Table

In order to take the minimum of the set, we will need to rank the flags.

Shown below are subjective weights used to rank the flags.

Table : Description of the Observation flag

Flags	Weights	Description
(blank)	1	Official Figure
T	0.8	Unofficial figure
E	0.75	Estimates
I	0.5	Imputed
M	0	Missing

Example 1

Lets take the computation of yield for example again, if we had production value collected from unofficial source (T) while area harvested were imputed (I).

Then provided with the previous table, we can compute the flag of yield as follow:

Flag Aggregation

$$\text{Flag of yield} = \min\{T, I\} = I$$

Example 2

A second example arises in the case when we need to calculate regional aggregate.

If we are trying to compute the total production of wheat of North America; with data from Canada and United States are unofficial (T) while the figure from Mexico is estimated (E).

Flag Aggregation

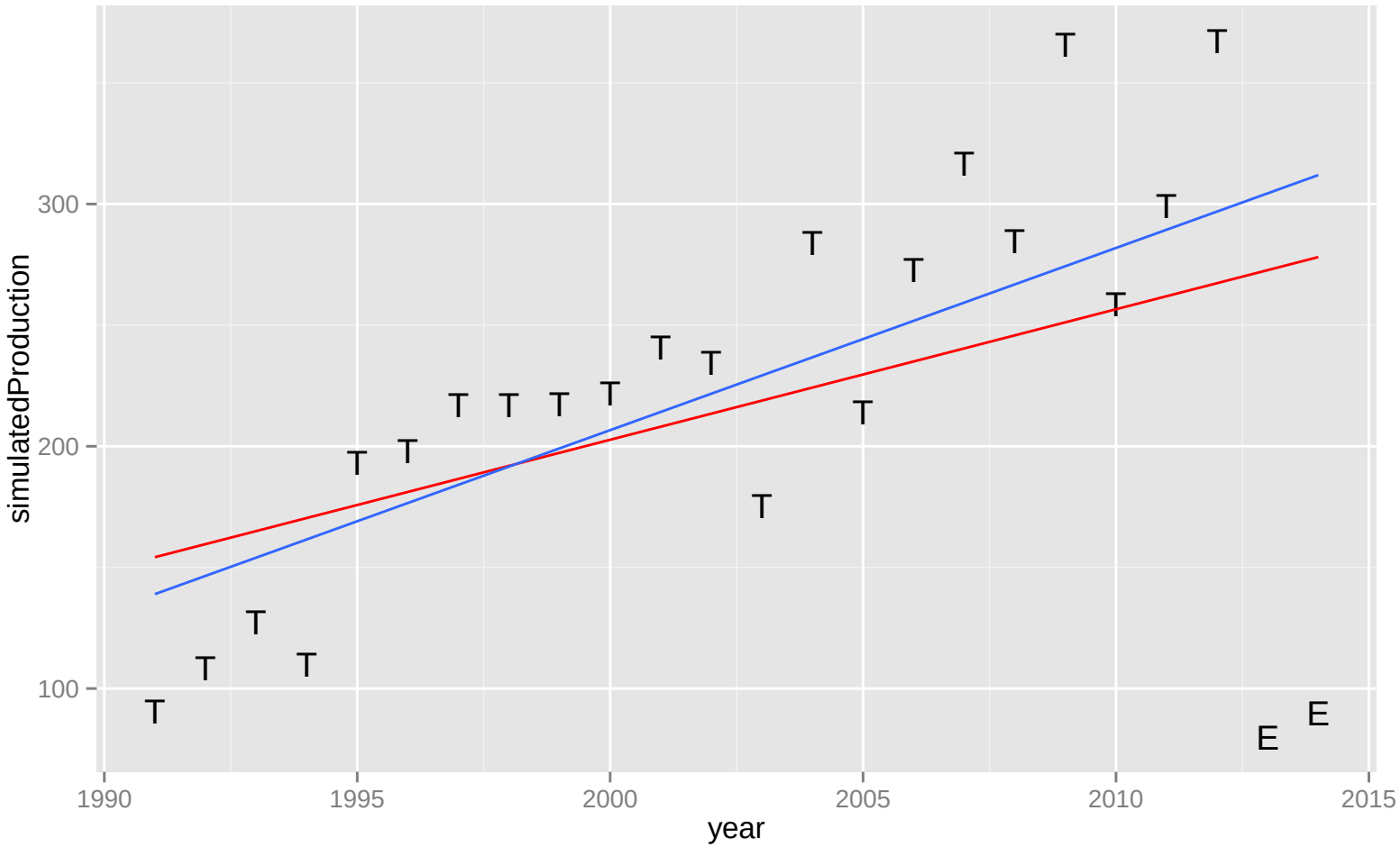
$$\text{Flag of aggregate} = \min\{T, E\} = E$$

Outline for section 3

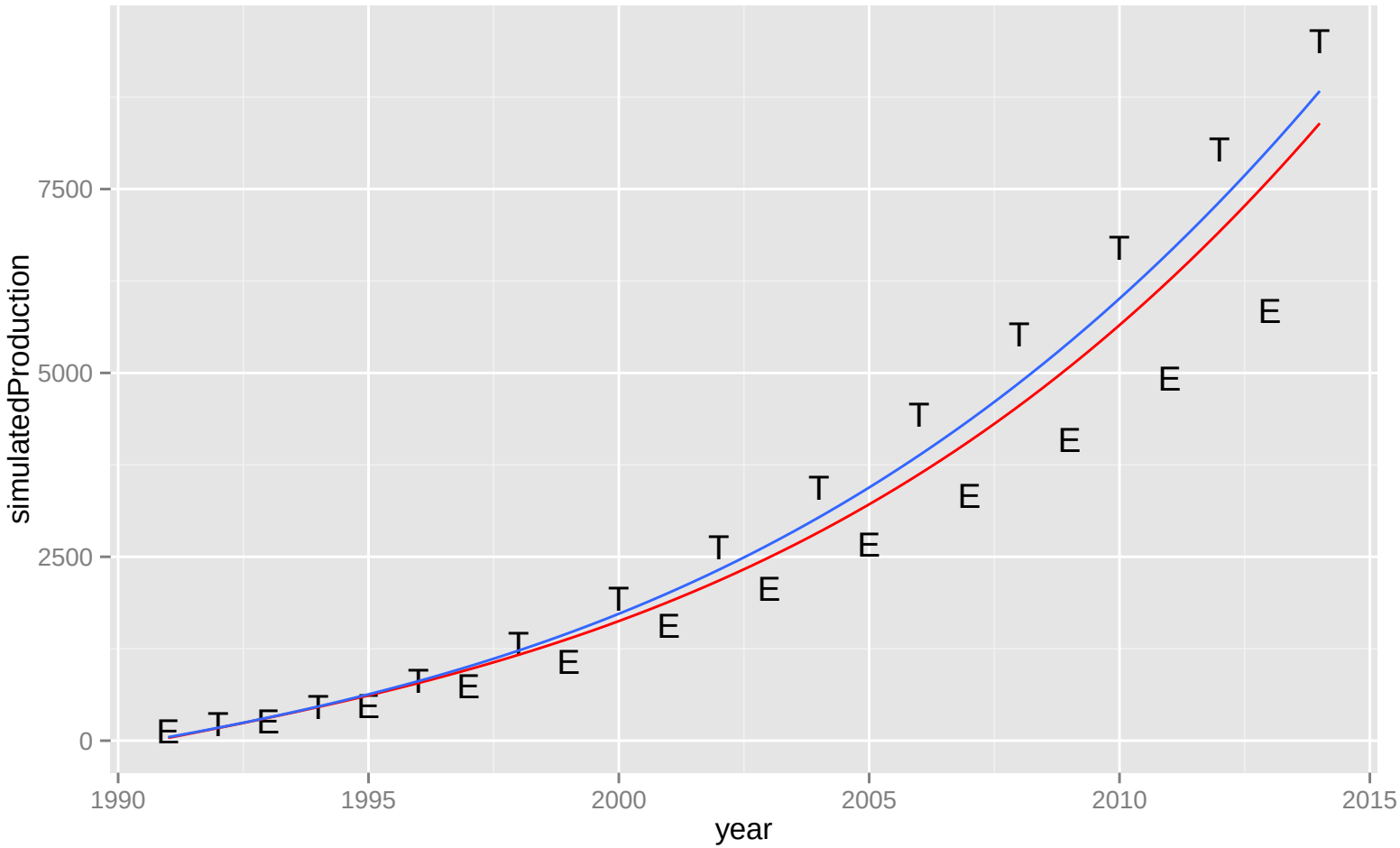
Flags have been collected for many decades, however, this piece of information have never been utilized.

The recognition of various data source and quantification of the information quality provides many potential application which can enhance subsequent analysis.

Robust From Anormalies



Combining Source of Different Information Quality



The potential associated with quantifying the quality of information is huge. It provides an additional piece of information for every single data point.

This is a general framework, no specific methodology is required since any model and methodology can utilize this framework by incorporate the additional information.

Outline for section 4

Currently, the weights are assigned subjectively by expert judgements in order to preserve a rank order.

Table : Description of the Observation flag

Flags	Weights	Description
(blank)	1	Official Figure
T	0.8	Unofficial figure
E	0.75	Estimates
I	0.5	Imputed
M	0	Missing

Objective Weights

A requirement for automatized and objective computation of weights is required due to the fact that we do not have the human resources to devote to assigning weights for each country, each agricultural domain and partition of the data.

Principle of Minimum Discrimination Information

Given derived information set, a new distribution q should be chosen which is as hard to discriminate from the original distribution p as possible; so that the new data produces as small an information gain as possible.

In another word, the principle states that if we have to choose another representation, the information set which result in the least amount of information gain or uncertainty should be chosen.

Cross Entropy

Cross-Entropy

$$\mathrm{H}(P, Q) = \mathrm{H}(P) + D_{\mathrm{KL}}(P \| Q)$$

Where

$$H(P) = -\sum_i p(x_i) \log p(x_i),$$

$$D_{\mathrm{KL}}(P \| Q) = \sum_i \log \left(\frac{p(i)}{q(i)} \right) p(i).$$

Compute the weights

However, since the entropy of $H(x)$ would be the identical. We can simple calculate the Kullback-Leibler Divergence $D_{\mathrm{KL}}(P\|Q)$.

After calculating the Kullback-Leibler Divergence, we can compute the weight according to the information gain.

Weights

$$\omega_i = \begin{cases} 1/(1 + D_{\mathrm{KL}}(P\|Q_i)) & \text{if } D_{\mathrm{KL}}(P\|Q_i) \neq 0 \\ 1 - 1e^{-5} & \text{if } D_{\mathrm{KL}}(P\|Q_i) = 0 \end{cases}$$

The New Flag Table

Shown below is the new set of weights for the wheat domain calculated based on the method above.

Table : New Observation Status Table of Wheat

Flags	Weights	Description
(blank)	1	Official Figure
T	0.9901	Unofficial Figure
E	0.97	Estimates
I		Imputed
M	0	Missing

The New Flag Table (Cont.)

Table : New Observation Status Table of Ginger

Flags	Weights	Description
(blank)	1	Official Figure
T	0.943	Unofficial Figure
E	0.873	Estimates
I		Imputed
M	0	Missing

Concusion

We hope this new change and framework will assist users to improve their work and utilize information which has been collected over several decades.