Balancing of Food Balance Sheets (FBSs)

Marco Garieri, Natalia Golini, Luca Pozzi
[name.cognome]@fao.org
November 11, 2013

Contents

Abstract

The balancing of FBSs represents a priority goal of FAO because they provide the main source of information for assessing world food situation and in order to establish the statistical base of FAO's plans for agricultural development. In this work we present a first attempt to solve the problem of balancing of FBSs. Taking advantage of the methodology developed in the statistical literature for a problem called "sampling contingency tables", we propose a version of the methodology that is suitable to the type of data we have available.

1 Problem Description

This paper describes the details of novel methodology for balancing FAO Food Balance Sheets (FBSs). A FBS represents a comprehensive picture of the pattern of a country's food supply during a specific reference period. The FBS contains information on the sources of supply and its utilization for each food item i.e. each primary commodity available for human consumption.
The sources of supply are of three different kinds: production, imports and stock at the beginning of the reference period. For the utilization several entities define the formulation: food, seed, feed, industrial use, losses, exports and stock at the end of reference period. As stated above Total Supply (TS) and Total Utilization (TU) are defined for each food item (i) in a given country (c) during the period (t) as follows:

$$TS_{i,c,t} = Production_{i,c,t} + Imports_{i,c,t} + Stock_{i,c,t-1}$$

$$\begin{aligned} TU_{i,c,t} &= Food_{i,c,t} + Seed_{i,c,t} + Feed_{i,c,t} + IndUse_{i,c,t} \\ &+ OtherUse_{i,c,t} + Losses_{i,c,t} + Exports_{i,c,t} \\ &+ Stock_{i,c,t} \end{aligned}$$

In order to have a "closed" FBS, for each item (or commodity) the sources of supply need to be equal to its utilization for a given country during a particular year. Written in mathematical terms, the following equation needs to hold:

$$TS_{i,c,t} = TU_{i,c,t} \quad \forall i \tag{1}$$

The main problem is that a FBS is assembled from a variety of different sources, both official and unofficial. For this reason we rearrange the balancing equation in 1 placing to the left of the equation the terms that come from official sources (hereafter called consolidated terms) and placing to right those that come from unofficial sources. Note that the consolidated terms may change from country to country. Thus, the aim of the project becomes the estimation the Supply Utilization Accounts (SUA) for a particular country, at a particular time for all the commodities, which are not consolidated.

1.1 Toy Example

In Table 1 we show the an example of FBS for a particular country at a particular year (i.e. Italy 2009). Here we assume that the consolidated terms are the production, imports and exports. Then we rearrange the balancing equation in order to have of the left side of equation the consolidated terms and the others on the right side, as in the following formula:

$$\begin{aligned} Production_{i,c,t} + Imports_{i,c,t} - Exports_{i,c,t} &= Food_{i,c,t} + Seed_{i,c,t} + Feed_{i,c,t} \\ &+ IndUse_{i,c,t} + OtherUse_{i,c,t} \\ &+ Losses_{i,c,t} - StockVar_{i,c,t} \end{aligned}$$

Item	Food	Feed	Seed	IndUse	OthUse	StVar	Tot
Cereals	9334	13715	765	902	371	2385	22702
Wheat	8686	2200	600	112	10	1644	9964
Rice	354	24	29	18	11	12	424
...	...	...	...	...	...	...	...
.. ...	...	...	...	...	...	...	...
Oats	18	340	35			37	356
.. ...	...	...	...	...	...	...	...
Potatoes	2370	50	150		184	-119	2873
Sweet Pot.	15	1				0	16
...	...	...	...	...	...	...	...

Table 1: FBS of Italy in 2009

where $StockVar$ represents the changes in stocks occurring during the reference period. And with the assumption that it can assume both positive or negative values.
Each line represents a different commodity. The last column (Tot) is the total of the previous columns given the following formula:

$$\text{Tot} = \text{Food} + \text{Feed} + \text{Seed} + \text{IndUse} + \text{OthUse} - \text{StVar}$$

In the Table 2 an additional column (Tot2) is added. Tot2 is described by the following formula:

$$\text{Tot2} = \text{Production} + \text{Imports} - \text{Exports}$$

Tot2 thus represents the trusted "consolidated term". A discrepancy occurs with the two different totals (Tot and Tot2) which leads to a non-closed FBS.
If and only if, for each line of the table, Tot = Tot2, the FBS has a closed form, and can be called Food Balanced Sheet.

Three additional information can help in order to solve the problem:

- A prior information on the possible value of the cells. The information for each cell is given in a distribution form (the details on that will be given in the next section)[1]
- Some cells are structural zeros, which means they are zero and they cannot take any other possible value
- An estimate of the interval where the totals of the different columns should fall into

[1]Several groups within FAO-ESS are working in specific estimate problems for Food, Feed, Seed, Losses, etc.

Table 2: FBS of Italy in 2009

Item	Food	Feed	Seed	IndUse	OtherUse	StVar	Tot	Tot2
Cereals	9334	13715	765	902	371	2385	22702	22794
Wheat	8686	2200	600	112	10	1644	9964	10035
Rice	354	24	29	18	11	12	424	423
...	...	...	...	...	...	...	...	...
.. ...	...	...	...	...	...	...	...	...
Oats	18	340	35			37	356	362
.. ...	...	...	...	...	...	...	...	...
Potat.	2370	50	150		184	-119	2873	2947
Sweet Pot.	15	1				0	16	15
...	...	...	...	...	...	...	...	...

2 Methodology

2.1 Theory

Due to the lack of strong constraints (the only one available is the total of the rows), we decided to chose a strategy able to find a solution row by row (commodity by commodity) and in the last step a validation of the results is applied on the total of the columns.
Let us assume we fix a particular country and a particular year, the procedure will be the same for different countries and different years. Then we have a table, where for each row there is a commodity C, and for each column we have the different levels L (Food, Feed, Seed, Losses, StVar, IndUse and OtherUse). For each commodity C, the only fixed value is the total of the row R, given by official information (e.g. this value is calculated as Production+Imports-Exports)

	L_1	L_2	...	L_j	...	L_s	Tot_rows
C_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1s}	R_1
C_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2s}	R_2
...	...	...	...	...	...	...	...
C_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{is}	R_i
...	...	...	...	...	...	...	...
C_r	x_{r1}	x_{r2}	...	x_{rj}	...	x_{rs}	R_r
Tot_cols	T_1	T_2	...	T_j	...	T_s	

Prior information:

- The totals of the rows R_i, given as fixed number

- The distribution of the different cells with are not structural zero $x_{ij} \sim \mathcal{TN}(\mu_{ij}, \sigma_{ij}^2)$, where μ_{ij} is the estimated mean and a fixed standard deviation σ_{ij} with bounds given as a prior information. The shape of the distribution will change dependently with the prior information we have for that particular commodity C in that particular level L. We know since the beginning that the estimates of particular levels L are more accurate than others, then if it is more accurate it will be closer to a normal distribution, otherwise it will be like a uniform distribution within the bounds

- The range of the possible outcomes of the columns' totals T_j, given as interval $(t_{j,min}, t_{j,max})$

Working independently row by row, we sample cell by cell from the given distribution until the last value of the row which will be given as difference from the total of the row R and the sum of all the previous sampled values (this value needs to fall within the given distribution for that cell, and thus a control for this is implemented). If we are successful with all the cells, the column's totals T are calculated. The table is acceptable if for all the columns, the observed value of T falls inside the given interval $(t_{j,min}, t_{j,max})$, otherwise we reject the table and start again.

2.2 Algorithm

1. For each row R_i (say commodity) of length s:

   (a) Set the last cell of the row x_{is} as StockVar, otherwise as the one with the biggest bounds

   (b) Sample all cells beside the last one, x_{is}

   (c) Compute x_{is} as difference from the totals minus all previous values: $x_{is} = R_i - \sum_{j=1:s-1} x_{ij}$

   (d) Check if x_{is} falls inside $x_{is} \sim \mathcal{TN}(\mu_{is}, \sigma_{is}^2)$, if not, sample again from the first cell of the row

2. Once all rows R_i are sampled:

   (a) Compute the column totals T_j

   (b) Check for all if T_j, $t_{j,min} \leq T_j \leq t_{j,max}$ is respected

   - If previous step succeed, the table is accepted as a solution
   - If previous step did not succeed, the algorithm starts from the beginning

3 Simulation on sample table

The first algorithm has been tested for a plausible FBS (with just seven commodities).

Item	Food	Feed	Losses	Seed	IndUse	StVar	Tot
Cereals	9230	12950	130	630	860	-350	24150
Starchy Roots	2300	190	135	155	0	-195	2975
Oilcrops	180	310	26	24	5169	258	5451
Vegetable Oils	1530	12	402	0	3	65	1882
Vegetables	12500	895	0	16	0	0	13411
Fruits	8990	0	4	0	7000	120	15874
Meat	5218	0	0	0	20	0	5239
Tot Cols	39948	14357	697	825	13052	-102	68981

where

$$Tot = Food + Feed + Losses + Seed + IndUse - StVar$$

We do suppose to do not know the actual values in each cell, but just the Tot.

The main object is to impute the missing values using information given by FAO staff. These information are given as expected value and level of uncertainty (percentage of gap from the expected value).

In order to evaluate the algorithm's behavior three different scenarios has been tested to simulate different levels of accuracy of the data given by FAO staff.

For each row (or commodity) the expected values of each cell need to be such that $Tot = Food + Feed + Losses + Seed + IndUse - StVar$ is equal or "really close"[2] to $Tot2 = Prodction + Imports - Exports$ (consolidated terms).

In the following table are shown the expected values hypothesized for each cell. Note how they do not differ substantially from the real value taken by each individual cell.

Expected Value	Food	Feed	Losses	Seed	IndUse	StVar	Tot	Tot2
Cereals	9210	12940	122	624	833	-344	24073	24150
Starchy Roots	2274	191	129	150	0	-175	2919	2975
Oilcrops	177	310	26	24	5169	277	5429	5451
Vegetable Oils	1527	12	402	0	4	65	1880	1882
Vegetables	12430	930	0	12	0	0	13372	13411
Fruits	9000	0	6	0	6965	90	15881	15874
Meat	5218	0	0	0	16	0	5234	5238
Tot Col	39836	14383	685	810	12987	-87	68788	68981

[2]Definition of "really close" is still under study.

3.1 Scenario I

In this scenario, the bounds given for each cell are really tight. In the following table both percentage and absolute value of gap from the expected values are shown for each cell.

±% (absolute)	Food	Feed	Losses	Seed	IndUse	StVar
Cereals	2 (184)	5 (647)	10 (12)	2 (12)	2 (17)	10 (-34)
Starchy Roots	2 (45)	5 (10)	10 (13)	2 (3)	0	10 (-18)
Oilcrops	2 (4)	5 (16)	10 (3)	10 (2)	2 (103)	10 (28)
Vegetable Oils	2 (31)	5 (1)	10 (40)	0	10 (0)	10 (7)
Vegetables	2 (249)	2 (19)	0	10 (1)	0	0
Fruits	2 (180)	0	10 (1)	0	2 (139)	10 (9)
Meat	2 (104)	0	0	0	10 (2)	0
Tot Col	20 (7967)	20 (2877)	20 (137)	20 (162)	20 (2597)	20 (-17)

3.2 Scenario II

The bounds, in this case, has a almost double size than in Scenario I.

±% (absolute)	Food	Feed	Losses	Seed	IndUse	StVar
Cereals	5 (461)	10 (1294)	20 (24)	5 (31)	2 (17)	20 (-69)
Starchy Roots	5 (114)	10 (19)	20 (26)	5 (8)	0	20 (-35)
Oilcrops	5 (9)	10 (31)	20 (5)	20 (5)	2 (103)	20 (55)
Vegetable Oils	5 (76)	10 (1)	20 (80)	0	20 (1)	20 (13)
Vegetables	5 (622)	5 (47)	0	20 (2)	0	0
Fruits	5 (450)	0	20 (1)	0	2 (139)	20 (18)
Meat	5 (261)	0	0	0	20 (3)	0
Tot Col	20 (7967)	20 (2877)	20 (137)	20 (162)	20 (2597)	20 (-17)

3.3 Scenario III

In this scenario, the prior bounds have an huge size comparing to the Scenario II.

±% (absolute)	Food	Feed	Losses	Seed	IndUse	StVar
Cereals	10 (921)	10 (1294)	30 (37)	5 (31)	5 (42)	30 (-103)
Starchy Roots	10 (227)	10 (19)	30 (39)	5 (8)	0	30 (-58)
Oilcrops	10 (18)	10 (31)	30 (8)	30 (7)	5 (258)	30 (83)
Vegetable Oils	10 (153)	10 (1)	30 (121)	0	30 (1)	30 (20)
Vegetables	10 (1243)	5 (47)	0	30 (4)	0	0
Fruits	10 (900)	0	30 (2)	0	5 (348)	30 (27)
Meat	10 (522)	0	0	0	30 (5)	0
Tot Col	20 (7967)	20 (2877)	20 (137)	20 (162)	20 (2597)	20 (-17)

4 Results

In the following table the execution times for each Scenario for 100 iterations are shown.

Scenario	user	system	elapsed
I	7.16	1.25	12.95
II	7.94	1.33	23.54
III	10.80	1.72	36.58

For the three Scenarios, the number of recurrent table is calculated. The main idea is to check if a particular table has more chance to be sampled. In all the three Scenarios and for different number of iterations (100, 1000, 10000) no table has a frequency more than one, thus all the sampled table are different from each other.

As a summary for the results of the different iterations for each Scenario, the Root-Mean-Square-Error (RMSE) and the Relative-RMSE (RRMSE) have been calculated. It is important to remark that this step makes sense just in a simulation study and not when the real tables will be sampled. The best iteration, considering the minimum RRMSE, are shown for the three Scenarios in the following tables:

Expected Value	Food	Feed	Losses	Seed	IndUse	StVar	$Tot = Tot2$
Cereals	9280	12944	118	636	820	-352	24150
Starchy Roots	2316	198	140	151	0	-170	2975
Oilcrops	175	307	24	23	5202	280	5451
Vegetable Oils	1500	12	433	0	4	67	1882
Vegetables	12473	925	0	13	0	0	13411
Fruits	9036	0	5	0	6927	94	15874
Meat	5220	0	0	0	18	0	5238
Tot Col	40000	14386	720	823	12971	-81	68981

Table 3: Scenario I. "Balanced" FBS ($RRMSE = 0.0874$)

Expected Value	Food	Feed	Losses	Seed	IndUse	StVar	$Tot = Tot2$
Cereals	9261	12933	123	625	846	-362	24150
Starchy Roots	2332	182	115	158	0	-188	2975
Oilcrops	172	309	30	28	5208	296	5451
Vegetable Oils	1517	11	409	0	3	58	1882
Vegetables	12437	960	0	14	0	0	13411
Fruits	9039	0	5	0	6927	97	15874
Meat	5223	0	0	0	15	0	5238
Tot Col	39981	14395	682	825	12999	-99	68981

Table 4: Scenario II. "Balanced" FBS ($RRMSE = 0.0944$)

Expected Value	Food	Feed	Losses	Seed	IndUse	StVar	$Tot = Tot2$
Cereals	9371	12798	115	648	872	-346	24150
Starchy Roots	2274	197	137	145	0	-222	2975
Oilcrops	166	310	27	26	5211	389	5451
Vegetable Oils	1518	12	431	0	4	83	1882
Vegetables	12486	909	0	16	0	0	13411
Fruits	8974	0	5	0	6995	100	15874
Meat	5221	0	0	0	17	0	5238
Tot Col	40010	14226	715	835	13099	-96	68981

Table 5: Scenario III. "Balanced" FBS ($RRMSE = 0.0942$)

5 Open Issues

There are several open issues related to the problem.

- The quality of a balance sheet and the coverage of the sampling vary considerably among countries and items (or commodities), depending on the distribution given for each cell
- Since there are not any insights from the literature for this type of problem, a proper statistical methodology is still not properly implemented
- The efficiency of the algorithm mostly depends on the accuracy of prior information and on the number of constraints
- A good definition of an objective function could help in order to find a solution to the problem. So far the objective function is a rejection of the sample in case the columns' totals do not fall inside the given intervals, but other possible solution to measure the quality of "acceptable" solutions are under revision
- So far if we fail at the last step, the columns' totals check, the information of the unsuccessful is not used in order to have a better estimate in the next sampling procedure.