Statistical Working Paper on Balancing Methodology for Food Balance Sheet (FBS)

Marco Garieri, Natalia Golini, Luca Pozzi
Food and Agriculture Organization
of the United Nations

Abstract

In this paper we describe a sampling strategy to generate balanced FBSs. A FBS is a collection of information from different sources (official and unofficial) prone to measures errors and uncertainties. Our method tries to solve the problem of the balancing of the FBSs in a flexible way, trusting reliable information and re-allocating the uncertainty over the non-reliable information. Considering agro-economical information in the form of constrains and objective functions, we are able to produce a unique solution (balanced FBS).

Keywords: Food Balance Sheet, FBS, Balancing, Structural Zero.

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 authors and are published to elicit comments and to further discussion.

This paper is dynamically generated on February 22, 2015 and is subject to changes and updates.

1. Introduction

The Food Balance Sheet presents the overall description of the food supply and utilization for a specific country in a defined time period. The information provided by the FBSs can be used to analyze the agro-economical condition of a country, in particular to assess food security trends over time for specific regions (FAO et al., 2013). Due to the importance of this information, FBSs must be precise and reliable.

FBSs represent the total food supply and utilization for a specific country in a specific year. Unfortunately, as described in "Food Balance Sheets: A handbook" (FAO, 2011), very often the quality and coverage of data is not optimal, varying considerably from country to country. FBSs are assembled from a variety of sources, official or unofficial, and inaccuracies and errors may also be introduced at every level. In case of non-reliable information the data are often estimated or adjusted while, in case of missing value, imputed. As a result, the FBSs are unbalanced, i.e., for each commodity in a specified country of the reference period the Total Supply (TS) is not equal to the Total Utilization (TU). Then, the balancing of FBSs is a primary problem within FAO.

In this paper we introduce a very flexible algorithm to solve the problem of balancing FBSs through sampling steps. The method proposed respect the balance equation for each rows (as explain in Section 2) and take in consideration of possible constrains, meaningful from a agro-economical point of view, and generate a unique solution on the basis of objective functions. The approach considers all prior information on measurement or estimate uncertainties available, in the form of mean and range around the mean (± sd) per cell or columns, in case only little information is available.

The paper is structured as follows. Section 2 describes the FBSs and the methods used to estimate its components. Section 3 shows an the new method to balance the FBSs. Section 4 presents a case study and the conclusions are in Section 5.

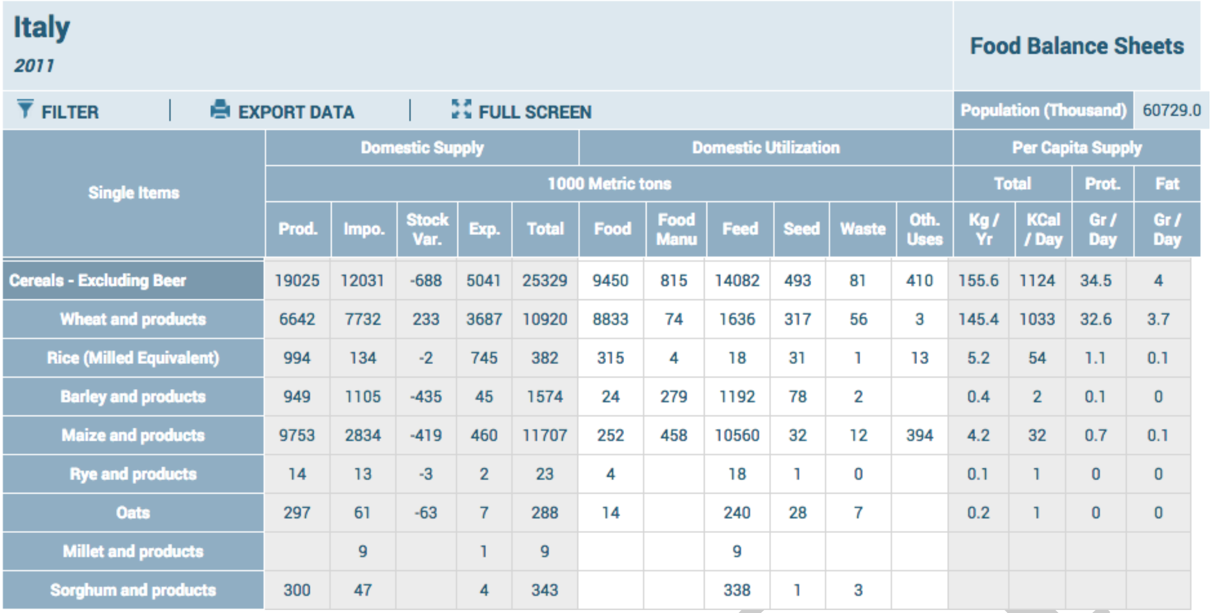
2. Background and Review of Literature

The balancing problem is well-known studied in economics literature. Often census-based Input-Output (IO) tables or social-accounting matrixes (SAM) are cases of not balanced tables, i.e., row sums different from and column sums. Some of the most used approach are the Generalized Maximum Entropy (GME) andthe Generalized Cross Entropy (GCE) techniques (Robinson et al., 2000, Britz and Wieck, 2002, Robilliard and Robinson, 2003). In Robinson et al. (2000), the authors present a flexible Òcross entropyÓ approach to estimate consistent SAMs starting from data estimated with errors. In this approach the error is a weighed average of known constants. Assumption on the weight, considered as prior, has to be given. However, in GME and GCE techniques, the interpretation of the prior information remains an unsolved problem. For this reason bayesian approaches have been proposed to overcome to this problem. In a Bayesian framework the prior information held by the researcher has a direct and interpretable formulation (Heckelei et al., 2008, Rodrigues, 2014). However, in this case, a considerable amount of prior information, not available to us, is necessary in order to have a unique solution.

The balancing problem has been associated with the solution of the contingency tables problem with fixed margins, trying to input individual cell entries and inferring the model's parameters underlying the table (see Dinwoodie and Chen, 2011, Chen, 2007, Chen et al., 2006, Dobra et al., 2006, Chen et al. 2005, Tebaldi and West, 1998). In Dinwoodie and Chen (2011), the authors propose a combination of a sequential importance sampling (SIS) with a sequentially updated normal proposal distribution. And thus, the maximum entropy distribution is sequentially approximated. The algorithm, has to sample all the cell conditionally to the constrains, and for this reason is high memory demanding and the speed depend on the table dimension. For these reason, and for the assumption of non-negative cell value, the method is not suitable for our problem of balancing FBSs.

3. FBS structure

As mention before, a FBS represents the TU and the TS of food for a given country in a specified period. Every row is a single (or aggregated) commodity and every column is each element of the Domestic Supply (Production, Imports, Exports and Stock Changes) and the Domestic Utilization (Food, Food Manufacture, Feed, Seed, Waste, and Other Uses). Each cell is a value expressed in thousands of metric tons (see http://faostat3.fao.org/download/FB/FBS/E). In Figure 1, the first lines of the FBS for Italy in 2011 are presented.



Italy

2011

Food Balance Sheets

Population (Thousand) 60729.0

Single Items	Domestic Supply					Domestic Utilization						Per Capita Supply			
	1000 Metric tons											Total		Prot.	Fat
	Prod.	Impo.	Stock Var.	Exp.	Total	Food	Food Manu	Feed	Seed	Waste	Oth. Uses	Kg / Yr	KCal / Day	Gr / Day	Gr / Day
Cereals - Excluding Beer	19025	12031	-688	5041	25329	9450	815	14082	493	81	410	155.6	1124	34.5	4
Wheat and products	6642	7732	233	3687	10920	8833	74	1636	317	56	3	145.4	1033	32.6	3.7
Rice (Milled Equivalent)	994	134	-2	745	382	315	4	18	31	1	13	5.2	54	1.1	0.1
Barley and products	949	1105	-435	45	1574	24	279	1192	78	2		0.4	2	0.1	0
Maize and products	9753	2834	-419	460	11707	252	458	10560	32	12	394	4.2	32	0.7	0.1
Rye and products	14	13	-3	2	23	4		18	1	0		0.1	1	0	0
Oats	297	61	-63	7	288	14		240	28	7		0.2	1	0	0
Millet and products		9		1	9			9							
Sorghum and products	300	47		4	343			338	1	3					

Figure 1: First lines of the FBS for Italy in 2011.

3.1. Balancing equation

The term "balancing table" refers to the attempt to balance accounts in each country for each commodity between its supply (Production, Imports, Stock Changes and Exports) and its uses (Food, Food Manufacturing, Feed, Seed, Waste, and Other Uses).
For each commodity i in a given country c at time t the Total Supply (TU) and the Total Domestic Utilization (TU) are described as following:

$$TS_{c,t,i} = Production_{c,t,i} + Imports_{c,t,i} + StockVar_{c,t,i} - Exports_{c,t,i} \quad (1)$$

$$TU_{c,t,i} = Food_{c,t,i} + FoodManu_{c,t,i} + Feed_{c,t,i} + Seed_{c,t,i} + Waste_{c,t,i} + OtherUses_{c,t,i} \quad (2)$$

Where with $StockVar_{c,t,i}$ we consider the difference between the level of stock at time t and time $t-1$. In order to have a "balanced" or "closed" FBS the following balance identity needs to hold:

$$TS_{c,t,i} = TU_{c,t,i} \qquad \forall i \quad (3)$$

The identity 3 is often not respected, due to the uncertainty of data, as explained in the previous section. Mahjoubi and Prakash (2012) and FAO (2014)

3.2. Methods for estimate TS and TU

Apart from Production and trade (Imports and Exports), which are collected through questionnaires, all other elements of the FBS are usually estimated and official data are available only for very few developed countries (Mahjoubi and Prakash, 2012, and FAO, 2014).

Production and trade

Due to these elements' reliability, these elements drive the FBS production system. Consequently, data collection efforts, especially in the Production domain, that rely on questionnaire response, remain key. The information that FBS supplies ultimately can only be as good as the core underlying data, and concerted action is required to improve data inflows.

More can be said about Production imputation, reference Michael's paper.

Food

This term represents the total supply of all agricultural and derived products available for human consumption. Food available can be reported in terms of primary product equivalent such as wheat and milk or in the form that the products may actually be consumed, such as bread and cheese. This number is typically the residual of the balance sheet. However, due to uncertainty surrounding many other elements of the balance and because of weak methods and outdated parameters, food is not the only residual element. In order to have an accurate estimate for food measurement a methodological framework for incorporating information from other existing statistical sources of information, i.e. Nationally representative Household Survey (NHS) is under development.

Jim work. Reference as soon as available.

Food Processing

This term, also known as Food Manufacturing, is the amount of a commodity processed for food purposes and for which separate entries are provided in the FBSs either in the same commodity tree or in another food commodity. This helps to maintain the concept of accounting for all foods (once and only once) and maintains the links in the various levels of the balance sheets.

Feed

This term represents the quantity of the commodity available for feeding to livestock and poultry during the year. Less than 15% of the countries respond to FAOÕs feed questionnaire and most of them are developed countries. For the rest of the world, a new method was developed. It is divided in two steps:

i determine feed requirements for metabolizable energy and protein

ii determine allocation of compound/concentrate feedstuffs to match requirements.

See FAO (2014) for more details, and reference Onno work.

Seed

Among the different categories of utilization, Seed use is one of the few categories which can be modeled a priori according to a deterministic rule, animal feed is possibly another. Seeding rates, and ultimately the demand for seed, can be modeled as a function of target plant density, establishment percentage, and seed weight. Multiplied by area planted, seed use can then be derived. However, with the rise in high-precision commercial farming, many farmers are choosing to use certified seed, purchased from specialized seed farmers. This trend requires the need to capture commercial seed production quantities.

Reference Josh work.

Waste

This term is the amount of the commodity lost through wastage during the year at all stages between farms and the household level in handling, storage and transport, but not including waste in the edible and inedible part of the commodity which occurs after the commodity has entered the household. The quantities lost during processing are also not included under this element because they are implicitly considered in applying the underlying extraction rate. Waste data have been reviewed for some 140 developing countries. It can be seen that the coefficients developed to calculate waste in FAOSTAT (waste is calculated as a percentage of total supply) have been constant for many years. Moreover, the review shows inconsistencies in the waste parameters among countries, and commodities within countries.

Reference Klaus work.

Other Utilization

This term is a miscellaneous category to account for all other uses not identified elsewhere (e.g. the use of maize to produce ethanol). This category also includes consumption by those who are not accounted in the countryÕs population (e. g., tourists).

Reference Jim work on tourism.

Stock Changes

Limited information exists on opening or closing stocks for many commodities in many countries, making stocks estimation complicated and imprecise. As a result, stock changes are often calculated to smooth supply and utilization. In practice, they are used as a balancing factor and they may, in part, be calculated residually by first estimating food available for consumption. As the item is partially and residually derived it then reflects not only stock changes but also other statistical errors in the food balance equation.

4. An alternative approach to balance FBSs

4.1. Problem formulation

A generale FBS looks like the following

Table 1: Scheme of the FBS

	L_1	L_2	...	L_j	...	L_s	$TotRows$
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
$TotCols$	T_1	T_2	...	T_j	...	T_s	Tot

Each row C_i represents a commodity and each column L_j represents a level, both supply or utilization. The table satisfies the following identities:

$$R_i = \sum_{j=1}^{s} x_{ij} \tag{4}$$

$$T_j = \sum_{i=1}^{r} x_{ij} \tag{5}$$

$$Tot = \sum_{i=1}^{r} R_i = \sum_{j=1}^{s} T_j = \sum_{i=1}^{r}\sum_{j=1}^{s} x_{ij} \tag{6}$$

First of all, we assume that for each commodity C_i, its total row R_i is fixed. This means that we rearrange the balancing identity shown in 3 placing to the left of the equation the terms that come from official sources, hereafter called consolidated terms, and to right those coming from unofficial sources. In our example, Production is the only consolidated term. Then, the balancing identity 3 can be rewritten in the following way:

$$\begin{aligned} Production_{c,t,i} = &- Imports_{c,t,i} - StockVar_{c,t,i} + Exports_{c,t,i} \qquad (7)\\ &+ Food_{c,t,i} + FoodManu_{c,t,i} + Feed_{c,t,i} + Seed_{c,t,i} + Waste_{c,t,i} + OtherUses_{c,t,i} \qquad (8) \end{aligned}$$

where $StockVar_{c,t,i} = Stock_{c,t,i} - Stock_{c,t-1,i}$. Then the FBS assumes the form of the following table:

Table 2: Scheme of the FBS

	Imps	StockV	Exps	Food	FoodM	Feed	Seed	Waste	Other	Production
C_1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	R_1
C_2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	R_2
...	...	...	...	...	...	...	...	...	...	...
C_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	x_{i8}	x_{i9}	R_i
...	...	...	...	...	...	...	...	...	...	...
C_r	x_{r1}	x_{r2}	x_{r3}	x_{r4}	x_{r5}	x_{r6}	x_{r7}	x_{r8}	x_{r9}	R_r
TotCols	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	Tot

In the same way, if we believe that the consolidates terms are more than one, i.e. Production and trade (Imports and Exports), the balancing identity can be rewritten as follows:

$$\begin{aligned} Production_{c,t,i} + Imports_{c,t,i} - Exports_{c,t,i} = &\, StockVar_{c,t,i} + Food_{c,t,i} \qquad (9)\\ &+ FoodManu_{c,t,i} + Feed_{c,t,i} + Seed_{c,t,i} \qquad (10)\\ &+ Waste_{c,t,i} + OtherUses_{c,t,i} \qquad (11) \end{aligned}$$

In this way, we consider data from official sources as highly accurate and reliable data, while those from non-official sources prone to potential measurement errors. The latter are often estimated or adjusted with a certain degree of error that depends on differing concepts, definitions, and methodologies involved in data gathering and generation among countries

(Jacobs and Sumner, 2002). For these reasons, the consolidated terms may change from country to country, and from year to year.
Then, we assume that all ammounts x_{ij} are measured with an additive error:

$$y_{ij} = x_{ij} + e_{ij} \qquad \forall i, j \tag{12}$$

where y_{ij} represents the true commodity value and e_{ij} is the difference between the measured value and its true value.
As discussed in Robinson et al. (2000), the classical assumptions made in regression analysis $Cov(x_{ij}, e_{ij}) = 0$ with $e_{ij} \sim \mathcal{N}(0, \sigma^2) \quad \forall i, j$ are extremely constraining when little is known about the error of structure and data are scarse. Even if a normal distribution can be attributed to e_{ij}, the assumption of zero mean seems not realistic in our case.
To take into account the measurement error of the individual cells (x_{ij}), we assume for these values a Normal truncated distribution:

$$x_{ij} \sim TN(\mu_{ij}, \sigma^2_{ij}) \tag{13}$$

Acknowledgement

This work is supervised by Adam Prakash with assistance from Josef Schumidhuber, Onno Hoffmeister, Salar Tayyib whom were crucial in the development of the methodology. The author would also like to thank the team members which participated in the previous discussions providing valuable feedbacks.

Annex 1: Supplementary Resources

The data, source code and documentation can all be found and downloaded from `https://github.com/marcogarieri/conSTable/tree/develop`, the package can also be installed by following the instruction.

Annex 2: Pseudo Codes

Algorithm 1: Imputation Procedure - function *swsProductionImputation*

```
Data: Production (element code = 51) and Harvested area (element code = 31) data
Result: Imputation
Missing values are denoted $\emptyset$;
Initialization;
begin
    if $A_t = 0 \wedge P_t \neq 0$ then
        $A_t \leftarrow \emptyset$;
    end
    if $P_t = 0 \wedge A_t \neq 0$ then
        $P_t \leftarrow \emptyset$;
    end
end
Start imputation;
begin
    forall the commodities do
        (1) Compute the implied yield;
                $Y_{i,t} \leftarrow P_{i,t}/A_{i,t}$;
        (2) Impute the missing yield with the yield algorithm ;
        forall the imputed yield $\hat{Y}_{i,t}$ do
            if $A_t = \emptyset \wedge P_t \neq \emptyset$ then
                $\hat{A}_{i,t} \leftarrow P_{i,t}/\hat{Y}_{i,t}$;
            end
            if $P_t = \emptyset \wedge A_t \neq \emptyset$ then
                $\hat{P}_{i,t} \leftarrow A_{i,t} \times \hat{Y}_{i,t}$;
            end
        end
        (4) Impute production ($P_{i,t}$) with ensemble;
        forall the imputed production $\hat{P}_{i,t}$ do
            if $\hat{Y}_{i,t} \neq \emptyset$ then
                $\hat{A}_{i,t} \leftarrow \hat{P}_{i,t}/\hat{Y}_{i,t}$;
            end
        end
    end
end
```

References

[1] Douglas M. Bates, *lme4: Mixed-effects modelling with R*, 2010.

[2] Data Collection, Workflows and Methodology (DCWM) team, *Imputation and Validation Methodologies for the FAOSTAT Production Domain*, Economics and Social Statistics Division, 2011.

[3] Nan M. Laird, James H. Ware, *Random-Effects Models for Longitudinal Data*, Biometrics Volume 38, 963-974, 1982.

[4] R Core Team, *A language and environment for statistical computing.*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org/, 2013.

[5] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar and the R Development Core Team, *nlme: Linear and Nonlinear Mixed Effects Models.* , R package version 3.1-108, 2013.

[6] Douglas Bates, Martin Maechler, Ben Bolker and Steven Walker, *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.0-4. http://CRAN.R-project.org/package=lme4, 2013.

[7] Donald B. Rubin, *Inference and Missing Data*, Biometrika, Volume 63, Issue 3, 581-592, 1976.

[8] Valentin Todorov, Matthias Templ, *R in the Statistical Office: Part II*, 2012.

[9] Nam M. Laird, James H. Ware, *Random-Effects Models for Longitudinal Data*, Biometrics, Volume 38, Number 4, pp.963-974, 1982.

[10] A. P. Dempster, Nam M. Laird, D. B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of Royal Statistical Society. Series B (Methodological), Volume 39, Number 1, pp1-38, 1977.

[11] Randy C. S. Lai, Hsin-Cheng Huang, Thomase C. M. Lee, *Fixed and random effects selection in nonparametric additive mixed models*, Electronic Journal of Statistics, Volume 6, pp810-842, 2012.

Affiliation:

Marco Garieri, Natalia Golini, Luca Pozzi
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: marco.garieri@fao.org
URL: https://github.com/marcogarieri/conSTable/tree/develop