Data Harmonization

EAC/FAO Advanced Training Workshop of CountrySTAT

Angela Piersante
Statistician

Lusaka, 12 – 16 November 2012

Initial issues

- The data scattering coming from multiple structures, which are responsible for producing statistics;
- The production of the same kind of statistics by different structures;
- The incompleteness of statistics;
- The absence or incompleteness of national classifications;
- The difference between the national classifications and international classifications of products;
- The lack of correspondence between national and international nomenclatures;
- The lack of an organized national level for the validation and harmonization of data;
- The weakness of data organization;
- The weakness of the technical documentation, that has to accompany the production data (metadata);

Cases of data issues

1. **Coherence between national sources**
2. **Missing and completeness data**
3. Anomalies in the **historical trends**
4. Incoherence **between related indicators**
5. Coherence between **Core** and **Sub-national** data
6. Consistency between the local and international **concepts and definitions**
7. **Correspondence** between **national and international classification**

1. Analysis of coherence between national sources

Check the **coherence between national sources**.

Production quantity of Primary Crops By Product And Year

Units: tonnes

Code	Product	Source	2004	2005	2006	2007	2008	2009	2010
15	Wheat	National Questionnaire for FAOSTAT	379425	368879	329193	322320	288642	129200	
15	Wheat	CountrySTAT	397005	365696	358061	354249	336688	219301	511994
27	Rice, paddy	National Questionnaire for FAOSTAT	49295	62677	64840	47256	63248	37198	
27	Rice, paddy	CountrySTAT	49290	57942	64840	47256	21881	42202	44468
56	Maize	National Questionnaire for FAOSTAT	2607139	2905559	3247200	2928793	2367237	2439000	
56	Maize	CountrySTAT	2454930	2918157	3247777	2928793	2369569	2442823	3464541
79	Millet	National Questionnaire for FAOSTAT	50467	53101	79207	119599	38462	54000	
79	Millet	CountrySTAT	75171	59481	79207	119599	38462	56417	53881

CountrySTAT data is different from the National Questionnaire provided to FAOSTAT by the Ministry of Planning and National Development. How can you deal with such discrepancies?

1)Data from the CountrySTAT web site

Number of Live Animals by Animal Type and Year

	2010
Cattle	16 577 962
Pigs	7 471 730
Sheep	35 519 759
Goats	56 524 075
Poultry	192 313 325

2) Data from the latest questionnaire of 2012

Code	COMMODITY	2010
		Total Stocks
	Live Animals	
866	Cattle	18871399
1034	Pigs	6040820
976	Sheep	37422554
1016	Goats	65651252
1057	Chickens ('000) (inc. Guinea Fowl)	138536162
1140	Rabbits ('000)	3689909
1126	Camels	277727
1096	Horses	101509

Analysis of coherence between national sources: possible solutions in case of contrasting data items from different datasets

As a premise, it is vital to have a **comprehensive understanding of definitions, methodology,** conversion factors and classifications. This can make the decision process more manageable and efficient.

In case for the reference period there is a discrepancy in the official data (statistics published by government agencies or other public bodies such as international organizations), consider the following suggestions:

- In case of the **National FAOSTAT questionnaire,** the Focal Point should participate in the TWG meeting, to introduce the compiled questionnaire. Then the TWG will discuss and validate the questionnaire, which will be published on the CountrySTAT website.
- Select the data item that provides **metadata** with more meaningful insight.
- Give precedence to data from **census, administrative records and sample surveys**.
- Select official data obtained by estimation based on the parameters and technical factors supported by the methodological documents.
- If all available **official data** are incoherent for the reference period, one **needs to analyze the reasons and find out the best solution, in order to reconcile the national official data.** You can, for instance, evaluate the opportunity to plan a reconciliation of the current statistics with the Census or another reliable primary source.

As an example, you can look at the case of Cote d'Ivoire, which describes the reconciliation of current data with the 2001 Census.

This situation include cases where direct data are available at different points of time or at a single point. In such cases techniques of interpolation or extrapolation may be applied. The possible methods could be the **Linear regression**, **Imputation method** and **Benchmark estimates**.

An example of FAOSTAT is shown with the method of '**Benchmark ' estimates**. They are extrapolated to other years until data for another point of time become available. The indicator should be disaggregated into its different components when these are known to have different growth patterns. In case of Production, the Yield and Area Harvested are analyzed selecting a model of Annual Growth rate of a neighboring country which shows a major correlation relation.

In this case, the selected growth model of Country2 has a correlation (between 1975 and 1984) of 0.6. It is higher than the other 2 cases. The Yield of 1984 of Country (Yt) is multiplied by the country Benchmark annual growth of a1985, as described in the formula below: 5441 x (-9.86%) = 4900 and so on until the end of the missing data series. The same procedure should be done for the Area.

With the **imputation method** the models can be created with product of the same category as first step, if data are not validated the program proceeds with neighboring countries as shown in this case.

Each country has proper national institute which produces statistics using the appropriate methodology that take into consideration the nation peculiarity.

$$Yt = (1 + T_R)\ Yt\text{-}1$$

With:

Rt: Yield at time t

T_R *: Annual Growth Rate*

Y_{t-1} *: Yield at time t-1*

Yield of CASSAVA					
Year	Country (Yt)	Country 1 neighboring	Country 2 neighboring	Country 3 neighboring	Annual growth rate of selected country as similar (country 2)
Correlation coeff.		0.4	0.6	0.3	
1975	3.321	5.2	6.667	5.6	
1976	5.366	5.3	7.276	6.111	
1977	5.062	5.4	7.806	6.333	
1978	5.18	6.111	8.064	6.556	
1979	5.254	7	8.032	6.667	
1980	5.233	8.143	8.077	6.667	
1981	5.339	9.2	9.833	6.667	
1982	5.254	9.2	8.635	6.667	
1983	4.879	9.2	8.229	6	
1984	5.441	9.2	8.8	5.042	
1985	4.90	8.289	7.932	6.164	-9.86%
1986	4.59	6.451	7.428	6.385	-6.35%
1987	4.33	6.679	6.998	7.998	-5.79%
1988	4.60	1.835	7.432	8.567	6.20%
1989	4.60	2.96	7.444	6.387	0.16%
1990	5.20	2.96	8.417	6.909	13.07%
1991	6.59	2	10.663	6.429	26.68%
1992	6.34	2.143	10.259	6.667	-3.79%
1993	6.94	1.946	11.231	6.125	9.47%
1994	7.16	2	11.578	6.25	3.09%
1995	7.41	2	11.992	6.034	3.58%
1996	7.44	2	12.039	6.5	0.39%
1997	7.34	2	11.878	6.517	-1.34%
1998	7.04	2	11.389	6.532	-4.12%
1999	7.741	2	12.253	6.51	

2. Analysis on missing data

Provide accurate information about the reason of **missing data**

Quantité de production des cultures primaires par Niveau administratif 1, Produits et Année

	2003	2004	2005	2006	2007	2008	2009	2010	2011
Bam									
Maïs	1 548	1 186	2 031	1 511	1 122	2 292	2 065	2 379	2 380
Soja	..	..	0	0	0	0	0	0	0
Igname	..	..	0	0	0	0	0	0	0
Mais Irrigue	.	6	12	26	0	6	34	0	466
Riz Irrigue	.	1	6	42	195	204	406	668	612
Tomate	..	386	..	..	..	2 603	...	...	...

- Data is provisional?
- Are statistics not collected, however the product exists in the country?
- Does the product not exist in the country because it is not produced any more?

This type of information is absolutely necessary, in order to make the best decision as of which solution to pursue when data are missing.

For example, by extending the survey to other main products, or analyzing an estimation methodology, etc.

How to provide information about missing data?

Category not applicable = . Data for these categories do not even hypothetically exist and/or data included in an other category

Data not available = .. Missing data (product exists but data not collected)

Not for publication = : Confidential data

Estimated value = "e"

Provisional or preliminary figure = "*"

Nil = - Absolute zero (data is equal to zero)

Less than 0.5 of unit employed = 0 (Insignificant data)

3. Anomalies in the historical trends

Check the **anomalies in the historical trends** comparing the data in relation to the different years

In the example below, the statistics of the DAP (diammonium hydrogen phosphate) utilization show some outliers in the time series, especially from 2007 to 2008 .

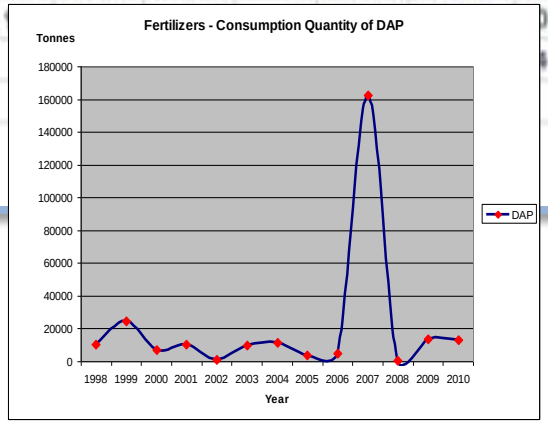
Variables et valeurs

Visualise les métadonnées de référence :

Engrais - Quantité consomée par Désignations et Année

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Urée	69 668	48 409	20 834	53 693	1 109	54 544	179 752	49 311	49 311	62 473	57 172	39 918	32 806
DAP	10 436	24 815	7 236	10 414	892	9 600	11 422	3 946	4 921	162 753	493	13 844	13 327
Engrais complexes coton	81 681	69 040	10 377	60 931	6	78 458	128 840	18 574	70 975	81 209	37 856	11 024	5 602
Engrais complexe céréales	24 253	34						748	24 961	12 887	12 408	4 628	15 857
NPKS	0							302	34 302	1 660	860	9 565	1 878
Engrais Sabugnuma	0							0	0	2 229	348	12	986

Note de bas de page:
LAST-UPDATED
2012-02-21

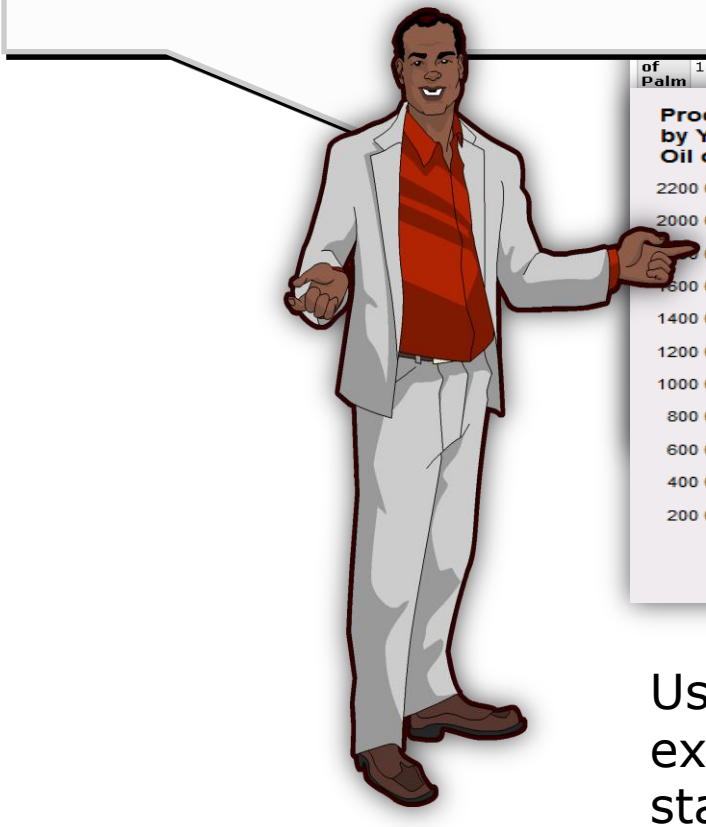


Anomalies in the historical trends: possible solutions

These are the key actions when finding anomalies in a time series:

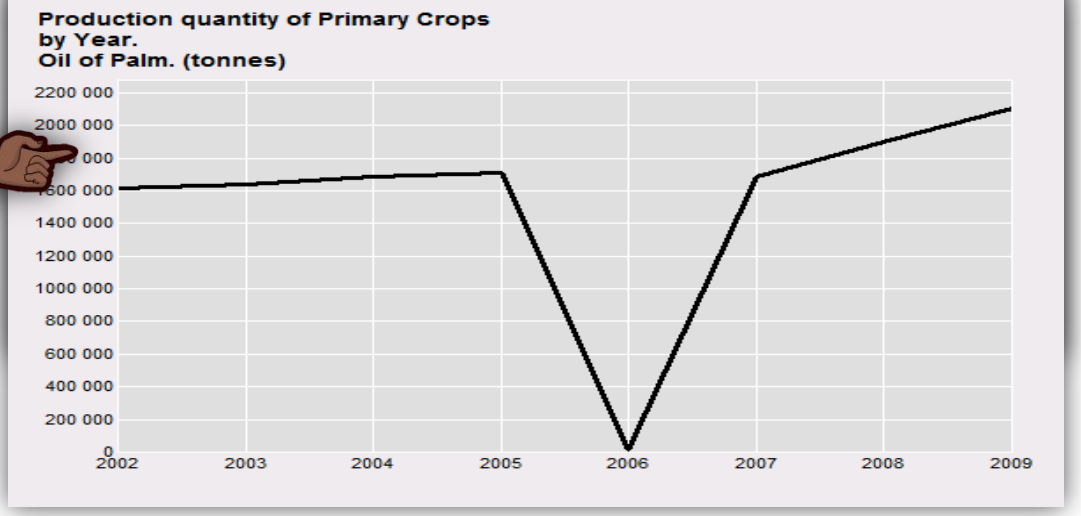
- Compare the data source with data disseminated on line, to verify if errors have happened during the data input phase;
- Analyze the social–economic and climate situation of the reference period, to find a comprehensive explanation that would justify the anomalous data;
- Compare the data source with other sources, to obtain a meaningful insight on the data, in order to evaluate the reliability of the sources.
- Another possible solution to solve the issue (if the outlier of one or more order of magnitude) is to replace it, using an appropriate estimation methodology.

The example shows that in 2005 the quantity is 1 712 600, but in 2006 it is 17 379.
As you can see, the quantity has decreased by two orders of magnitude from 2005 to 2006.



ops by Crop and Year

	2004	2005	2006	2007	2008	2009
of Palm	1 612 700.00	1 640 100.00	1 686 800.00	1 712 600.00	17 379.00	1 684 500.00	1 896 800.00	2 103 600.00



Usually a order of magnitude is expressed by 2 times the deviation standard

Possible estimation methodology

This table shows a data distribution that could be affected by errors or missing data.
How to address this issue?

1) Estimate by calculating a simple average between the year before and after the gap.
2) Estimate the missing data by calculating the series for the annual growth rates of the year before the gap, and apply the results to the following missing year.

Production Quantity of Peanuts			
Year	Pt = (1+Tp) Pt-1	Original time series with missing data	Tp
1991	121117	121117	
1992	102070	102070	-15,73%
1993	99344	99344	-2,67%
1994	117613	117613	18,39%
1995	137613	137613	17,00%
1996	**161014**	..	
1997	**188394**	..	
1998	210503	210503	52,97%
1999	184364	184364	-12,42%
2000	196702	196702	6,69%
2001	203587	203587	3,50%

$1+T_{1994/1995}$	1+17,00% X
P_{1995}	137613 =
P_{1996}	**161014**

With:
P_t : *Production at time t*
T_p : *Annual growth rate of the production*
P_{t-1} : *Production at time t-1*

The example here refers to the gap in Production Quantity of Peanuts, for 1996 and 1997. to calculate the estimates for 1996:

4. Analysis of incoherence between indicators that are in relation

Analyze the coherence between **indicators that are in relation** as they should have the same trend. For example, if the number of live cattle increases, the expectation is that the data meat production should increase proportionally. If this doesn't happen, **there is a need to revise the data**, or to **provide an explanation**.

Number of live animals by species and year

2005	2006	2007	2008	2009	2010	2011
6 770 000	6 973 100	7 182 293	11 408 740	11 751 002	12 103 532	12 466 638

Production of meat by product and year (tonnes)

2005	2006	2007	2008	2009	2010	2011
147 000	160 000	174 150	169 950	175 049	180 300	185 709

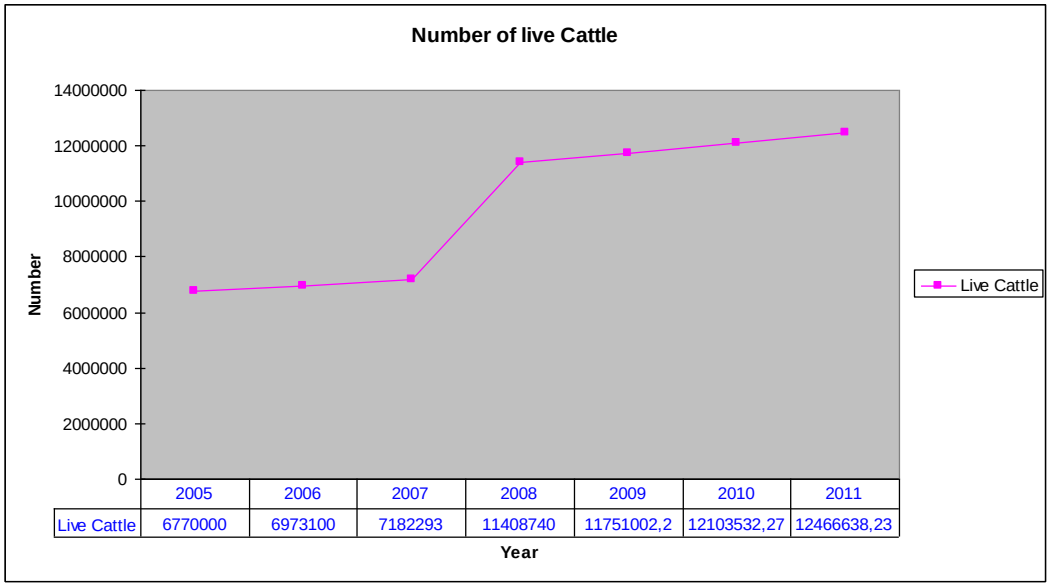
- The **first issue** is that in 2008 the cattle live animals have increased by 50%, which affects the following series (it looks like we have a broken series).
- The **second issue** is that the meat production decreased by 2.41%.

These two times series in relation show two different issues to be addressed!

Analysis of incoherence between indicators that are in relation: possible solutions: Solution for point 1

The **broken time series** for " *Number of Live animals*" took place because the data collection methodologies used were different:

1. Before 2008, data were collected from administrative records,
2. In 2008, data were collected by the Livestock Census,
3. After 2008, data were estimated by the growth rates based on the 2008 Census.



	2005	2006	2007	2008	2009	2010	2011
Live Cattle	6770000	6973100	7182293	11408740	11751002,2	12103532,27	12466638,23

Based on this information, in order to make the time series homogeneous, one can carry out a data reconciliation of the time series before 2008. It is possible to use the Census from 2008 as a reference. Look again at the Cote d'Avoire example to see this methodology in practice.

Analysis of incoherence between indicators that are in relation: possible solutions: Solution for point 2

To address the second issue, which is the lack of a coherent relation between the two indicators, one needs to revise the Meat Production time series, on the basis of the Number of Live animals time series. This can be done by estimating data using national conversion factors.
A suggestion would be to refer to the FAOSTAT conversion methodology, as shown in the example below:

Number of live cattle		FAOSTAT take-off rate		Estimated Number of Slaughered Cattle		Carcass weight Average (kg)		Estimated Meat production (Kg)
7182290	X	11,3%	=	811003	X	150	=	121650396

Take-off rate : The number of animals taken out from the national herd during the year to be slaughtered in the country or exported alive. These numbers are expressed as a percent of all animals of the same species which were present in the country at the time of enumeration of the same year, including newborn animals, for example, day old chicks.

$$\frac{\text{Slaughered animals - Imports + Exports}}{\text{Stock of animals}}$$

If the take off of the country is not available, it is advisable to evaluate the adoption of the take off (possibly based on official data) of a neighboring and similar country.

Carcass weight – This means dressed carcass weight, i.e. the weight of the carcass after removal of edible and inedible parts, particularly hides and skins, offals and slaughter fats.

5. Analysis of the coherence between the data in the CORE and Sub-national sections

Analyze the **coherence between the CORE and the Sub-national module**

Production quantity by National level			
2004	**27 Riz**	**56 Maize**	**79 Mil**
	74 501	481 474	800 630

Production quantity by Administrative level 1			
			79 Mil
2004	40276	Region 1	156 542
	40277	Region 2	6 209
	40278	Region 3	20 809
	40279	Region 4	101 114
	40280	Region 5	67 760
	40281	Region 6	114 362
	40282	Region 7	85 647
	40283	Region 8	107 775
	40284	Region 9	42 210
	40285	Region 10	81 573
	40286	Region 11	44 374
	40287	Region 12	53 298
	40288	Region 13	55 955
		TOTAL	937 630

This situation can occur for many reasons as the sources of the two administrative levels (National and Sub national) are different, and/or if they use different calculation methods.

It would be ideal to disseminate data from the same primary sources, to avoid inconsistencies between CORE and Sub-national modules.

Coherence between data in the national and sub national sections

As a first activity, the TWG should assess whether an error had occurred in the data input phase or during the official report preparation, by comparing the data source with the data disseminated on line.

Here are some ways in which the TWG can address and solve incoherencies **in case the sources are different**:

- Select the data item that provides metadata with more meaningful insight.
- Data from a census, administrative record and sample survey should take precedence.
- Once the source which includes Sub national data is analyzed and evaluated as reliable, then the National data should be derived by the calculation of the sub national data.
- Once the source which includes national data is analyzed and evaluated as reliable, then the sub national should be recalculated with a data reconciliation analysis, using an appropriate estimation methodology (as previously mentioned).

On the other hand, when the sources use **different collection methods**, the responsibility is more institutional, because data have already been validated by high authorities (and perhaps used during official events).
In this case, a data reconciliation should be elaborated, and later validated and made official through a national process.

The best case

Production quantity by Product at CORE			
2004	27 Riz	56 Maize	79 Mil
	74 501	481 474	937 630

Production quantity by Regionel 1			
			79 Mil
2004	40276	Boucle Du Mouhoun	156 542
	40277	Cascades	6 209
	40278	Centre	20 809
	40279	Centre-est	101 114
	40280	Centre-nord	67 760
	40281	Centre-ouest	114 362
	40282	Centre-sud	85 647
	40283	Est	107 775
	40284	Hauts-bassins	42 210
	40285	Nord	81 573
	40286	Plateau Central	44 374
	40287	Sahel	53 298
	40288	Sud-ouest	55 955
		TOTAL	937 630

Tableau 06 : Poids relatif des productions réalisées pour chaque céréale

Cultures	Campagnes agricoles 1999-2000 à 2003-2004		Campagne agricole 2003-2004		Campagne agricole 2004-2005	
	En tonne	Proportion en %	En tonne	Proportion en %	En tonne	Proportion en %
Mil	947 428	33	1 184 283	33,2	937 630	32,3
Sorgho	1 276 167	44,5	1 610 255	45,2	1 399 302	48,2
Maïs	541 920	18,9	665 508	18,7	481 474	16,6
Riz	91 014	03,2	95 494	02,7	74 501	02,6
Fonio	10 811	00,4	8 741	00,2	9 066	00,3
Total	2 955 707	100	3 564 281	100	2 901 973	100

Source : MAHRH/SG/DGPSA/DSA – Résultats définitifs de la Campagne Agricole 2004/2005

The best practice is to disseminate the CORE data in line with the disaggregated data coming from the primary source

6. Analysis between the local and international concepts and definitions

Sometimes, the national concepts and definitions are different from those used at international level. An analysis of the concepts and definitions at national and international level is necessary, to respond to the international requests and ensure data comparability. The example below shows significant discrepancies on the disseminated data, because of the differences in the definitions of the indicator 'Seed Utilization'.

Utilization of Seed by Product and Year

Units: tonnes

Code	Product	Source	2004	2005	2006	2007	2008	2009	Concepts and definitions
27	Rice, Paddy	FaoSTAT	2628 (Fc)	2210 (Fc)	2027 (Fc)	4005 (Fc)	4612 (Fc)	4612 (Fc)	Calculated Data, based on FAO concepts and definitions which includes 'Certified and Uncertified' seeds
27	Rice, Paddy	CountrySTAT	18	56	283	174	877	1541	Data collected by the country, based on ' Certified' seeds (imported or produced by specialized firms)

Since the quantity of uncertified seed is missing, how can we calculate the missing quantity to be added to the certified seed, in order to estimate the total national data?

Analysis between the local and international concepts and definitions

A possible solution is to calculate the quantities of non-certified seed by multiplying the area planted with the **Utilization seed ratios** (provided by the country). The table below is an example of utilization seed ratios of non-selected seed.

Example for Maize:

Area Sown of 2001 (ha)	334682 X
Seed rate utilization of non certified Seed (kg/ha)	18 =
Estimated non certified seed of 2000 (kg) (the rate is applied to the area of following year)	6024276

Estimated non certified seed of 2000 (tonnes)	6024 +
Certified Seed of 2000 (tonnes)	495 =
Total of Utilized Seed of 2000	6519

This will be the final data to be published

Utilization seed ratios of non-certified seed

Products	Seed weight (kg/ha)
Millet	8
Sorghum	12
Maize	**18**
Rice paddy	65
Fonio	35
Groundnuts	125
Sesame	5
Cow peas	23
Pois Bambara	125
Yam (tuber)	109

7. Analysis of the correspondence between national/international classification

The analysis on the correspondence between the national and international classifications can sometimes show incoherencies, due to the differences in the commodities lists.

The example below shows that a large category such as "Tropical Fruits" can report different time series, because of different commodities list in the correspondence between commodity classifications described in the questionnaire and CountrySTAT data.

This source includes only **Passion Fruit** in the Tropical Fruit category

Production quantity of Primary Crops By Product An

Units : tonnes

Code	Product	Source	2001	2002	2003
603	Fruit, tropical fresh nes	FaoSTAT	41041	33484	33686
603	Fruit, tropical fresh nes	CountrySTAT	114323	110929	115484

This source considers as Tropical Fruit several products: **Passion Fruit, Guava and Paw Paw**

This kind of discrepancy, gives you the opportunity to revise the accuracy of your Correspondence Table. In this case, the question would be: why is there such a discrepancy? Is our Correspondence Table accurate and comprehensive enough, or is it leaving out some products?

Check and revise the accuracy of classifications correspondence

The solution is to create an accurate Correspondence Table between the two classifications, based on an analysis of the concepts and definitions related to the commodities, at national and international level.
A coherent correspondence will provide unique and more consistent data.

International commodity codes	International commodity description	Local commodity codes	Local commodity description
603	Fruit, tropical fresh nes	060301	Guava
603	Fruit, tropical fresh nes	060302	PawPaw
603	Fruit, tropical fresh nes	060303	Passion Fruit

Thank you !!!