12 RESPONSIBLE CONSUMPTION AND PRODUCTION

Food and Agriculture Organization of the United Nations

Supplementing Data and Knowledge With Model Based Estimates

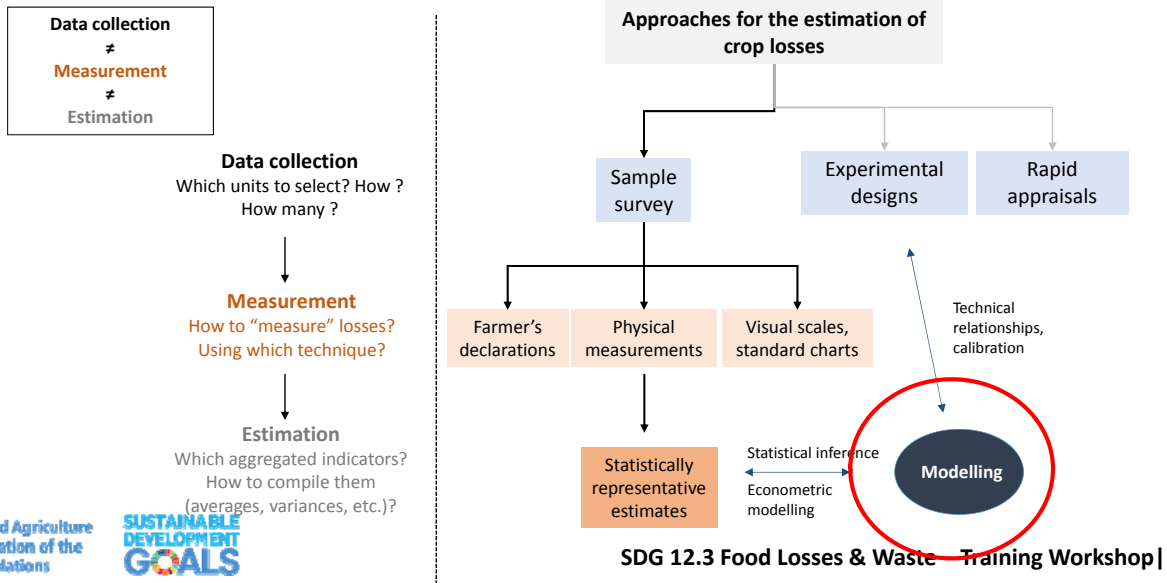
FAO Workshop on: Monitoring SDG 12.3.1 Global Food Loss Index
April 26-27 2018

Presenters:
Ms. Carola Fabi and Dr. Alicia English
FAO Statistics Division

Outline

- Modeling structure
- Efforts to Model Post-Harvest Losses
- ESS – SUA (Pre-Balancing) Loss Model
  - Commodity Groupings
  - Variable Selection
  - Model specification
- Estimates for the FLI Baskets

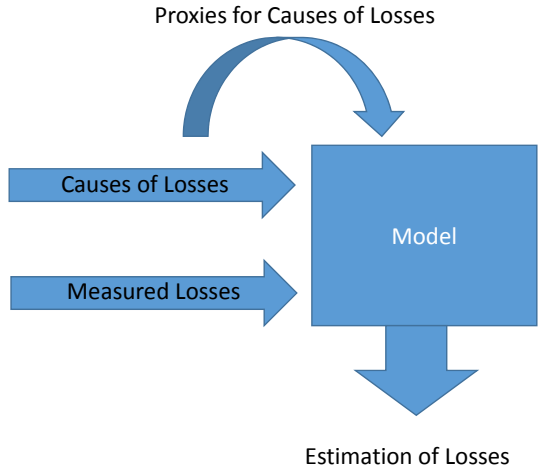
Guidelines for tools to use in new data collection



Connection of Data to Models

- Data Resolution and Integration - Levels of Modeling
  - National
  - Stage of value chain
  - Strata within stages
- Representability/Causes/Correlation with losses
- Appropriate Model Framework
  - Types of data collected/needed
  - Disaggregation/Aggregation
  - Model Selection (Parameterized Models, ANOVA, Ordinary Least Squares)

Connecting the Data Collection Strategy to a Model Framework



A model estimates the impact of the independent factors on a depended variable.

- Is there a relationship?
- How strong is the relationship?
- Does it make a positive or negative impact?
- Is it linear/What is the appropriate functional form?

12 RESPONSIBLE CONSUMPTION AND PRODUCTION

Previous Model Efforts at the Global Scope

FAO 30% Loss Study

- The Global Food Loss and Waste – extent, causes and prevention
- FAO (SAVE FOOD) to the Swedish Institute of Food and Biotechnology (SIK) from August 2010 to January 2011. The estimates from this paper conclude that the global food loss and waste is approximately 30% of all food produced.
- The study uses a mass flow model to quantify the annual volumes of food loss and waste at a global scope.
  - It divides the world into three main categories (low-, medium- and high-income countries)
  - Food Balance Sheet (FBS) groupings (Cereals; Root & Tubers; Oilseeds & Pulses; Fruit & Vegetables; Meat, Fish & Seafood; Milk & Eggs).
  - Conversions factors were used to convert food available for human consumption to their equivalents based on the literature available on non-edible quantities of different commodities.
  - Included animal feed and consumer and retail
  - Is not replicable, given the high uncertainty of where the conversion factors on the supply stages originated

APHLIS

- African Post-harvest Losses Information System (APHLIS)
  - Cereals Oriented (maize, sorghum, wheat, millet, rice, barley, teff, oats, fonio) in Sub-Saharan African countries (SSA).
  - Mix of expert opinion, studies
- Scenario Based Model
  - user may adjust these variables to create scenario outcomes on harvest periods, pest incidences, weather, etc
  - But the impact from theses variables are not dynamic

SUA/FBS Loss Model

- The FBS is a time-referenced food accounting framework whereby supply equals utilisation (in quantities):

Total Supply = $\boldsymbol{Total\ Utilization}$

Total Supply = $Production + Imports - \Delta stock$

Total Utilization = *Food + Feed + Seed + Loss + Industrial Use + Tourist Consumption + Residual Other Use*

- Food Balance Sheet (FBS/SUA) Methodological Update & Improvements
  - 4% of countries have any reported losses data
  - Each module has been undergoing review and improvements
  - Previous model relied on production, commodity perishability and country to estimate losses

SUA/FBS Loss Model

$$\log(\text{Loss}_{ijk}) = \alpha_1 t + \alpha_{2ijkl} \log(\text{Production}_{ijk} + 1) + A_{ijk}$$

Does not change

Does not change

Does not change

Does not change

where:
i is country,
j is commodity,
k is food group

The 1990-2016 FBS Model relied heavily on production to estimate losses. Other factors commodity group and country were used to adjust the losses (quantities)

α_{2ijkl} is the only element that changes from year to year

Can't be used for SDG monitoring or for policy making

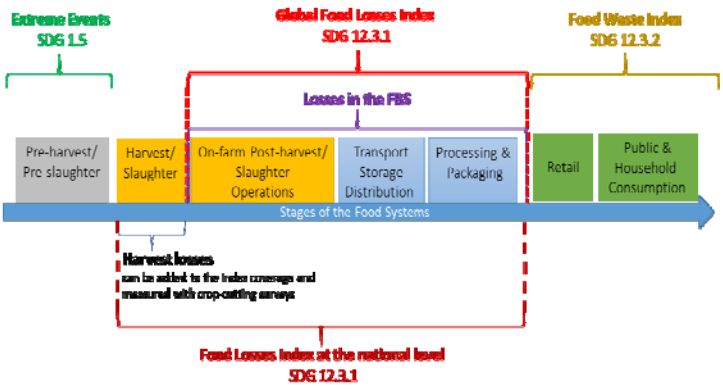
Food Loss Indices under SDG 12.3.1

12.3.1 Food Loss

"…reduce **food losses** along **production and supply chains**, including post-harvest losses."

<u>Steps to compiling the Index if the data exists:</u>

1. Select Basket of commodities and compile weights
2. Compile Food Loss Percentages
3. Compare Food Losses over time



Indicator 12.3.1 - Countries' Food Loss Percentages (FLI)

Step 2: Compile the **Food Loss Percentage (FLP)** of the whole basket of commodities at country level:

The FBS data and model outcomes are what calculates loss percentages and provide q_0

$$FLP_{it} = \frac{\sum_j l_{ijt} * (q_0 * p_0)}{\sum_j (q_0 * p_0)}$$

Economic weights

- The FLP is composed of several commodities
- The FLP is the average loss of these commodities
- Not all commodities have the same importance - weights

Justification for New Approaches

- Provide a mechanism for aggregating subnational stages data to a national estimate for each country/commodity/year
- Build off existing efforts, include more policy relevant variables
- Create a comparable, transparent method for countries that do not have officially reported data and to estimate loss while addressing many of the previous modeling limitations
- Solve the undercoverage issue
- Incorporate country feedback
- Foster a standardized, homogenous approach for estimating losses and selecting explanatory variables based on commonly found factors that contribute to loss and be able to encapsulate the variability needed.

Conflicting Data Needs

- FBS need *Loss data* to estimate Food Availability
  - Under coverage of the data means that these estimates will likely decrease
  - More data collection will likely increase the loss estimates
  - Altering the time series of data
- The SDGs need Loss data for estimating 12.3 on Food Losses, but also the related indicators on Food Security, measurement of undernourished, etc.
  - Also need to be able to show the changes in loss related to changes in policy, interventions etc.
  - Food systems approaches

Loss Model

ESS – SUA (Pre-Balancing) Loss Model

Clustered Commodity Groups

- Assume that commodity groups will have similar causes and rates of losses within the commodity groups than across them.
  - E.g. Corn and lentils vs. Corn vs. fresh milk
  - Don't want the higher losses of one commodity impacting the modeling framework.
  - Types of commodity value chains and key stages where commodities are impacted and solutions are similar within the groups
  - But there may be factors that impact multiple supply chains – e.g. electricity, infrastructure

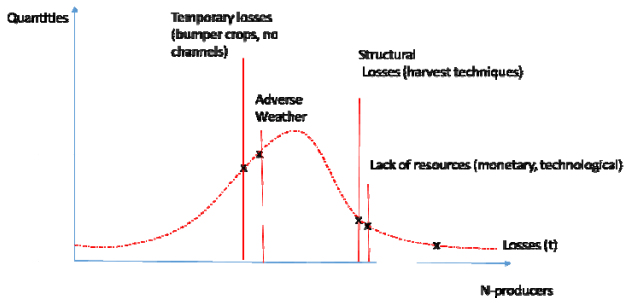
Commodity Baskets

1. Cereals & Pulses;
2. Fruits And Vegetables;
3. Roots, Tubers & Oil-Bearing Crops;
4. Animals products; Fish and fish products
5. Other crops (stimulants, spices, sugar, etc.)

Consistent with the FLI Baskets, which will help improve coverage & will help the model run

Explanatory Variables

- Set of 200+ Explanatory Variables based on the causes of losses outlined in the 40 years of post-harvest literature.
  - Transport, costs of energy, access to capital in agriculture, logistics, etc.



Challenge: Micro-level studies and the connections to policy relevant indicators.
- How do we make connections between studies and policy?
- What impacts multiple stages/commodities?
- What are good proxies for things that cannot be measured directly?

Variable Selection

- Literature provided theoretical basis for the explanatory variables
- Used these as proxies to find variables that were collected at the national levels. Focusing on:
  - Storage
  - Transportation
  - Input Costs (Fertilizer, etc.)
  - Energy
  - Investment/Monetary
  - Social/Economic
  - Weather/Crop

The variables were sourced from international organizations. For example:
- Quantities of Energy used in Agriculture from IEA by country
- Metal and Fertilizer prices from the World Bank Pink Sheets)
- Capital Stock values from FAOSTAT
- Access to electricity (% of population)
- Etcetera

Variable Selection

- Data can be improved in these categories
  - Crop Calendars
  - Temperature/Rainfall
  - Logistics Measurement
- Data does not exist at the national level for some of the factors found in the literature at the farm level or any other stage (e.g. Hermetic storage bags)
  - Model can be used at the national level to model losses, replacing/adding datasets
  - Challenge
    - Lack of data
    - Estimating by stages (e.g. academic studies)
    - Domestic prices at the different stages
    - Connecting behaviors to policy-relevant choices (e.g. handling as a cause, but fixing it as a policy?)

New Model Specification

Loss % = function (commodity, time, country, random effect)

$$y_{ijt} = \alpha + x_{ijt}^{T}\beta + z_{ij}^{T}\gamma + u_{ijt}$$

where:

y_{ijt} is the percentage of food losses for the country i, for a given commodity, j, at time t

x_{ijt}^{T} is the k-dimensional row vector of time and commodity varying explanatory variables

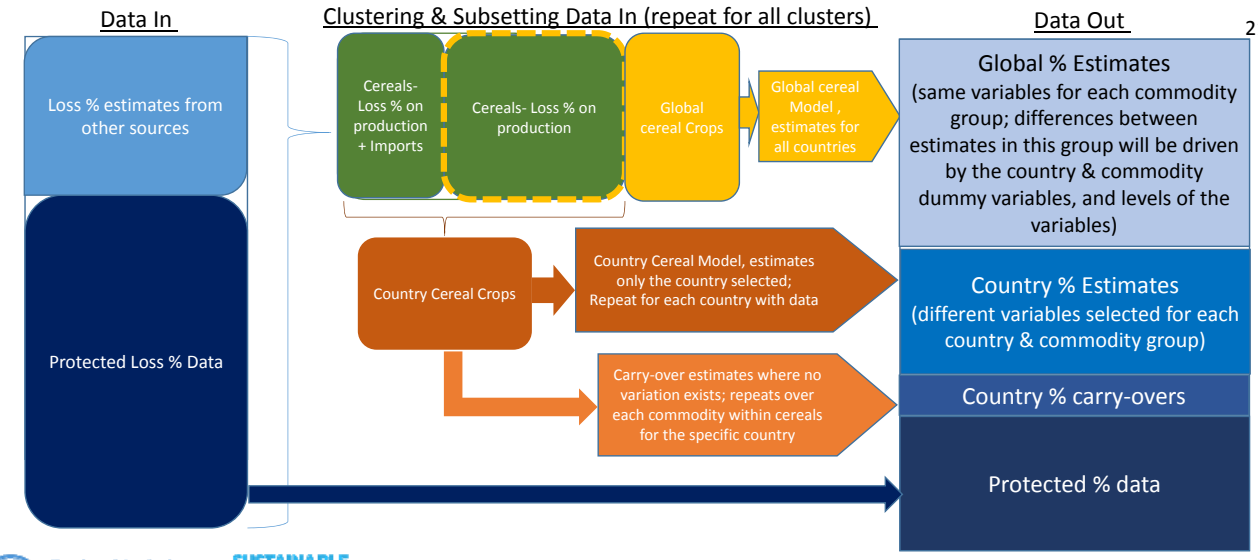
$z_{ij}^{T}$is a M-dimensional row vector of time-invariant dummy variables based on the indices i,j

$u_{ijt}$is the idiosyncratic error term

α is the intercept

- Fixed Effects Model
- Within the clusters, the fixed effects are on country and commodity
- Applied first to countries with data to estimate a country model and then to the world
- Vector of 8 variables selected from the larger dataset – chosen by the commodity group
- Model performs better in clusters with more data
- When not enough data exists a simple average is applied

Visualization of data used in the Models

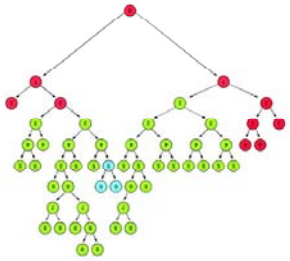


Variable Selection

- Variables were chosen base on their predictive performance in a RandomForest algorithm, as to allow for variability between different supply chains and countries
  - by each country in the cluster
  - intersected to find the common variables across the cluster
  - Second order effects were taken into account with variable first and second order interactions
- Random effects model was used to estimate the coefficients, allowing for variation within the cluster on the indices to decrease heteroskedastic errors
- Testing is being done on other variable selection methods to see if performance can be improved

Random Forests

- Standardizes the selection process of variables across clusters
- Bootstrap samples from the data to iterate over variables to find best fit
- Uses decision trees to select the best variables
  - Select m variables at random out of all M possible variables (independently for each node).
  - Find the best selected m variables and ranks them based on fit
  - At the next node, choose another m variables at random from all predictor variables and do the same
  - Trim the forest for collinear variables
- Strengths and weaknesses:
  - Able to deal with unbalanced and missing data.
  - Cannot predict beyond the range in the training data
  - May over-fit data sets that are particularly noisy



Handling Missing Explanatory Data

- Missing data in the explanatory variables occur
  - The datasets used in the explanatory variables are smoothed before undergoing the variable selection for consistency
  - Various statistical packages can estimate missing data
    - Smoothed by country
    - *Use sparingly* and consistently
- Example – The Logistics Performance Indicator is only collected every 3 years, data in the interim needs to be imputed, for it to be relevant.

Protected Loss estimates – Adjusting for import dependent countries

- One adjustment that was made for import dependent countries was to apply losses to production plus imports (105 unique country/commodities)
- Rule applied:
  - This was done if the $\frac{loss}{production} - \frac{loss}{production+imports} > 10\%$
  - This was not done for all countries – as it deflates loss percentages, exacerbating the low loss levels in the system.
  - These loss percentages including imports are only used for the specific country and commodity cluster and not used in the global dataset.

Known Unknowns/Gaps

- Few reporting on Fruits and Vegetables in the Annual Production Questionnaire
  - Also no reporting on Milk, Meat and Fish – but studies in these areas are more rare, regardless
- Countries have losses data that haven't been incorporated or collected
  - Canada, UK and Turkey report under estimation of the data and have additional sources
  - Most data collection happens at the subnational stages and then is mixed in sometimes unknown ways to get a loss estimate.
- Challenges getting losses right with differences in export/imports/production dynamics

Loss Estimates – by stage, region, source

- In the process of improving the data collection on losses a dataset has been put together from a variety of sources.
- These sources have been tagged by how the data was collected, and the following were used in the modeling effort.

Type of Data Collection	Included in the model
Expert Opinion	x
WRI Protocol	
APHLIS	x
Field Trial	
Estimates Existing in the SWS*	x
Survey	x
FBS/APQ	x
Crop Cutting Field Experiment	
National Stats Yearbook	x
Rapid Assessment	
Lit Review	…
Laboratory Trials	
Census	x
NationalAcctSys	x
Case Study	
Modelled	